



Importance of Customer Churn Prediction Using Machine Learning and Providing Recommendation Plans in Mobile Telecommunication Industry in Sri Lanka

**A dissertation submitted for the Degree of Master
of Business Analytics**

**IMMB Illangasinghe
University of Colombo School of Computing
2019**



Approval for Submission

This is to certify that the Dissertation on “Importance of Customer Churn Prediction Using Machine Learning and Providing Recommendations Plans in Mobile Telecommunication Industry in Sri Lanka” by IMMB Illangasinghe has been accepted in partial fulfillment of the requirement of the Master of Business Analytics degree program of the University of Colombo School of Computing, Colombo, Sri Lanka.

Approved for Submission.

Dr. D.A.S Atukorale

Supervisor

Date: 2021/09/10



.....

Dr. D.A.S Atukorale

Deputy Director,

University of Colombo School of Computing,

University of Colombo,

Sri Lanka.

Declaration of the Candidate

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.


To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: IMMB Illangasinghe

Registration Number: 2018/BA/015

Index Number: 18880153

Signature:



Date: 2021/09/10

Table of Contents

Table of Figures	iv
List of Tables	vii
Acknowledgement	viii
Abstract.....	ix
01. Chapter - Introduction.....	1
1.1 Chapter Introduction	1
1.2 Overview	1
1.3 Background of the Study.....	4
1.3.1 History of Cell2Cell	4
1.4 Motivation of the Study.....	5
1.5 Objectives of the Study	6
1.6 Research Question.....	6
1.7 Scope of the Study.....	6
1.8 Feasibility of the Study	7
1.9 Structure of the Study.....	7
02. Chapter - Literature Review.....	9
2.1 Chapter -introduction	9
2.2 Traditional Methods of customer behaviors prediction	9
2.3 Churn Prediction in Telecommunications.....	10
2.3.1 Customer Satisfaction.....	14
2.3.2 Factors influencing customer satisfaction.	18
2.3.3 Customer loyalty.....	21
2.3.4 Customer Retention	22
2.4 Churn Management Framework.....	26
2.5 Identification of the Most suitable variable.....	27
2.6 Feature Selection	29
2.6.1 Search Phase	29
2.6.2 The Evaluation Phase	32
2.7 Statistical Models	32
2.7.1 Naïve Bayes	32
2.7.2 C 4.5 Decision Trees	34
2.7.3 Support Vector Machines	40
2.8 Model Performance evaluation	43

2.9	Research Gap and Conclusion.....	45
03.	Chapter – Methodology	47
3.1	Chapter -introduction	47
3.2	Stakeholder Analysis.....	47
3.3	Requirement Gathering Techniques.....	48
3.3.1	Data Gathering Technique.....	48
3.4	Project Management Methodology	48
3.5	Anticipated Risks and Mitigations.....	49
3.6	Data set analysis	50
3.7	Application Frame Work.....	59
3.7.1	Application Framework model.....	59
3.8	Data set visualization	62
3.9	Data set preparation.....	63
3.9.1	Missing values and outliers	63
3.10	Feature selection.....	64
3.11	Output data format data recording and evaluation methodology.....	64
3.12	Data analysis methodology.....	66
3.13	Role of the researcher.....	66
3.14	Validity.....	66
3.15	Generalizability	67
3.16	Installation methodology.....	67
3.17	Configuration and generalization of use	68
3.18	User interface and generalization of use	69
3.19	Legal, Ethical, Professional and Social Issues (LEPSI).....	71
3.19.1	Legal	71
3.19.2	Ethical.....	71
3.19.3	Social	71
04.	Chapter – Results and Analysis	72
4.1	Chapter -introduction	72
4.2	Row data Set analysis.....	72
4.2.1	Numerical Variables. Descriptive Statistics.....	72
4.2.2	Categorical Binary variables	73
4.2.3	Missing values in Numerical variables.....	75
4.2.4	Missing values in Categorical (Boolean) variable.....	75

4.3	Naïve Bayes Results analysis.....	76
4.3.1	Naïve Bayes and Missing data processing	76
4.3.2	Naïve Bayes and Outlier data processing.....	78
4.3.3	Naïve Bayes and feature selection.....	80
4.3.4	Naïve Bayes Stabilization.....	82
4.4	Decision Tree Results Analysis.....	83
4.4.1	Decision Tree and Missing data processing	83
4.4.2	Decision Tree and Outlier data processing.....	85
4.4.3	Decision Tree and feature selection.....	88
4.4.4	Decision Tree Stabilization	89
4.5	Support Vector Machine Results Analysis	90
4.5.1	Support Vector Machine and Missing data processing	90
4.5.2	Support Vector Machine and Outlier data processing.....	93
4.5.3	Support Vector Machine and feature selection.....	95
4.5.4	Support Vector Machine Stabilization	97
4.6	Ensembled classifiers Analysis	99
4.6.1	Naïve Bayer and Decision tree Ensembled classifiers	99
4.6.2	Decision tree and Support vector machine Ensembled classifiers	101
4.6.3	Support vector machine and Naïve bayes Ensembled classifiers.....	103
4.7	Accuracy differences analysis.....	105
4.7.1	Single Classifiers performance evaluation	106
4.7.2	Ensembled Classifiers performance evaluation.....	108
05.	Chapter – Conclusion and recommendations	111
5.1	Conclusion.....	111
5.2	Summary and Future Research Areas	113
5.2.1	Accuracy evaluation methodology	113
5.2.2	Ensembled classifier	114
	References.....	115

Table of Figures

figure 1: proposed model for customer retention	14
figure 2: proposed model for customer retention	14
figure 3: churn management framework	26
figure 4:key blocks of feature selection	30
figure 5:plane separating the data point	41
figure 6:the confusion matrix	44
figure 7:mean number of dropped or blocked calls	51
figure 8: month distribution	53
figure 9:highest credit rating	54
figure 10:mean monthly revenue.	55
figure 11: charhem distribution	57
figure 12: application framework model	60
figure 13: classifiers (1,2)	61
figure 14: nb classifies evaluation list wise deletion.....	61
figure 15:the confusion matrix.....	65
figure 16: installation diagram.....	67
figure 17: configuration console initial appearance.....	68
figure 18: configuration console connect with server.....	68
figure 19: configuration console without put accuracy values	69
figure 20:: user interface default view.....	69
figure 21:: user interface after connect.....	70
figure 22:test	70
figure 23: user interface after load view	70
figure 24::: user interface get accuracy	71
figure 25: model of testing nb in missing values.....	76
figure 26: nb missing data processing	77
figure 27: nb accuracy lwd and mmi difference distribution.....	77
figure 28: outlier trimming range.....	78
figure 29:model of testing nb in outlier.....	79
figure 30:nb outlier data processing.....	79
figure 31: nb accuracy and outlier factor	80
figure 32: model of testing nb in feature selection	81

figure 33: nb feature selection	81
figure 34: model of testing nb accuracy with 3 variables	82
figure 35: nb accuracy distribution.....	83
figure 36: model of testing dt in missing values.....	84
figure 37: decision tree missing data processing.....	84
figure 38 : dt missing data processing	84
figure 39: dt accuracy lwd and mmi difference distribution	85
figure 40:outlier trimming range.....	86
figure 41: model of testing dt in outlier	86
figure 42: dt outlier data processing.....	87
figure 43: dt accuracy and outlier factor	87
figure 44: figure 26: model of testing dt accuracy	88
figure 45: dt feature selection.....	89
figure 46: model of testingdt accuracy with 3 variables	90
figure 47: model of testing svm in missing values	91
figure 48: svm missing data processing	92
figure 49:svm accuracy lwd and mmi difference distribution	92
figure 50: outlier trimming range.....	93
figure 51:model of testing svm in outlier	94
figure 52: svm outlier data processing.....	94
figure 53: svm accuracy and outlier factor.....	95
figure 54: figure 26: model of testing svm accuracy.....	96
figure 55: svm feature selection	96
figure 56: model of testing svm accuracy with 3 variables.....	97
figure 57: base diagram for nb, dt ens classifiers	99
figure 58: graphical representation of accuracy nb&dt, nb, dt with missing value	101
figure 59:base diagram for dt, svm ens classifiers	101
figure 60:accuracy difference distribution dt& svn, dt, svm with missing value	103
figure 61: base diagram for nb, svm ens classifiers.....	104
figure 62:: accuracy difference distribution nb&svm, nb, svm with missing value.....	105
figure 63:single classifiers roc curve	106
figure 64: accuracy number of time maximum and minimum	107
figure 65: single classifier accuracy difference comparison	108

figure 66: ensembled classifier roc curve	108
figure 67: accuracy number of time maximum and minimum	110
figure 68: ensembled classifier accuracy difference comparison	110
figure 69:the confusion matrix.....	113

List of Tables

table 1:sector wise break down	11
table 2:customer retention as churn prediction accuracy	22
table 3:comparison of accuracy, learning time and number of support vectors	40
table 4:risks and mitigations.	49
table 5:cell2cell data set overview	50
table 6:customer care service attributes in cell2cell dataset	50
table 7:dropblk mean number of dropped or blocked calls	51
table 8: customer demography data	53
table 9: months distribution	54
table 10:customer credit score	54
table 11: high credit rating	55
table 12:bill & payment analysis in cell2cell data set.....	55
table 13:summary statistic revenue.....	56
table 14:customer usage pattern (behavior pattern).....	56
table 15:summary statistics chargem	57
table 16:value added services in cell2cell data set.....	58
table 17: churn dependent variable	58
table 18: all combination of classifiers	65
table 19: numerical variables. descriptive statistics.....	73
table 20: categorical binary variables. descriptive statistics.....	74
table 21: missing values in numerical variables	75
table 22:categorical (boolean)variable missing values	75
table 23: ens, nb, dt comparison.....	100
table 24: ens, dt, svm comparison.....	102
table 25:ens, nb, svm comparison.....	104
table 26: individual classifiers comparison.....	107
table 27: ensembled classifiers comparison.....	109

Acknowledgement

To all people in this world who tries make a better world to all.

Abstract

This research study is focusing on predicting customer churn using machine learning in the Sri Lanka telecommunication industry. This model can apply to the telecommunication sector of Sri Lanka to manage or early identification of the customer churn which has major commercial impact in this industry.

The review of the literature is referred to identify the previous related to customer churn and predictive model development and how research works were conducted to predict customer churn prediction using machine learning and other statistical methodologies. And what significant factors contribute and what are the algorithms used and performance and accuracy of the output were considered from published academia.

This research is done using Cell2cell openly available dataset in the telecommunication domain which is a standard featured dataset which can be generalized to the Sri Lanka Telecommunication industry based on the assumption of generalization.

Dataset is subjected to preprocessing with missing value and outlier treatments. Filter method is used for the feature selection using Chi-square and Anova for numerical and categorical variables.

The research study will be conducted using supervised learning algorithms Naïve Bayes (NB), Decision Trees (DT), and Support Vector Machines (SVM). Here single classifiers performance is analyzed. After that under the voting method ensembled classifiers are formed using two singles classifiers NB &DT, DT & SVM, SVM & DT and performance are evaluated.

Performance of the classifiers are examined using ROC curve and Discrete statical method in this research. Overall accuracy of the predictive models is the key output parameter which is related to customer churn which was analyzed by Discrete statical. All the features of the data set visually represent with histogram and box plot tools to get an overview of the features.

Application developed on distributed architecture which data set and predictive model is running on server and remotely admin and user interface can connect to it. Admin interface is supported for deep analysis of dataset and user interface can be used for future prediction.

01. Chapter - Introduction

1.1 Chapter Introduction

This chapter presents framework of the overall research study, which is to predicting customer churn prediction using machine learning in Sri Lanka telecommunication industry using supervised learning approaches. This present a brief overview of scope of the research work which is covered by the study of objective and literature review of the study. This research proposal elaborates the context of the telecommunication industry as customer that benefited of the study, in methodology and methods will use to conduct this study accordingly.

1.2 Overview

In last three decades there was a continuous growth in information communication sector in worldwide and telecommunication industry was boomed in technology, Size of market and Capital investments. Telecom companies are starting to realize the effectiveness of churn prediction as a way of generating more profit, especially when compared to other approaches. An increasing investment was made in the study area of churn prediction during past years (Ascarza, 2016). Under the telecommunication device industry sector mobile phone is the one of leading industry is recognized as the rapid increase of the market share and penetrate size of market (Keramati, 2011). With the high demand in market was caused a competition among the telecommunication operators to be acquired maximum market share. In enhancement of the technology the capabilities of telecommunication operators were increased continuously. It was caused to offer enrich value added service to customer in the economical telecommunication product such as mobile device and packages. These all factors leads a to saturate the available customer's market and in spite of the beginning that the mobile telecommunication industry is said to be in, the competition has driven to take a new path to acquisition of new service subscribers to retention of the existing consumers. With saturation of the market was drives every telecommunication operator to heavily invest on the marketing with aim of attracting customers from competitor companies (Hashmi, 2013).As a result every

telecommunication operator started to experience customer churn impacting many negative encounters (Hashmi, 2013) and continuous incremental marketing cost.

Customer churn is defined as the early identification of an existing consumer to terminate contract with the current service provider (Chitra, 2013). Due to impotency and the urgency of the needs to manage customer churn and identifying the influencing factors, churn prediction of the customers has become inevitable for mobile operators, focus on the mobile telecommunications industry. Customer churn prediction means the identification the potential subscribers who are probable or potential to swap the service providers due to a number of reasons (Coussement, 2008). It is presented that the main focus in the churn prediction models is to early identification on customers that are potentially to transfer in order to guarantee that strategies of customer retention are focused to target them and prevention and mitigation of probability of transferring to competitors mobile operators (Lu, 2003). This consideration may help telecommunication companies in flourishing through improvement of the revenue and stability of the organization.

Different sectors in the industry are struggling with customer churn in different ways and churn prediction has become a serious issue that has been highlighted in different aspect. However, in the telecommunications sector has subjected to many number of studies in customer churn (Verbeke, 2008), (Huan, 2012) (Hashmi, 2013), customer churn can be traceable with aid of many attributes parameter or implications. However, a growing of bad debt, operational costs increases, and reduction of revenue generation is a main predicament that telecommunications companies badly paining point of customer churn. In the other hand acquisition of new customer is a heavily costed marketing operation. For example, (Lu, 2003) shows that the commercial impact of losing a customer is that it would cost over ten times or more with respect to new customer acquisition in comparisons to retaining active customers. In 1995 USA a conducted a survey-based study program, showed that the total cost of acquiring new customer was about USD 300 in comparison to USD 20 per person for retention of active consumers. However in 2004 there was a similar type of study conducted as an outcome, the cost of on boarding a new customers remained the same while that of retaining active ones rose to USD 25 per person in 2004 (Brown, 2004) A US market survey on telecommunication industry demonstrated that these costs later reverted respect to original originals. According to (Seo, 2008) it implies that it cost fifteen times more to acquire new customers compared to retaining existing .

It was clear that acquisition of new customer from competitor account is a costly operation and it also subject to continuous increased which creates service providers gather data and analyses of data focusing the consumer to get better understanding and insights of their customer base, insight on consumer behaviors and preferences.

In addition to that, onboarding of new customers frequently implies taking such customers from competing service providers accounts in well mature market situations (Keramati, 2011) based on that, it is more important for telecommunication organizations to focus on customer retention is really worth from a financial gain and marketing perspective. There was a research finding, the telecommunications industry service experiences a course of between 20% and 40% of churn rate annually. According to one research (Barrett, 2003). In the US sole, researcher (Seo, 2008) assessed that the churn rates of closer to 46 percent for one particular mobile network provider in 2001 alone. The churn rate annually was reported to be between 15% and 20% by the Telecom Regulation Authority in Norway and Norwegian Post (Svendsen, 2013).

Telecom companies are starting to realize the effectiveness in cost saving in churn prediction as a way of generating more profit, especially when compared to other approaches and which define the stability of the market. An increasing the investment in subject of churn prediction in was made in during past years (Ascarza, 2016) With evidence from research and experience of growing churn rates in the telecommunications sector, the factors that associated to churn are still not explicitly clear to most of telecommunications companies. The complexity to rationalize customer churn growth with distinguish classification of churn into two class named partial and total churn (Colgate, 2001).

Partial churn can be introduced as service providers lose relationship the customers in one business segment and this makes it considerably difficult to identify this in data set (Siddiqui, 2011) As an example, a customer may further progress to use the voice services of the current telecommunications operator, but change or shift to using the internet services of an another operator in competitor domain. This is caused to create to partial churn. However, full churned customer is usually not as difficult to identified because customer closed their entire access accounts with a mobile provider and shift to another mobile operator to enjoy the service (Gurjeet Kaur, 2012).

Customer churn is influenced by internal and external factors and as per (Gurjeet Kaur, 2012) shows that an estimated 35% churn rate is considerable to external factors that out

of the control of a mobile service provider from point of customer satisfaction. Internal factors will be the caused for remaining customer churn that are generally within the direct control of the mobile service provider. These service provider controllable factors are product problems, location ,product or pricing structure , lack of quality in service attributes, user convenience, merger, and problem identification and resolution in service operation (Trubik, 2000). Based on above studies, mobile service providers have direct control on the factors that initiated switching in over 60% of consumers who shift or churn. Thus, the improvement of strategies for betterment or enrich of these factors besides identification of churn customer is more important in mitigating customer churn rates, early identification, and its associated harmful implications.

This research is a based on quantitative research approach on system generated logs attempting at recognizing and modelling the key factors of mobile performance criteria from a customer retaining perspective, measure the direct impact of the key customer-perceived mobile services performance dimensions on and intentions of churning, and compare Cell2 Cell mobile service operator customer-perceived mobile performance with respect to major competitor.

1.3 Background of the Study

1.3.1 History of Cell2Cell

Cell2Cell is the 6th largest mobile communication service provider in the US, with 10 million subscribers approximately. It covers with more than 2900 cities and communities, serves more than 210 metropolitan markets, and covers almost all 50 states in USA. Cell2Cell has one of the largest retail store networks in the country and it has over 20,00 employees for its operations.

The company announced public in 1992. Just after, its stock price grew rapidly and strengthen Cell2Cell's power to enter its most ambitious mergers & acquisitions period. The major milestone was appeared in 1992 with the FCC auction of new digital service licenses, when it surprisingly won 18 major trading areas (MTA). With these major achievements Cell2Cell was funded by major public and private organization and it began and expansion process that led to key player in the market.

The greatest strengths of the Cell2cell are its network infrastructure and marketing capabilities. The company secured coverage with its own network in the primary target markets while the affiliates brought service to secondary markets. Cell2Cell was operating in 150 retail stores with full-service stock and Cell2Cell phones and services available in more than 8,000 retail outlets presenting large distribution system of stores throughout the country. The company's products also pushed at Best Buy, CompUSA, Wal-Mart, and Office Depot. Recently it has expanded its services online through Barnes & Noble.

Like other competitors, Cell2Cell is vulnerable to the current state of the stock market. The company has cancelled one of large stock offering recently and is currently focusing on other ways to generate funds. Also involved in the 4G race, Cell2Cell experiences the financial crunch from investing in the requisite network infrastructure with these technology advancements. In addition to that, the company has recently reported a decrease in new customer acquisition and a growth of customer churn.

1.4 Motivation of the Study

In last decade Sri Lanka mobile market was expanded and saturated with respect to citizen and mobile penetration. With respect to this condition there will be a smaller new market segment from who will be the child enter into mobile market. This market size is not larger enough for new customer acquisition of 4 larger telecommunication companies in Sri Lanka. This led to all operators try to attract other operators' customer into their own account and initiated the customer churn. With poor service quality, bad debt and some other reasons make it more complex.

All the telecommunication operator experienced that there will be a huge investment need to be allocated to for new customer acquisition and churn management is much easier straight forward and economical with respect to it.

This finding is led telecommunication industry to early identification or prediction of potential customers who might be convert to churn customer. This advance identification process is name as churn prediction and under it is necessary to identify behaviors attributes of these types of potential churn customers.

Once they identified in accurate manner there will be cost effective ways to retain them in the network which will be the aim or goal of this process.

1.5 Objectives of the Study

To find out the best classification algorithm for predicting customer churn in telecommunication industry and make model for identify the best practices or based on attribute which can be significantly impacted on customer churn.

1.6 Research Question

There are many classification learning algorithms that used to predict identification on the customer churn in mobile telecommunication industry and it is important to study whether the different algorithms behave differently. Therefore, following research question need to be addressed.

“To examine different classification algorithms and identify the best classification algorithms out of examine to predict the customer churn and identify how to retain the existing customer base.”

1.7 Scope of the Study

Churn prediction was a luxurious requirement which nice to have for the operators. With the saturation of the mobile market and perception changes of customer due to mobile experience caused the churn rate. In current environment customer churn monitoring and prediction is an essential requirement in telecommunication industry in Sri Lanka.

In this research study is referring historical research work under the literature review and understanding the available knowledge on this area and methodology of research were conducted. Based on these facts it can highlighted the areas where the knowledge gaps are arising.

In this research is using the Cell2Cell dataset, using supervised learning method identifying the probable classifiers which can predict the customer churn with better accuracy. Further identify the key variable contribution based on statistical method to identify what variable define or impact to the customer churn.

Scope of research is limited to Naïve Bayes, decision trees and support vector machines classifiers and postprocessing or optimization with feature selection approaches.

Generalization of the finding of Cell2Cell will be apply for Sri Lanka telecommunication industry is the ultimate scope boundary of this research.

1.8 Feasibility of the Study

When telecommunication operator end user or customer is considered systematically, he can be defined in large number of variables. Customer demographic data, customer operational data, customer related service inquires service, usage and packages related data and payment related data will be the main categories and under each there are number of variables. With the time also it can be a seasoned or cyclic way of customer behaviors.

With large number of attributes or feature list it is necessary to extract the relevant features which has more negative or positive relationship or contribution to customer churn. With the Cell2Cell data set and feature selection method in machine learning it can be select the statistically most optimum feature set.

Cell2Cell dataset has missing values and outliers which need to be considered under the preprocessing of the dataset. So, data cleaning and validation part pays a major role in the research and by using python in built libraires this process can be implemented.

Cell2Cell Data set has over 70,000 data record which is sufficient to train the classifiers and validate the output with the training data set. Dataset no off instance is larger enough for the research work with respect to classifier programing.

Cell2Cell is USA based 4 the largest operator and which produced this dataset actual basis. Feature or attributes in this dataset is like other any telecommunication operator's dataset so this research work can be generalized too.

1.9 Structure of the Study

This research thesis is arranged in following order

- **Chapter One – Introduction**

Overall research study reporting framework is summarized and presenting in this chapter. Under this chapter it is presenting overview of the research study and what is the background to perform this research study under customer churn prediction in telecommunication industry in Sri Lanka. Further it is presenting objectives of

research work and scope of the study with research boundaries and limitation. Feasibility study of the dataset and its attributes, quality of the data and how it can generalize to local context is explained under the feasibility study.

- **Chapter Two – Literature Review**

This chapter is presenting how churn constructed in the organization environment and what are the available academic findings based on customer churn prediction in telecommunication domain. What are the classifiers used and how supervised learning is applied on dataset and find out the prediction accurate? Based on these academic findings research gaps are going to address is describe in this chapter.

- **Chapter Three - Methodology**

This Chapter will be presented details of methodology used to analyze the data set and used techniques for the identify the prediction. In addition to that describe about training data set and test data set and attributes in relating to selected data set.

- **Chapter Four – Analysis and results**

In this chapter analysis and evaluates the outcomes of the implemented machine learning techniques and proposes a framework that can be used to predict the credit card defaults.

- **Chapter Five – Conclusions and recommendations**

In this chapter summarizes what is the key contributions or output of this research and highlights opportunities for future research in addition to recommendations and conclusions.

02. Chapter - Literature Review

2.1 Chapter -introduction

Literature Review chapter presenting review of available literature in the areas of customer churn, satisfaction or dissatisfaction of customer, and the other various attributes or factors caused for customer satisfaction and retention, especially in the mobile communication industry. How to process the data set and feature selection methodology related theoretical background will be considered under this chapter. And here it is presenting the supervised machine learning algorithm and classifiers, combine classifiers. Performance evaluation methods of the classifiers also review in this chapter.

Additionally, the author reviews and presents literature related to commitment and proves the theoretical interconnection between customer satisfaction and customer churn. The review primarily identified the potential of minimizing churn based on increased commitment as feature or attributed to more customer satisfaction. This chapter presents with a section covering literature about customer retention, satisfaction, attributes which are related to customer churn in telecommunication industry.

2.2 Traditional Methods of customer behaviors prediction

There are some older techniques to dealing with customer behavior prediction include a semi-Markov process as used by (JENAMANI, 2003). In here it is proposed a model that considers customer behavior. Designed of the discrete-time semi-Markov process was based on probabilistic model for which can use in analyzing complex dynamic systems. Slotnick and Sobel (SLOTNICK, 2005) also used a semi-Markov process.

Mixture transition distribution (MTD) was introduced by (PRINZIE, 2004) to investigate purchase-sequence patterns. Estimations Markov chains in higher order were allowed by Mixture transition distribution and facilitating interpretation of management aspect using smaller transition matrix. Markov chains were also consist with by Ma et al (MA, 2008) for relationship marketing to customer's lifetime value (CLV) estimated value.

Customer satisfaction index scores (SAT), year of data (YEAR), customer loyalty index score (LOY), the average ease of comparing quality differences (EQ) and relative quality importance (RQI) are five firm-level variables which is used by Auh and Johnson

(CHIANG, 2003) and introduced a simple learner model for customer behavior prediction. In the research work of Prinzie and Van Den Poel (PRINZIE, 2004) did not offer clarity on how these variables are selected for this model. Customer's purchase and consumption habits could be evaluated as cumulative evaluation with Customer satisfaction index scores. The relative impact that the perceived quality and value has on customer satisfaction was determined by relative quality importance.

As an alternative regression was used to test for linear equality restrictions. Qwn algorithm for determine potential churners were introduced by Chiang and team (CHIANG, 2003), which was named 'goal oriented sequential pattern' by the team. This work was based on association rules, defined as a technique that mining relationships amongst variables. Finding out association rules was defined as two steps in the research work. In two step process first step was used to identified the large itemset and second step is to create the association rules by deep exploiting the large itemset a priori algorithm were considers in the second step for the exploration of association rules. The a priori algorithm interprets rules using a direct sequential process to conclude relationships in the database.

For the purpose of understanding customer behavior Tao and Yeh (TAO, 2003) document two database marketing tools named USC (Usage Segment Code) and NRE (Net Revenue Equation). It is described that customer retention is one of the non-marketing activities that these two tools can be used for. These tools are not used for predicting most likely defectors, but to help in deciding on the marketing strategies to be used if a customer calls the company wishing to cancel the subscription.

2.3 Churn Prediction in Telecommunications

According to this research (Pawar, 2011) biggest revenue loss of the telecommunication organization is the customer churn happen trough out the year and caused a huge financial lost and ultimately may lead to sickness of the company and early identification of potential churn customers was the key objective. It has used in house customer databases, External sources, Research survey data sets with following breakdown.

Category	Number of records
Government	35
Business	125
private	735

TABLE 1:SECTOR WISE BREAK DOWN

Data recodes were subject to data cleansing and remove record with missing values call_date, call_time, and call_duration and process using MathLab Software. As per result Approximately 82.33% were not churning and 17.67% churning customers during research period. As conclusion the model can be used to identify the potential churning customer and with further work, the scope of model can be developed to identify insolvency prediction of telecommunication customers.

This paper presents (Li-Shang Yang, 2006) the first efforts of a major Taiwanese telecommunications provider how to acts with market situation in which they were operating was unheralded anywhere in their industry (Li-Shang Yang, 2006). From a base rate of only 7%, the penetration rate (mobile users per 100 people) had escalated to 106.15% within 5 years. It creates Unavoidable (involuntary) (Li-Shang Yang, 2006) churn and Voluntary churn in the industry. In the data set Over 170 variables were identified in data set If a pattern was identified was tested with a chi-square test. Any variable whose univariate test had a p-value <0.25 was used as a candidate for inclusion in the prediction model. training and test were data sets which ranged in size from 50,000 to 1,000, 000 records. Data set evaluated under decision tree and logistic regression and data mining model effectiveness, even with a low churn rate (0.55%) the model provided lift figures of near 100.

Many researchers have defined the customer churn is the leaving the existing regular customer by terminating the relationship with the service providers (Hwang, 2004), (Berson, 2000). Customer churn can be either internal or external in nature of classification as per (Mattison, 2006). When a customer terminates the currently using service with mobile service provider and shift to another service from same service provider is define as internal customer churn. External customer churn is defined as fully termination of all service with current service provider and switch to another service provider and may be

involuntary or voluntary. These customer churn subtypes are discussed in research work (Yang, 2006).

When a customer shift to another alternative service provider due some reason such as change of customer location, cancelling of services or perceived superior value can be considered as. Voluntary churn in telecommunication industry. It may be exactly similar or deliberate (Jahanzeb, 2007). When service providers ending customer contract or accounts due to various reasons such as due usage bills, fraud underutilization can be considered as Involuntary churn (Mattison, 2006), (Yang, 2006).

customer churn and link between churn and its determinants were mainly investigated by previous studies (Bolton, 2004), (Ng, 2000), (Wei, 2002)). As an output of these studies there are eight churn related determinant were found out. They are swapping barriers' adjustment effect on churn (Yang, 2006) customer satisfaction , (Bolton, 2004) customer loyalty (Bolton, 2000) customer loyalty and customer satisfaction (Gerpott, 2001)and ending interviews with customers ending a service using post hoc analysis (Kon, 2004).

To predict customer churn and switching of brands some researchers used complex tools in opposite way of doing a descriptive analysis about potential customer to be churned. (Wei, 2002), (Ng, 2000), (Weerahandi, 1995). Most of previous studies highlighted the identification of direct variables or attributes effect on customer churn. (Ahn, 2006) researchers use different way of evaluating customer churn and direct variable by using consumer status as a mediator.

Base on Norwegian telecommunications customer (Svendsen, 2013) conducted a research study in form of two-wave longitudinal and investigated that perceived switching costs, customer satisfaction and customer demographics on customer churn in telecommunication industry. As an outcome of this research they have pointed that gender has no statistical relationship with the customer churn and significant effects of age, satisfaction, and interaction between customer and provider has direct relationship with customer churn. Provider side identified attributes were brand image or reputation, switching cost customer satisfaction and customer demography were controlled in this research work.

The researchers derived that a strong brand image cause improve the confidence and reduce the uncertainty of a service provider to churn specifically churn began due to customer

satisfaction (Svendsen, 2013) Previous research work help to find out the other factors influence to customer churn in telecommunication industry.

As per the research conducted by (Paulrajan, 2011) product or service price and interaction are main influencing or motivating or influencing factors for customer retention or customer churn in the India mobile market.

In this studies communication was covered areas of site of geographical coverage, voice quality of a call, and call drop rate and basically factors determine the quality of the mobile network. According to (Seo, 2008) geographical coverage and voice clarity are key factors directly influence customer churn in the telecommunications industry. Geographical coverage is key influencing factor to churn of mobile service subscription as per (Turki, 2010). Poor mobile signal, which is directly define the bad network coverage was the main reason for customers shifting to another alternative mobile network as per the (Turki, 2010). According to Rahman, Haque, and Ahmad (Rahman, 2011) argue that mobile network quality of services are critical factor related with customer satisfaction and key reason behind the customer churn in mobile services. Similar thing was presented by (Min, 2009) 13.7 % of Korean mobile subscribers or accounts subject switched their network or churned due to problem associated with signaling and network coverage issues. Mobile traffic level and traffic saturation condition also considerable impact on the customer churn in mobile network (Haque, 2010), (Munnukka, 2008) research shown that the perception of a subscriber about value for money other word price is main concerned with the perception of vales, beliefs, quality. Mobile tariffs condition and saturation were also caused to be a key impacting factor that influencing subscriber wiliness chose another subscriber (Haque, 2010), (Min, 2009), (Rahman, 2011) shown that traffic level in mobile network were an influential factor in considering subscriber switching patterns and behaviors in the Korean market (Kollmann, 2000).

Customer service and its qualitative values has also been identified as a key factor in customer churn. (Kim, 2004) pointed out many activities involved in customer service which has an impact on customer churn. Factors identified in the customer services were friendliness during reporting, processing speed, reporting ease, customer support systems and the process of handling the customer complaint and way of resolving it. This

highlighted the importance of dedication, courtesy, employee professionalism, and friendly (Söderlund, 2008)

Customer service quality was used to promote customer satisfaction, which ultimately related with customer loyalty, customer churn, as well as second time purchases. (Almana, 2014) also commented that inappropriate or slow response to customer complaints rise the probability of churning in the telecommunications domain. Malaysian subscribers perceptions about service providers directly related with customer service quality and founded by (Rahman, 2011) Telecommunication service providers brand image is highly impact improved with innovativeness as a critical aspect of user’s perceptions based on the studies of Malhotra and Malhotra (Malhotra, 2013).

As per (Oladapo, et al., 2018)they have present a framework that Customer satisfaction is subjected to customer loyalty and as final out put customer retention was improved. In other word customer chur in reduced. As per research they have proven their model using machine learning logistic regression algorithm.

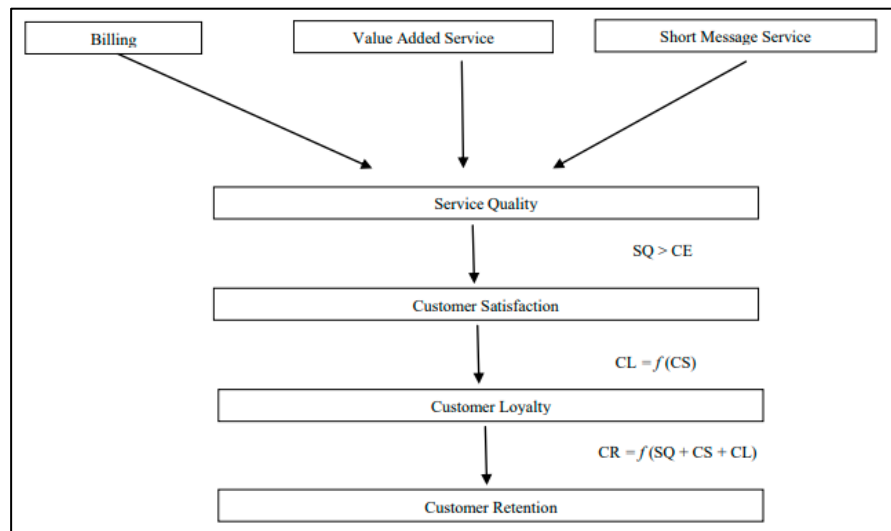


FIGURE 1: PROPOSED MODEL FOR CUSTOMER RETENTION

2.3.1 Customer Satisfaction

(Cai, 2018)as per the researcher among industries in the world telecommunication is mostly hit with churn , Mobile Telecommunication comes on top with more than 30% of defection in Europe; reaching 60% in African sub-Saharan countries and in Asia (Cai, 2018). For Example, China, if a customer’s category spends CNY 200 per month and with 10 million

customers in relevant category, then 0.5% churn rate is CNY 1 million/month of lost income directly. Also, the cost of acquisition a new customer is 5 to 10 times bigger than the cost of satisfying and keeping an existing customer (Cai, 2018) so customer satisfaction is playing a key role. Logistic regression and voted perception method were applied on the data set and it was key target to apply the hybrid model and evaluated the performance. Data set was consisting with 2000 subscriber 6 months all transactional activity which represent with 23 data variables. Base on the analysis. The evaluation of the model presents that its accuracy is better than when using single model and that the results could be ameliorate when the data distribution are less skewed. And it shown that customer churn is related with customer satisfaction too.

The availability of marketing research based on customer satisfaction and customer dissatisfaction prove that large no of research that has been conducted based on this subject area (Rahman, 2014), (Almossawi, 2012), (Eshghi, 2008). There are less no of academic research compromise as to what customer satisfaction really is and the application of numerous vast ranges of definitions for satisfaction of customers proves as much. For an example, (Westbrook, 1978) well-defined that customer satisfaction is an emotional or feeling reaction which is a subjective matter. They further presented that customer satisfaction results from a complex process that needs background knowledge of psychology of subject customers. As per them, the range of emotion is covered vast area with, for example surprise, pleasure, contentment, or relief. Satisfaction is impacted, in the end, by customer expectations and the gap between experiencing quality and expected quality, called "expectancy disconfirmation".

As per NBRI (NBRI, 1982) satisfaction of customer is the service operator ability to satisfy the business needs, emotional needs , and psychological needs of its end customers. In (Burdiek, 1993), customer service has been shown as the single factor, that if improved in the short run, yield better long lasting customer satisfaction that are vital to overall corporate success. (Burdiek, 1993) further shows that a model for customer service quality in mobile communication service is not yet defined, even though a large amount of capital has been allocated to model other network performance attributes.

(Westbrook, 1978) can be triggered following the understanding that the psychology of customers is not the single, ultimate factor of customer satisfaction. For an example,

(Leelakulthanit, 2011) developed a quantitative research study examining customer satisfaction related factors among 400 Thai mobile phone users. They identified marketing promotional value, corporate image of the organization, and service quality of the customer to be the most curtail factors. Thus, customer satisfaction is more than just sensible thought in the researcher's opinion. In fact, other factors like service price, service quality, product quality, special discounts, and so on also impact on satisfaction. NBRI's (NBRI, 1982) defined customer satisfaction seemed as satisfactory. However, the NBRI (NBRI, 1982) study not matched properly the influencing factors in each of the business, emotional, and psychological requirements of customer following the process to either qualitatively or quantitatively verify them. (Burdiek, 1993) researched to customer satisfaction is one-sided, emphasizing heavily on customer service for appropriate customer satisfaction. However, it has been claimed that a business requirements to get positive advantage to have a concerted approach by enhancing the interrelated services such as customer related service, network related service and operational related service to improve the total customer satisfaction in today's world (Almossawi, 2012). (Kotler, 2009) described satisfaction as the sense of disappointment or gratification that each customer has and that outcomes from get comparison an offered product's perceived outcome or output performance against customer expectations. Additionally (Kotler, 2009) portray satisfaction as a sense whose initiation is trackable to relatively analysis of expectation against supply or real experience of an end product. This definition has used support from many number of research studies and apparently matches the definitions provided in studies earlier the research by (Kotler, 2009) such as (Oliver, 1997)

Other academia (Hansemark, 2004); (Wells, 1996); (Solomon, 1996) presented that customer satisfaction is the subjective personal human quality or emotional feeling that customers shows towards a offered service or product after using it. The way it defines is somewhat responds to the rhetoric theorized in the previous paragraph as it overlooks the need for biased or anticipated outcomes towards an offered service or product. Altogether, literature says that many of customer satisfaction research developed the Disconfirmation of Expectations Model in shaping and rationalizing customer satisfaction. The Disconfirmation of Expectations Model assumes that customers assess product or service quality or performance in comparison method of anticipated quality or performance value against a base expectation (Motley, 2003). In other words, customers turn out to be satisfied when their perceived performance surpasses or at least matches their expectations.

With other consideration customer dissatisfaction happens when identified performance falls below customer expectations (Mittal, 1999), (Banker, 2000). Even with the contentions how it's define the customer satisfaction, which is a primary cognitive measure that attribute or characteristics predominantly in consumer and literature (Zorn, 2010). With widening of the expectations that enhance customer dissatisfaction or satisfaction as an result of comparison is also presented in literature showing that expectations of customers towards a particular service or product are bounded to the perceptions consumers had on the organization. These perceptions may contain former experienced pricing fairness and creditable performance (Dover, 2006). Nonetheless, customer expectations have time dependency, which express why they demonstrate highly complex levels of variability. This is very generally in the service-oriented industry (Chandrashek, 2007) likes as the mobile telecommunications industry. further, consumer behavior can be labeled as stable throughout a large range of distributed satisfaction till it reaches anticipated level or falls below a pre define threshold level (Chandrashek, 2007).

Number of researchers (Olsen, 2003) identified that satisfaction may assume two forms. These two forms consist of transaction-specific and overall satisfaction. The former is linked with an instantaneous after-purchase scenarios. After that it has to do with customer evaluation of their transactions' narration in its entirety and encompasses satisfaction with specific firm qualities like physical facilities, and customer satisfaction with services or goods purchased (Zorn, 2010). Based on repeated experiences, transaction-specific satisfaction has lower influence on impact with respect to overall satisfaction (Olsen, 2003) In fact (Zorn, 2010)s claim that overall customer satisfaction can also be told as generally service quality in literature in the mobile service industries. Furthermore, this research contend that overall satisfaction is constitutes a principal antecedent to customer retention and replicate purchase.

Generally, expects in the market said that both overall and specific satisfaction prior length of relationship and behavioral loyalty (Ganesh, 2000), (Dover, 2006). However customer satisfaction is not by itself and in itself a proved of loyalty among target customer base (Ganesh, 2000), (Dover, 2006).. Without having necessarily pre-empting the following the literature related to the loyalty and customer retention, it is worth noting that the association between loyalty and satisfaction is sophisticated and heavily nonlinear especially due to the presence of other factors that apply influence on the time period of the relationship.

These influencing factors are recognized from existing literature and consist pressure imposed from society, quality of product or service (Oliver, 1997), commitment and assessments of pricing (Bolton, 2004) among others. Moreover, (Chandrashek, 2007) presented that the rigidity of satisfaction conclusions made by a customer suffices as a representative in the relationship between retention and satisfaction. Alternative way, customers that had less satisfaction are not likely to show loyalty and higher potential on churn. For instance, interruption in services could cause uncertainties and afterward weaken the satisfaction decisions of a consumer even after the mobile operator solves the problem in its entirety (Zeithaml, 2000).

2.3.2 Factors influencing customer satisfaction.

Service quality

In the mobile telecommunications industry, large number of literature and empirical proof underscores consistently the instrumental and overarching position of service quality in affecting customer satisfaction. For an example, a study by (Chou, 2006) following a case study way about China Mobile operated that perceived value, perceived customer expectations, perceived ease of use of offered service, and perceived quality were directly impacted by service quality and established critically customer satisfaction factors in mobile service requirements. Another academic research has been done by (Nimako, 2010) although using a method of cross-sectional survey design dealing with 1,000 subscribers drawn from 04 different mobile telecommunications networks discovered that service quality was the principal reason of customer dissatisfaction in this domain.

(Boohene, 2011) has done a research in a different way in Ghana involving 460 Vodafone clients. It was designed and developed that although satisfaction did not necessarily indicate loyalty, customer churn, service quality factor had a strong, positive association with loyalty. Altogether, (Almossawi, 2012) presented, “customer satisfaction is already built into service quality”. Simply, customers are likely to experience the service quality in positive light if their leverage band of satisfaction with the service provider is better and the services offered to customers are excellent quality.

Service quality which related customer churn is an important decision-making factor in mobile operators was also recognized (Rahman, 2011) 04 major Malaysian cities were offered involvement of 400 mobile telecommunication customers for their quantitative

research. Key influencing factor was network quality which was describe overall service quality for these studies.

Price

Price pays a significance important value as per the above literature and it is one of marketing Mix factor in marketing domain too. However the research conducted by (Herrmann, 2007) about satisfaction stated that role of the pricing strategy or the product price and service strategy or the quality of the service has yet to appeal to extensive research studies. The application and relevance of the price is highly related variable among product and service in the telecommunication domain (Chandrashek, 2007) Considering all the facts the key influence in price on the customer satisfaction is not clearly describe in the justifiable and judgmental way. Fair level pricing strategy may be shown as value for money for the customer experience service and it cost finally derive the avoiding the customer retention due to justifiable customer satisfaction. In other word customer satisfaction is a justifiable judgment of the price of the service or product which provide the feeling that it is accepted with respect to paid value. value (Herrmann, 2007)

Balance pricing strategy is improving expectations from the service or product and caused to higher customer satisfaction. Fairness of the pricing strategy is positively correlated with the customer satisfaction as per the academic studies of (Ali, 2010)else customer churn can be happened by shifting to other telecommunication operators whose has per the fair pricing strategy as per the eyes of customers. This is expressed an indication that with fair pricing strategy cause to retain the customer for long time and improve the customer satisfaction which ultimately reduce the customer churn in telecommunication industry. (Ali, 2010).

There are some argument from researchers like (Butt, 2011) stated that the price also can be cause to attracted or preventing the customer from the product or service impaction the customer satisfaction and customer churn. He mentioned the price is the exchange medium in between the organization and customer for the tangible product and intangible payment plan and service. Customer mostly willing to pay for the product and services and hesitate pays for the satisfaction (Sattar, 2012) As a result, low price has positive direct relationship with higher customer satisfaction which lead to prevent customer churn and improve the loyalty. Customer purchase decision is highly correlated with price in both product and

service purchasing and its scope of quality as per (Ahmed, 2010) As per (Garrette, 2010) customers must be interested in and have positive actions to pay the set price, which should be justifiable with the quality and his expectations.

Corporate Image

Corporate image or the business goodwill is one of the factors that impact on customer satisfaction, retention as well as customer churning. Corporate image is bound with reputation and in hand reputation is a reflection of the image in the stakeholders and specially customer how do they interpret their mobile service provider. Image can be present as corporate image or brand image subject to relevant market conditions. As per the (Nguyen, 2001) corporate image is based on the behavioral scenario and physical attributes within the telecommunication organization. Those are included architecture, various services or products, the business name, and the scenarios of which each individual interacting with customer or target market communicates the quality impression. There is another research work from (Malik, 2012) proposing that brand image is an active tool which creates the customer satisfaction and reduce the customer churn. (Foxall, 1997) have presented customer perception on product and service traits depends on the branding and brand perception. Organizations are enjoying huge short term and long-term profit by creating strong brand image on its own product and services on the target customer base (Malik, 2012). Reputation of the telecommunication service provider or seller has an effect on the service quality perception of the customer in product and service offered (Cabral, 2000).

(Rahman, 2014) has shown that customer confidence and familiarity is a key factor that depends upon the usage of the product or telecommunication service under that brand and the history of the organization. With long period historical operation of an organization develops familiarity with customer and brand images and linked with past experience of product and service for customer satisfaction (Wen-yeh, 2004). Academia from (Donnavieve, 2002) presented that the customer confidence on the brand is derivative of consumer perception on brand which directly related for buying behaviors. It means that buying behaviors of customers is a result from the attitude of customer having about that brand.

2.3.3 Customer loyalty

Customer loyalty and dimensionality are defined in many of literatures [88], [98], [102] with similarities and some differentiation also. When customer loyalty is defining it should be present both behavioral perspective and attitude perspective. Most of academic researchers were based on behavioral perspective such as repurchasing and purchasing frequency to interpret the customer loyalty [88]. As per the academic research of Ganesh, Reynolds, and Arnold [98] customer loyalty can be form of passive or active. Former way of defining the customer loyalty is maintaining the same customer base withing the service provide even the service scope in getting fewer positive situations. On other hand some researchers had shown that customer loyalty is a process and is not one-time output. As per Oliver [38] there are four stages of loyalty process and namely cognitive loyalty, conative loyalty affective loyalty and action loyalty.

Two main segments of customer loyalty are behavioral loyalty and the attitude loyalty. There are different presentation or definition was given by different research explaining these two concepts. [89]. Behavioral loyalty directly related with the repeated buying behaviors of the customers. It can compute as a percentage to total transaction to make it into numerical presentation [88]. On the other hand, attitude loyalty can be described as positive affect to continue the current relationship with the organization and wiliness to persist in the relationship. Else how much customer wiliness to tolerate expectation on behalf of the organization to continue the relationship.

This attitudinal loyalty is defined in equivalently as the relation with direct customer [90]. As per Ball, Coehlo, and Machás [88] a traditional loyalty is a subjective topic and cab be measured using questionnaire methods. Sales transaction and details can be used to interpreted the behavior transaction in most of cases [88].some researcher s like Oliver [38] presented that attitudinal loyalty is also contribute considerable amount because of both behavioral and attitude loyalties are increase same time intertwined. This provide an idea about repeat purchase or repeat buying behaviors are evident that customer has a positive affect or positive attitude loyalty with concerned organization. Attitude loyalty is clear represent or interpret by level of customer intention on repeated purchasing [88] argued that both conative and attitudinal loyalty can be types of attitudinal loyalties.

2.3.4 Customer Retention

(Study, 2018) developed an academic research on machine learning model to compare its performance of machine learning algorithm on customer retention. In here it has identified that Customer 1) identification; 2) attraction; 3) retention; and 4) development are the key variables to define customer relationship with an organization and ultimately to prevention of customer churn from the organization. Following statistical methods are used to identify the customer retention.

- 1) Regression analysis: logistic regression.
- 2) Decision tree–CART.
- 3) Bayes algorithm: Naïve Bayesian.
- 4) Support Vector Machine
- 5) Instance – based learning: k-nearest Neighbor.
- 6) Ensemble learning: Ada Boost, Stochastic Gradient Boost and Random Forest.
- 7) Artificial neural network: Multi-layer Perceptron.
- 8) Linear Discriminant Analysis

They have used telecommunication transactional data based with 17 attributes with 3333 sample data set for the evaluation studies. Customer retention is presented using the customer churn rate and models produce accuracy of the prediction.

Model	Min	Max
Random forest	0.931138	0.963964
ADABOOST	0.934132	0.963964
Multi-layer perceptron	0.93329	0.944
Stochastic Gradient Boosting	0.8900000	0.9439552
Support Vector Machine	0.904192	0.94012
K-Nearest Neighbor	0.873874	0.915916
CART	0.852853	0.903904
Naïve Bayes	0.864865	0.882883
Logistic regression	0.844311	0.873874
LDA	0.832335	0.867868

TABLE 2: CUSTOMER RETENTION AS CHURN PREDICTION ACCURACY

Based on the findings of this research ensemble – based learning techniques were recommended as both Random forest and Ad boost models gave the best accuracy.

(Oladapo, et al., 2018) conduct a research on how to use predictive analytics to improve customer loyalty and retention in telecommuting industry. In here use machine learning logistic regression for consumer complaint from the Nigerian Communication Commission statistics and reports which had three variables worth of information about 18,711 customers, with customer retention status and period. As a result, this model explained 89.3% (Nagelkerke R Square) of the variance in customer retention and correctly classified 95.5% of issues and it has concluded that model gives more than 95.5% accuracy for the prediction of customer retention.

(Hargreaves, 2019) has conduct a research on customer retention and customer churn base on the machine learning algorithm and framework. Her aim was developing a churn reduction system, while maintaining existing valuable customers and maximizing the revenue of the organization. It has used The Machine Learning Technique: Logistic Regression for the prediction model development used The training and test data sets (Hargreaves, 2019) contain 2409 and 1350 observations, respectively. Data record contain total of 21 variables: 3 numerical variables, 17 categorical variables, including the response variable, Churn, and 1 string variable Customer_ID. Data cleaning technic were used to handle the missing data and accuracy of the Logistic Regression Model used to conclude the model accuracy and Overall, the model was able to predict 76.7% of the observations accurately and predicted 80.3% of churners accurately. She has concluded that the crafted retention strategies were able to do churn reduction rate by five times, from 49.9% in the test dataset to 9.9%, and save more than ten times the costs that were invested for the retention strategies.

This section focus on what kind of researches have been done academia on customer churn and customer retention within the mobile telecommunications industry and then step by step widens this investigation by focusing on the same subject through other service sectors and how they operated. This consideration should assist to develop a clear idea on the developments and limitations of current academia with the aim of establishing the required actions in developing an improving churn prediction approach. There has been a significant number of academia into improving a methodology for the prediction of churn under the mobile telecommunications industry.

It is an obvious challenge in the mobile telecommunications industry is competitive because of the diversity of services and different segment of customer. (HAHM, 1997) have evaluated the issues by creating a decision support system which gives use of data related to customer satisfaction.

This research undertakes dividing customers into different segments by applying a 04-step process. The first step presented is the process of dividing customer in to groups, based on their applicability for target services, e.g. a standard telephone service has customer demographic variables such as occupation, age, revenue, sex, and level of education and also behavioral variables such as amount of data usage per month and benefits required attached. Step 2 is to focus understanding the search space.

This level is basically determining how each variable or attributes should be constrained. E.g. geographical location may be limited to city/street and ages of customer may be grouped into ranges like 18-25, 26-35, etc. Step 3 is to collect the information carried from step 2 and recognizes what are the variables are most crucial for breakdown of the market. Step 4 takes identifying the correlation between the critical variables so that a matrix presenting the features of each category can be constructed. Correlation analysis is done on the identified segments and measured over a period. It identifies the factors of overall satisfaction scores.

Hung and team (HUNG, 2006) were conducted a research which was focus to predict customer churn in telecommunications industry. This research aimed on the mobile Taiwan telecommunications industry. Reason behind was the geographical area where this program was based, they were not limited to the same restrictions as the researcher and have been able to conduct it indicting their research on service usage and demographic data. They have evaluated for differences using method decision tree, propagation neural networks and C5.0 to check how all that algorithm performs for the prediction of mobile customer churn. Clustering algorithm K-means clustering was also used under clustering to divide customers into groups based on their usage or billing amount. Derived groups were used to estimate customer value. A churn predictive model is developed for derived each cluster using both prediction model decision trees and neural networks. That activity reported that neural networks accuracy a marginally better performance on the segmented derived clusters but significantly accuracy performance on the full un-segmented data.

Appropriate method of reported their results, with respect to results on the top 10% of mobile customers with the lowest level of loyalty index ranked for each cluster. Actual churn was assessed against misclassifications for this 10% cluster only. subjected mobile customers were reported to give an over accuracy of 90% churn prediction accuracy. (AHN, 2006) examine the factors of customer churn associated to mobile telecommunications industry based on massive customer transactional logs and billing data using supervised learning logistic regression models. Churn prediction outputs were not given form this research focus and precisely. however, the key attributes leading to churn have been identified. Researcher has identified key attribute parameter of customer churn in the mobile telecommunications is variable “dropped calls”, i.e. calls that were disconnected or lost while having the conversation due to low signal coverage or any other technical issues related to network.

(Coussement, 2008) and team also aimed their academic research on concluding an early prediction of mobile customer churn; any how their research background service section was in subscription of newspaper services. The concept that have been the focus of their study are grid search and cross validation. The most crucial factor or attribute in their research are subscription duration, length of time since last renewal and contract ending month explaining that their project was mainly focused on features like billing and user subscription.

As per (TAO, 2003) they evaluated the accuracy of the top 10% of customers with most probable to churn. Among the focus group they claim a 95% confidence in derived churn prediction model, but researchers did not prove how far into the future (if at all) target group of customers churned after being evaluated .There is another academia focused on the churn identification within the BFSI sector (banking and insurance). Again, the prediction model was constructed on customer demographics data, behavior data and purchase/billing data. This research highly focus on examine the association of the variables in relation to customer churn with help of the Kaplan-Meier estimate of the survival (Seo, 2008) constructed a two-level prediction model for customer retention for the USA mobile telecommunications industry.

More accurately researchers’ team had access with permission a dataset consists of the data records for 31,769 customers entries from which 46% where subjected to churn. The key variables applied in their analysis were ‘gender of the customer’ and ‘age group’. The prediction model primarily used the discrete statistical approach of mean and deviation f-

test for to validate linearity, to test what was the linear relationship between retention and customer satisfaction. This test provided positive correlations among that two variables. After that binary logistic regression was applied to test if focus key variables had a direct or indirect relationship to retention. Those variables confirmed to have an indirect influence or effect on customer churn. To conclude, three models were developed based on linear regression. Researchers didn't produce exact churn prediction results in academia publication; anyhow team has presented that all independent variables significantly predicted customer retention with an overall group retention accuracy of average of 0.988.

2.4 Churn Management Framework

(DATTA, 2001) was introduced a customer churn management framework with five stage and illustrated in Figure 3

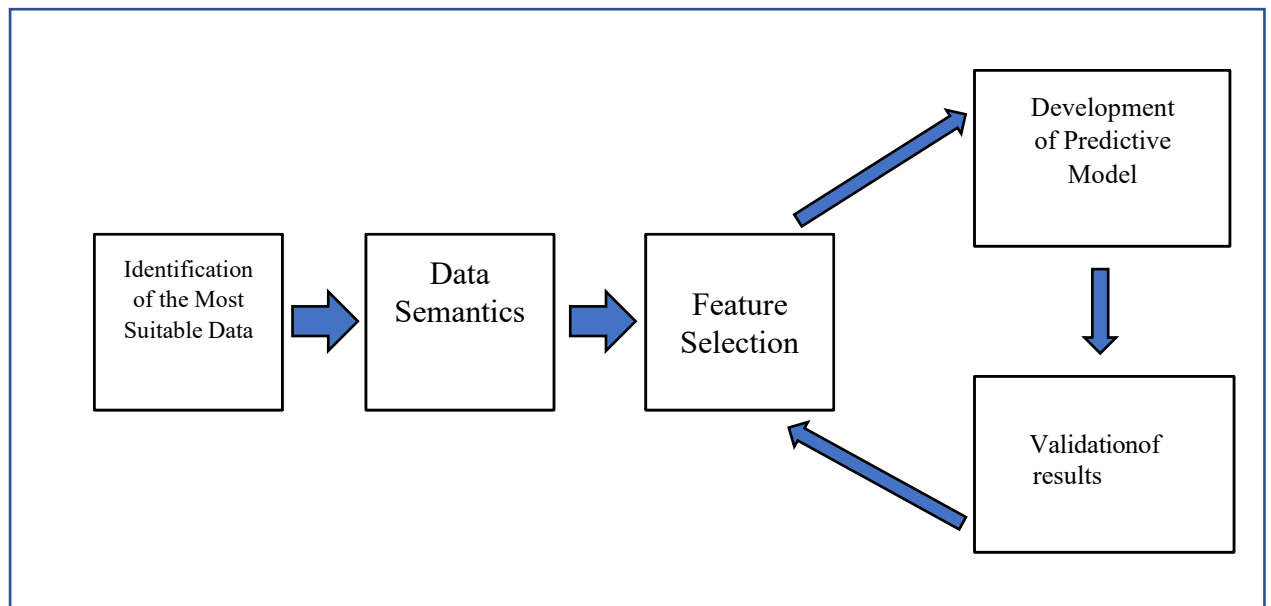


FIGURE 2: CHURN MANAGEMENT FRAMEWORK

as shown in (DATTA, 2001). The data quality is one of critical factor for to better analysis of the churn. Once data is extract form data warehouse will have better an accuracy for the prediction process and duration of the data is one key factor need to be considered. Based on the situation it can be used all available accessible data or limit to focus time duration (DATTA, 2001) .This is the first step of “identification of most suitable data”.

Step 02: in here it is necessary to get better understanding of the data semantic. It means what are the attribute or variables in target data set and what are the complete interpretation

or the definition of each. So, this stage 02 has direct relationship with stage 01. Data quality is related with many issues and paying important role and it can easily use for misinterpretation. Data semantic also covered similarities of the data attributes and its variability.

Same variable could contain different values types related to it. As an example length can be counted in units cm, inches, feet or meters, and the date can denoted in different date format (mm/dd/yyyy, dd/mm/yyyy) where these two formats could be misinterpreted this considered as heterogeneity. Another issue with the data is in same meaning but different value. For example, month represents single time unit, but it can be 28,29,30,31 days. So, its size is not a fix one and known as Ontological heterogeneity (MADNICK, 2006)

Stage 3: this is basically covering feature selection. A definition for feature selection has been taken from (MA, 2008) describe as Feature selection is how to identify the useful features to describe an application domain. Selecting enough number of features with relevancy will cause to effectively represent and index the given dataset. it is an important action to solve the classification and clustering problems with more effective way". Minimum Redundancy and Maximum Relevance(mRMR) based feature selection method is one of method to extracts the subset of features which have a higher dependency on the class labels (Peng, 2005)

Stage 4: In this level is the how to develop the prediction model. There are number of models applied for defining the prediction of an expected event including based on classification statistical, and soft computing approaches.

Stage 5: The final stage involves how to validate the model to ensure that it is deriving a higher accurate prediction for the customer churn.

2.5 Identification of the Most suitable variable

As per the figure 3 effective way of developing customer churn prediction framework is subjected to identification of most suitable variables represent the interested behaviors of data set. When it is narrow down to feature subsets they have different analytical representation for each one and research intention is to find the best suit subset has highest analytical power.

target data set can provide better interpretation for various problems and service sectors. According to, (Ng, 2000) transaction or usage data in the Internet Service Provider service sector and telecommunications industry can be mined to identify the customer churn associated with industry. Usage transactional data can be used for examining the e-customer behavior of internet user (JENAMANI, 2003) and predicting mail order based repeat purchasing behaviors (PRINZIE, 2004). One of the best ways to predict repeat purchasing behaviors of telecommunication related service or product is usage of the historical data based on findings of research studies of (VERHOEF, 2001)

This argument is followed by (HSIEH, 2004) and he proposed with support of account data and customer data, analysis of transactional data could provide the insight about promotional discounts to the customers be an effective marketing strategy. This is strengthening the importance of the stage 01 of identification of the appropriate data set for the development of model to predict the customer churn using the churn management. Framework. Researcher has identified the quality of the data will be key determinant of the accuracy and power of the final model.

In general customer data sets consists of large no of variables or attributes. There is a number of research that proposed recency, frequency and monetary (RFM) related attribute or variable produced a good source for mining the data insight and predicting customer behavior (HSIEH, 2004), (VERHOEF, 2001), (LIU, 2004)

This concept is agreed by (KITAYAMA, 2002) who state that, "What is often given as an example of a measurement of the value of the customers is the RFM analysis approach". Many insight can be derived from an analysis of the perception of the customer on service satisfaction based on mobile service provider, predictions developed based on analysis of customer spending patterns, the level of likelihood to churn if spending mobile service suddenly and decreases at once.

(LIU, 2004) was defining the RFM variables in his research studies as followings. Recency variables collected information regarding the timeline between purchases or use of service. In here if it is a lower value it represents higher probability of the customer doing a repeat purchase. Frequency variables directly related how often customer using this service from service provider. If it is consisting with higher frequency means customer has better satisfaction and/or confident with the product and services. But in other side customer user requirement may be higher due to nature of the customer environmental factors.

The expenses or cost incurred for the customer can be analysis under the monetary variables. It is denoting the total cost incurred during the consider time unit or based on actual usage. Based on the customer category (retail/ corporate) mobile service provider is interesting the customer who were bearing the higher bills for the services and always want to retain these customers. Focus example given by (RYALS, 2002)states “20% of a retail banks account holders may contribute for over 100% of its profits.

Customer loyalty data can be utilized to identify the degree of customer satisfaction within the online retailing industry by using the customer complain related logs or data (HSU, 2008) sponsoring company has advised that the United Kingdom based telecommunications regulation impose or govern by Ofcom who has defined monopoly telecommunications regulations avoiding any type of log or CDR data being the key source of analysis that could deter fare assessment and competition and the positive growth of the mobile telecommunications sector. Analysis of transactional and demographic data could break these imposed regulations through the monitoring of service chargers of competitor offers; therefore, customer reported issues and service repairs related data would be a more common basement can apply for model development.

2.6 Feature Selection

According to (SUN, 2004) Feature selection is a crucial process of identifying the attributes which are more appropriate for customer churn prediction model development. In here it is going to be target sub set of features from the data set which has capability of highlighting the important features or help to identify output class or clusters, redundancy of noisy and less informative ones and progressing data cleansing and data reduction (YAN, 2004) Feature selection process is a two stage process and first step is searching strategy or algorithm to identify the optimum subset of features and second step is how to evaluate its integrity and validity.

2.6.1 Search Phase

feature selection is initially subjected to search phase and which can be categorized in to three as optimal, heuristic, and randomized searches. Optimal search is exhaustive search method and straight forward too. Under Optimal search considered many numbers of

possible subsets grows rapidly which caused to cannot be used even medium size featured data set. Some optimal search methods that avoid the exhaustive approach, the branch and bound algorithm can be considered as an example (SUN, 2004).

Most common empirical methods are sequential backward selection (SBS) algorithm and sequential forward selection (SFS) algorithm. SFS starting from null feature set and adds the top single feature to it. SBS works in the totally other way around, by starting with the full feature set, and at each step, removing the feature that contribute less for the final performance. Some time as a hybrid method of feature selection SFS and SBS are combined. There are two variations of these hybrid methods are named as sequential forward floating search (SFFS) and sequential backward floating search (SBFS) (SUN, 2004).

probabilistic steps, or sampling techniques are used by Randomized search feature selection approach and this also method is called the relief algorithm. In here features are assigned with weights and exceeding a user-defined threshold are chosen to train the classifier.

genetic algorithm (GA) is randomized search method, attracting more popularity. The research proposes that a GA can offer better accuracy at the expense of greater computational effort. The key blocks of feature selection using a GA are shown in Figure 4 (SUN, 2004)

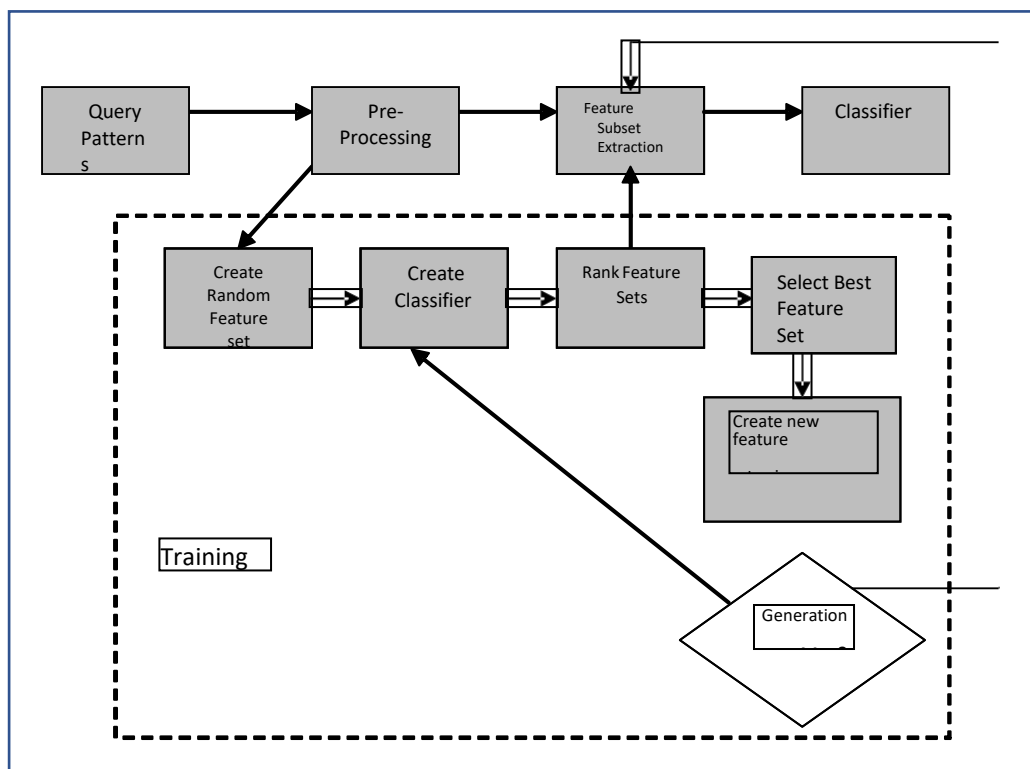


FIGURE 3:KEY BLOCKS OF FEATURE SELECTION

feature selection methods presented and used by various researchers such as Sun [120] suggest a GA search approach stating that a GA can provide a simple, general, and powerful framework for selecting good subsets of features.

Due to the characteristics of a GA, number of feature subsets are created and tested and out of all the best feature subset will be used for fitness assignment. Receiver operating characteristic (ROC) curve for feature selection was proposed by (Yang, 2006) they presented the area under the curve (AUC) can be used for selection of feature. The AUC is measured using the Wilcoxon–Mann–Whitney statistic formula. According to (DATTA, 2001) it is necessary to initially finding a feature subset from a data warehouse or target data set by manually based on those seems to be most suitable for the task. (MEYER-BASE, 1998) suggested how to use the neural networks for feature selection. This research examines the use of radial basis neural networks with class of three-layered, feed-forward network. However, it was necessary to include an extra layer to the traditional architecture to obtain a relevant features representation.

An induction algorithm was run by (Ng, 2000) for feature selection withing given dataset. (DATTA, 2001) has derived a method to predict churn for the cellular telecommunication service sector. It was a two-step process and first step use forward feature selection (FFS). According to their error rate a decision tree sorts the features to predict customer churn done by using a single feature. The first 30–50 features with the lowest error are then presented to the modelling system. Adding selected feature to existing generations is the second step of this process. A GA is used to find features subsets that are more accurately predictive than using single features alone. logistic regression was used by GA to evaluate each possible set of features. Some of the features that tend to be valuable for churn prediction with the mobile telephone service sector.

decision tree was used by (DATTA, 2001)to sort the features according to their error rate. anyhow by (DATTA, 2001)preferred the decision tree method they mention that K-nearest-neighbor (KNN) also subjected for their experiment and found that there is no differences in the performance accuracy and of the models. The Decision tree approach for feature selection could be offer highly accurate, good performance as per the studies of Ng and Liu [92]. Extracted feature should be completely independent need to be validate with the model to avoid the risk of over flitting

2.6.2 The Evaluation Phase

Filter method and wrapper method are the two method of evaluation strategies can be an applied. A wrapper evaluation method is performed using a learning algorithm that is used in the classification design to find the appropriate feature subset. Filter approach uses evaluation of feature subset which is external to the classification design and this is faster and more efficient with respect to wrapper method based on the fitness of features using criteria that can be tested faster. Both of these approaches could result to non-optimal features, when the features are incorporate on the classifier, based on poor classifier performance (SUN, 2004)

2.7 Statistical Models

(T. Vafeiadis, 2015) and team has conducted a research on “A comparison of machine learning techniques for customer churn prediction” in the telecommunication domain. They have target with Artificial Neural Networks, Support Vector Machines, Decision trees, Naïve Bayes, and Logistic regression. They have done the evaluation UCI Machine Learning Repository, included in the package C50 (R language). Data set consists with 5000 records which were allocated 67% for training and 33% for testing and subject to every Monte Carlo realization. It has given the accuracy Artificial Neural Networks (94%), Support Vector Machines (93%), Decision trees (94%), Naïve Bayes (86%) and Logistic regression (86%).

2.7.1 Naïve Bayes

As per the research (Clement Kirui, 2013) was conducted, customer churn was analysis and developed a prediction model using two probabilistic data mining algorithms Naïve Bayes and Bayesian Network, and their results compared to those obtained from using C4.5 decision tree. Confusion matrix is used as evaluation criteria and data set was 3 months user transactional data from European telecommunications in 1997. This dataset had 106,405 record which was covering 112 attribute variables with 5.6% churning rate of subscribers. As per result It was also observed that the two probabilistic classifiers (Naïve Bayes and Bayesian Network) has better performance with respect to C4.5 decision tree.

Using Naïve Bayes algorithm, it can generate a probabilistic model of the target data set. Even it is a has simplicity property Naïve Bayes has capability to compete with some advanced complex algorithm such as neural network or decision tree in some domains [124], [125]. $[x_1, x_2, \dots, x_d]$ is represented as a vector of features each. Given a training set of instances and expected outcomes is to predict the most probable class $y_j \in \mathcal{C}$ for future instance or new instance where the label is unknown. Probabilities of the classes which are derived from Bayes's theorem is used by the Naïve Bayes algorithm.

$$P(y_j|x_1, x_2, \dots, x_d) = \frac{P(y_j)P(x_1, x_2, \dots, x_d|y_j)}{P(x_1, x_2, \dots, x_d)}$$

$P(y_j)$ is the prior probability of class y_j . it is estimated as its frequency of occurrence in the training data. After observing the data, the posterior probability of class y_j is denoted by $P(y_j|x_1, x_2, \dots, x_d)$ is. $P(x_1, x_2, \dots, x_d|y_j)$ is the conditional probability of getting an instance with the feature vector $[x_1, x_2, \dots, x_d]$ among those being class y_j . And $P(x_1, x_2, \dots, x_d)$ is the probability of getting an instance with the feature vector $[x_1, x_2, \dots, x_d]$ without considering of the class. Since the sum of the posterior probabilities entire classes is one $\sum_{y_j \in \mathcal{C}} P(y_j|x_1, x_2, \dots, x_d) = 1$, the denominator 's right hand side is a normalizing factor $P(x_1, x_2, \dots, x_d)$ and can be removed from equation.

$$P(y_j|x_1, x_2, \dots, x_d) = P(y_j)P(x_1, x_2, \dots, x_d|y_j)$$

An instance will be labeled as the target class which has the highest posterior probability y_{MAP} .

$$y_{MAP} = \arg \max_{y_j \in \mathcal{C}} P(y_j)P(x_1, x_2, \dots, x_d|y_j)$$

It is necessary to estimate the term $P(x_1, x_2, \dots, x_d|y_j)$ by counting frequencies, one needs to have a large training set where it is include every possible combinations $[x_1, x_2, \dots, x_d]$ appear many times to obtain justifiable estimates (Mitchell, 1997) Naïve Bayes provides a solution for this problem by using Naïve assumption that features that define instances are conditionally independent given the class. Therefore the probability of getting the combination $[x_1, x_2, \dots, x_d]$ is simply the product of the probabilities of getting each

individual feature value $P(x_1, x_2, \dots, x_d|y_j) = \text{Product of } P(x_i|y_j)$. applying this assumption into equation to derive the Naïve Bayes classification rule.

$$y_{MAP} = \arg \max_{y_j \in \mathcal{C}} P(y_j) \prod_{i=1}^d P(x_i|y_j)$$

For nominal variables or features, the probability is estimated as the occurring frequency over the training data. For continuous variables or feature, there are two solutions can be taken. The first one is to convert to discretization on those continuous variables or transferring them into nominal ones applied the case one. The second solution is to assume that they are in a normal distribution and follow the calculation.

The term $P(x_i | y_j)$ is estimated by the fraction $\frac{\#D(x_i|y_j)}{\#D(y_j)}$, where $\#D(y_j)$ is the number of

instances in the training data set having class y_j , and $\#D(x_i|y_j)$ is the number of these instances having feature value x_i and class y_j . If the training data doesn't contain any instance with this combination of class and feature value, $\#D(x_i|y_j)$ is count will be zero. The estimate probability according to equation will be zero for every similar cases. To manage these cases, a m-estimate correction method is introduced (Gutkin, 2008)

$$P(x_i|y_j) = \frac{\#D(x_i|y_j) + mP(x_i)}{\#D(y_j) + m}$$

If the prior probability $P(x_i)$ is unknown, uniform distribution is assumed feature has k possible values, then $P(x_i) = 1/k$. The parameter m can be arbitrary smaller value.

2.7.2 C 4.5 Decision Trees

A per research study (V. Umayaparvathi, 2012) all operators invest large some of revenue on expand the market and get new customers (V. Umayaparvathi, 2012) all This caused to switching customer each other and as a result every operator experience the huge rate of churn. 6 months user transactional data be used. It has Customer Demography, Bill and Payment, Call Detail Record, Customer Care related variable were subjected to data mined. In here used Decision tree and neural network method and performance evaluated by Confusion matrix. Decision tree model produced accuracy and error rate 98.88% and error rate of 1.11% and Neural Network produced accuracy and error rate 98.43% and 1.56% respectively. it is necessary to selecting the right combination of attributes, it is observed that decision tree model surpasses the neural network model in the prediction of churn.

decision tree learning was first introduced (Hunt, 1966), has become one of the most useful algorithm in researches in machine learning methods. Decision trees generate interpretable and understandable models with clear visibility of each nodes. decision tree can be introduce as building a treetop-down using branch divide and conquers strategy. The end goal is recursively partitioning the training set, choosing one after one feature to split into nodes along the branches each time until all or most of instances in each partition belong to the same end class.

A decision tree structure develops based on four main elements: a root; branches represent to possible outcomes of feature value; decision nodes shows features used for splits; and at end leaves that specify expected value of the class. Each leaf is allocated to the class that has the most of instances inside it. To classify a new instance, it is applied to the root and follow a tree path lead by the nodes and branches downward, end at a corresponding leaf and the instance is assigned a class as per the leaf.

Decision tree learning algorithm is the mainly based on metric which measures the best way of a split. Information gain heuristic is used in ID3 algorithm as the fundamental rules of CLS (Quinlan, 1986). ID3 is further improved to C4.5 by same researcher (Mirowski, 2008) considering drawbacks. In C4.5 it is focused on concept of every split is to generates with child nodes that are purer than and its parent node. To estimate the impurity of a group (Quinlan, 1986). used entropy measure in information theory.

$$\text{Entropy}(D) = - \sum_{j=1}^c p_j \log_2 p_j$$

In here it is the proportion of D within to class j and “c” is the number of different classes. A completely pure node in includes same class has entropy equal to zero as per the definition. In the case of only two classes, a positive and a negative one, the highest entropy will be generated in a node where there is equal number of positive and negative instances. Way of evaluating the goodness of a split, entropy is compared by ID3 in a parent node to the weighted sum of its child nodes’ entropies after the split. Gain criterion function is used to compute the entropy reduction cause by a split according to a particular feature.

$$\text{Gain}(D, F) = \text{Entropy}(D) - \sum_{v \in \text{Values}(F)} \frac{|D_v|}{|D|} \text{Entropy}(D_v)$$

According to (Mitchell, 1997) where set of all possible values for feature F is denoted Values(F), D_v is a subset of D denoting all instances whose value of F equals to v. In continuous progressing step, we choose a split that result the largest entropy reduction which is equal to the largest information gain. There is a limitation on the decision tree when we consider primary key type unique features like customer ID. Or these features to split the data set results in many subsets, and each leaf only contain one records with higher information gain and zero entropy. This is useless on prediction purpose due to no prediction power. (Mirowski, 2008).

$$\text{SplitInfo}(D, F) = - \sum_{v \in \text{Values}(F)} \frac{|D_v|}{|D|} \log_2 \frac{|D_v|}{|D|}$$

C4.5 focus a so-called gain ratio criterion that provides balances between the two targets of achieving the purest child nodes and using the fewest branches.

$$\text{GainRatio}(D, F) = \frac{\text{Gain}(D, F)}{\text{SplitInfo}(D, F)}$$

Back to an extreme type of example above, the entropy of a training data set with respect to this ID no of features are larger than other all values. If there are n instances with n customer ID. SplitInfo is present with to $\log_2 n$. It interprets that this ID feature will be eliminated in the selection split selection according to the GainRatio function. However, in a case of a feature that varies very slightly where almost all instances have the same feature's value creates a problem of evaluation. The SplitInfo is then close to zero and the GainRatio approaches infinity. This can solve with first to select only features for split contain the Gain values that are exceed the average Gain of all tested features.

Among the features set GainRatio is then calculated for all of them and found the single feature which had maximum gain. The discussion so far is limited to nominal features.

Since the target task is classification, the class feature should be nominal. Continuous features can be handled by C4.5 other than the class feature. It looks for a threshold value v of a continuous feature F which gives the best split. According to base value V it split into child nodes F is above v and other creates with instances whose value of F is equal or below.

There is another problem that needs concentration is the existence of records with missing feature values. C4.5 gives separated solutions for three steps in the three construction steps: calculation of GainRatio, records partition and records classification. When only a fraction of records in set D have known value for feature F and value of F is missing in the rest of records, an amendment is done using information gain calculation.

$$\text{Gain}(D, F) = r * \left\{ \text{Entropy}(D) - \sum_{v \in \text{Values}(F)} \frac{|D_v|}{|D|} \text{Entropy}(D_v) \right\}$$

Where r is the part of records in D with known value for feature F . missing value is treated as an extra possibility of outcome in the estimation of the SplitInfo. After a feature for split has been chosen, the task is where to send records with unknown value for this feature downward the child nodes. each record will be included with weight which represents the probability that it attached child node. A record with a known value v for feature F is assigned to a corresponding child node. The probability that it attaches to this node is one, so it gets the weight $w = 1$. Only a portion of a records with unknown value is sent to each child node $w = w * p$. The portion p is estimated as the sum of the weights of records in a child node divided by the sum of the weights of records with known value for feature F . A same procedure is used to classify a new record whose feature value is missing. At a decision node, since an outcome is unknown, the algorithm discovers all possible paths which lead to leaves with different classes. A record is assigned to as the class with the highest probability.

Decision tree is grown fully in C4.5 algorithm and it continues to split until a node is completely pure. Anyhow, a node with too few records has very less prediction power. So, there is a trade-off between generalization and accuracy, an equilibrium between the pureness and the size of a node. With target of to scale down version of a tree in the last phase of tree construction, tree pruning decreases the complexity, not allows over-fitting, and improves the accuracy. There are two methods of constructing tree pruning: Theses

two methods are pre-pruning and post-pruning. Pre-pruning is carried out while the tree is being induced. It avoids the formation of node which has too smaller number of records therefore it has insignificant generalization ability. The selection of the minimum allowable number of records per leaf define based on each data set, often found by experiments.

Post pruning method applies on a fully induced tree. After the full-size tree has been constructed, it will be subjected pruning backward. Each node in the tree is applying for pruning whether the node containing its sub-tree might be converted to one leaf. All records that consist to the sub-tree are transferred to the newly introduced leaf. From the training set a separated data set used for tree constructing called pruning set is prepared. If the pruned tree output predictions \as good as or better on the pruning set, then subcomponent below that level are removed. When a separated data set is applied to identify the classification error rate, it is named as reduced-error pruning.

Another approach contains in C4.5 named pessimistic pruning. Considered a leaf contains N records, E of them are misclassified so the identified error rate for this leaf is E/N. Assuming that the error rate in a binomial distribution, its expected value being in inside a confidence interval which is bounded by two limits. For a significance level, the upper limit $U_\alpha(E, N)$ is denoted as an estimate of the error rate. Therefore, the expected number of records that are misclassified in a leaf is $e = N * U_\alpha(E, N)$. And the estimate of the error rate of a tree is derived as follows.

$$e_t = \frac{\sum_{i=1}^l N_i * U_\alpha(E_i, N_i)}{\sum_{i=1}^l N_i}$$

Where r is the part of records in D with known value for feature F. missing value is treated as an extra possibility of outcome in the estimation of the SplitInfo. After a feature for split has been chosen, the task is where to send records with unknown value for this feature downward the child nodes. each record will be included with weight which represents the probability that it attached child node. A record with a known value v for feature F is assigned to a corresponding child node. The probability that it attaches to this node is one, so it gets the weight $w = 1$. Only a portion of a records with unknown value is sent to each child node $w = w * p$. The portion p is estimated as the sum of the weights of records in a

child node divided by the sum of the weights of records with known value for feature F. A same procedure is used to classify a new record whose feature value is missing. At a decision node, since an outcome is unknown, the algorithm discovers all possible paths which lead to leaves with different classes. A record is assigned to as the class with the highest probability.

Decision tree is grown fully in C4.5 algorithm and it continues to split until a node is completely pure. Anyhow, a node with too few records has very less prediction power. So, there is a trade-off between generalization and accuracy, an equilibrium between the pureness and the size of a node. With target of to scale down version of a tree in the last phase of tree construction, tree pruning decreases the complexity, not allows over-fitting, and improves the accuracy. There are two methods of constructing tree pruning: Theses two methods are pre-pruning and post-pruning. Pre-pruning is carried out while the tree is being induced. It avoids the formation of node which has too smaller number of records therefore it has insignificant generalization ability. The selection of the minimum allowable number of records per leaf define based on each data set, often found by experiments.

Post pruning method applies on a fully induced tree. After the full-size tree has been constructed, it will be subjected pruning backward. Each node in the tree is applying for pruning whether the node containing its sub-tree might be converted to one leaf. All records that consist to the sub-tree are transferred to the newly introduced leaf. From the training set a separated data set used for tree constructing called pruning set is prepared. If the pruned tree output predictions \as good as or better on the pruning set, then subcomponent below that level are removed. When a separated data set is applied to identify the classification error rate, it is named as reduced-error pruning.

Another approach contains in C4.5 named pessimistic pruning. Considered a leaf contains N records, E of them are misclassified so the identified error rate for this leaf is E/N . Assuming that the error rate in a binomial distribution, its expected value being in inside a confidence interval which is bounded by two limits. For a significance level, the upper limit $U_u(E, N)$ is denoted as an estimate of the error rate. Therefore, the expected number of records that are misclassified in a leaf is $e = N * U_u(E, N)$. And the estimate of the error rate of a tree is derived as follows.

$$e_t = \frac{\sum_{i=1}^l N_i * U_{\alpha}(E_i, N_i)}{\sum_{i=1}^l N_i}$$

2.7.3 Support Vector Machines

Author has conducted the research (Sindhu M E and Vijaya, 2015) on customer churn using well known data mining methodology CRISP-DM (cross-industry standard process for data mining) to identify network usage behaviors of the ISP (Internet service providers) subscribers in Taiwan. A system developed in (Sindhu M E and Vijaya, 2015) based on genetic programming methodology to predict churn and proposed the intelligent churn prediction concept to develop a new hybrid model to improve the performance and accuracy for churn prediction. In here used data set was consisted with 23 attributes and 750 customers in formation. Based on the result of different type of SVM predictions models. After applying the model following result were derived.

Classifier	PSVM	LSVM	ASVM
Accuracy (%)	100	85.5	86.25
Learning time (in secs)	0.34	0.43	0.54
Average number of SV	352	203	309

TABLE 3: COMPARISON OF ACCURACY, LEARNING TIME AND NUMBER OF SUPPORT VECTORS

based on its predictive accuracy performance were determined, learning time and number of support vectors. It was found to be PSVM based churn prediction model is the best performing churn model and was recommended for predicting whether the telecom customer is churning or non-churning.

[142] as per the research mobile telecommunications industry, the churn term, refer as customer attrition or churning of subscriber, refers to the phenomenon of loss of a customer [142-2]. It has used a dataset is from University of California, [142-13] and consist with 3333 subscribers with 21 variables each. IBM SPSS is used to develop the statistical model based on Support vector machines (SVM). It has used different kernel function of SVM such as Radial Basis Function kernel (RBF), Linear kernel (LIN), Polynomial kernel (POL), Sigmoid kernel (SIG). and POL has the best correct overall accuracy of 88.86% and RBF has second correct overall accuracy of 85.63%. they have concluded that

practically these three models (RBF, LIN, and POL) have a better performance (around 80%) in predicting churners too.

support vector machines (SVM) is consider as third algorithm to use for customer churn prediction. In here training sample d and data points $X_i, i = 1, 2, n$ and it should be divided the data set into j regions (j may be the class variable). The data is subjected segmentation by a set of hyper planes into j regions (Mirowski, 2008) .The data points that have smallest distance to the plane that divides the data into j sub regions known as support vectors. There will be large number of the hyper planes that segment the data into j sub region but real requirement is one hyper plane (line) that maximizes the region between the support vectors (J. Friedman, 2008). This is happened because maximizing the region between the support vectors minimize the likelihood of misclassifying new data points.

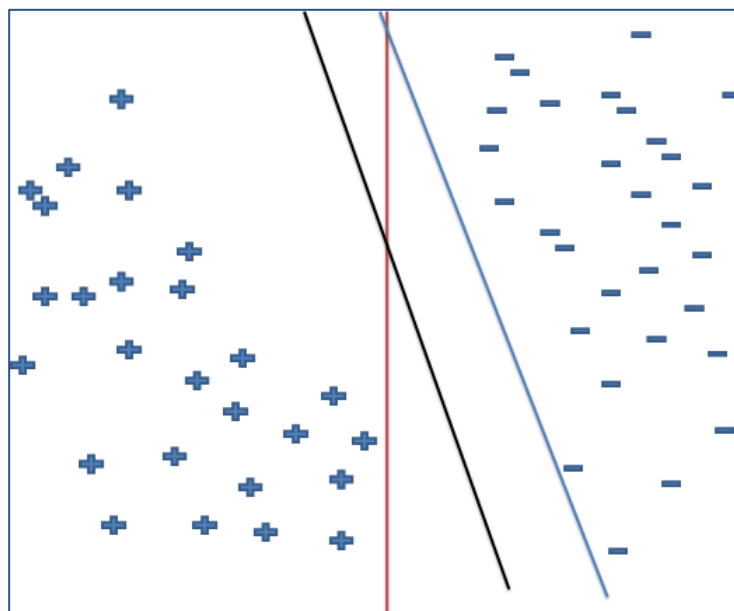


FIGURE 4:PLANE SEPARATING THE DATA POINT

Above figure presents three planes that can divide the positive and negative data points without any misclassification rate (Mirowski, 2008) The black plane separates point in the best way in the direction of bigger margin between the two groups of points. A bigger margin is the best expectation because there is a higher probability that if a new data point is imputed classified correctly. the best plane (line) that separates these points is an optimization problem which can be solved using Lagrangian methods. Sometimes these data points may not be linearly separable. If the training data set.

$$D = [(x^1, y^1), (x^2, y^2), \dots, (x^l, y^l)], \forall x \in R^n, y \in [-1, 1]$$

where x^L , $i = 1, 2, \dots, l$ is the vector of individual and variable, and y^L are regions of owing for each individual X^n . These points can be separated into 1 or -1 by a hyper plane $\langle w, x \rangle + b = 0$ where b is the distance from the point to the corresponding plane, w the weights vector and $\langle w, x \rangle$ is the dot product. The separating hyper plane must satisfy

$$y^i[\langle w, x^i \rangle + b] \geq 1, \forall i = 1, 2, \dots, l$$

and the distance of x to the hyper plane which is

$$d(w, b; x) = \frac{|\langle w, x \rangle + b|}{\|w\|}$$

The optimal hyper plane is the one that minimizes $\phi(w) = \frac{1}{2}\|w\|^2$ combining and forming a Lagrangian equation with alpha parameter it drives to finding a solution of

$$\phi(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^l \alpha_i (y^i[\langle w, x^i \rangle + b] - 1)$$

which complies the Karush Kuhn Tucker condition for an optimal value it is the first order condition. It finds the first partial derivative with respect to b and w and equal to zero for an optimal solution. The solution to the problem is then given by

constrained by $\alpha_i > 0$ and $\sum_{i=1}^l \alpha_j y_i = 0$

For an assumption hyper plane cannot be linearly separated the data and suppose now that there is an error then the $\psi_i, \forall i = 1, 2, \dots, l$ constraint equation can be updated to

$$y^i[\langle w, x^i \rangle + b] \geq 1 - \psi_i, \forall i = 1, 2, \dots, l$$

and the optimal plane is found by w that is minimum.

$$\phi(w, \alpha) = \frac{1}{2}\|w\|^2 - C \sum_{i=1}^l \psi_i$$

where given C subject to constrains. The Lagrangian equation now presents.

$$\phi(w, b, \alpha, \psi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \psi_i - \sum_{i=1}^l \alpha_i (y^i [w^T x^i + b] - 1 + \psi_i) - \sum_{j=1}^l \beta_j \psi_j$$

in here β and α are Lagrangian multipliers. The equation is solved in similar fashion applied earlier and the solution is presented by

$$\alpha^i = \underset{\alpha}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum_{k=1}^l \alpha_k$$

constrained by

$$0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^j \alpha_j y_i = 0$$

As per above optimization problem is solved and need to identify the hyper plane to be fitted. When fitting a SVM model kernel functions can be used to map the data into high dimension with the target of making the data more separable. There are few several kernel functions are available and here followings are considered:

- Radial Basis Function: $k(x, x') = \exp(-\sigma \|x - x'\|^2)$
- Polynomial: $k(x, x') = (\text{scale} \langle x, x' \rangle + K)^N$
- Hyperbolic Tangent: $k(x, x') = \tanh(\langle x, x' \rangle + K)$
- Laplace: $k(x, x') = \exp(-\sigma \|x - x'\|)$

Base on the data set choice of the kernel really depends. The data set determines the parameter choices of K, N (degree) and σ . Furthermore, in R Software tool if these parameters are not given, the program will select the best "parameter" values for you (Karatzoglou, 2013)

2.8 Model Performance evaluation

There is no algorithm is superior or faster to another on average over all domains (Almana, 2014) it can applied. One algorithm may work well in particular domains while the others

shows better performance in elsewhere. To build a classifier in in this study, learning algorithms are applied on training data set. Different learning algorithms performances of the classifiers are compared on this target data set (Gutkin, 2008). What are the parameters or attribute of defining a better classifier need to be assess here? Appropriate properties are generality confidence level and accuracy of prediction. Confusion denotes the four results which classifier is applied on a set of instances.

		Actual Class	
		p (+)	n (-)
Hypothesized class	p (+)	True Positive	False Positive
	n (-)	False Negative	True Negative
Column totals		P	N

FIGURE 5: THE CONFUSION MATRIX

A correctly classified if its actual class is positive or negative which instance is counted as a true positive (TP) or a true negative (TN) respectively. A positive instance which is misclassified as negative is considered as a false negative (FN). And a negative instance which is misclassified as positive is considered as a false positive (FP). The total number of positive instances in the data set is $P = TP + FN$, and the total number of negative instances is $N = TN + FP$. Based on a confusion matrix, the most commonly used for evaluation metrics are overall accuracy, true positive rate and false positive rate.

The overall accuracy (OA) is the ratio of the correctly classified instances.

$$OA = \frac{TP + TN}{P + N}$$

The hit rate or true positive rate is the ratio of positive instances that a classifier capture.

$$TP \text{ rate} = \frac{TP}{P}$$

And false alarm rate or the false positive rate is the ratio of negative instances that a classifier wrongly flagged as positive.

$$\text{FP rate} = \frac{\text{FP}}{N}$$

Receiver operating characteristics ROC graph is obtained by plotting TP rate is plotted as y against FP rate as x. All relevant classifier is represented by a point on ROC graph. A most accurate classifier is represented by point (0, 1) on ROC graph which classifies all positive and negative instances correctly with 100% TP rate and 0% FP rate. The diagonal line $y = x$ presents classification that is based completely on random guesses (Fawcett, 2006).

It can easily achieve the required TP rate but same time it also gains equally high FP rate which is not acceptable. churn prediction is to the main goal in this analysis. Therefore, an appropriate classifier is the one consists with higher TP rate and low FP rate given that churn is the positive class. This type of classifier is denoted at the upper left corner of ROC graph (Fawcett, 2006)..

A classifier gives output in probabilistic form $P(y = +1|x)$, the probability that an instance owns to the positive class. If this probability is greater than the predefined threshold $P(y = +1|x) > \theta$, an instance is group in positive class, otherwise negative class. A classifier which is using high value for θ is considered “conservative”. It classifies positive instances only with strong proof, so it makes few FP misclassification but at the same time has low TP rate. A classifier which is using low value for θ is considered “liberal”. It classifies positive instances with weak evidence so it achieve high TP rate but also consist with many FP mistakes (Fawcett, 2006) (Svendsen, 2013). When the classifier performance is plotted on ROC graph with value of θ varied from 0 to 1, a ROC curve will be plotted on it. It demonstrates the trade-off between TP rate and FP rate.

2.9 Research Gap and Conclusion

Under literature review chapter discussed about previous studies done by the researchers in customer churn prediction analysis evaluation and finding in telecommunication industry. statistical techniques such as Logistic regression and Naïve Bayes and classical machine learning techniques such as Decision Trees, K – Nearest Neighbour, Artificial

Neural Network and support Vector Machine in addition researchers were used by most of researcher.

When we consider the classifier, they can be combined or individually perform and produce the prediction results with different accuracies. But when there is lack of studies which derived all combination of (nPr) of target classifiers evaluated for the performance.

In this research work with Naïve bayes, C4.5 decision tree and SVM three classifiers will be evaluated for same data set to in all combination available.

Also, there are few Studies in relating to develop models in the credit card application process default prediction.

03. Chapter – Methodology

3.1 Chapter -introduction

In this chapter research is focusing on the research development methodology in different aspect. Cell2Cell data set is based for this research and it tis over 70,000 records contain dataset. In this chapter dataset analysis also a key objective.

3.2 Stakeholder Analysis

In this research and its outcomes are focus or related with different level of stakeholders. Project has direct impact and indirect impact to stakeholders and will be discussed over this subsection.

First tier stakeholders

Researcher: He will be gathering the technical skill academic qualification on this research

Supervisor: There will be one new research which is interested by him and conducted by him

University: There will be research on churn prediction and supervised learning area

Second tier stakeholders

Telecommunication Operators: they can get a proven way of analysis their gathered data and identify the customer behaviors and what are the factors or variables significantly contribute to customer churn.

Customers: Based on the actions of the telecommunication operator customer satisfactory and minimize the effort of shifting the network

Third tier stakeholders

Government: it will be reduced the ATL and BTL marketing investment on customer acquisition and churn management which will reduce the cost of operation and ultimate reduce the cost on telecommunication expenses and living cost of people.

Society locally and Globally: Less resource utilization for operation and save the environment which has positive effect to the world. And improve the quality of life of peoples and which drives the better living conditions.

3.3 Requirement Gathering Techniques

3.3.1 Data Gathering Technique

For this research I have used Cell2Cell data set from Kaggle data set repository. So, this is secondhand dataset used to develop the research and based on the generalization concept it is suggested that this an appropriate selection.

Cell2Cell is the 6th largest mobile communication service provider in the US, with 10 million subscribers approximately. In the Sri Lankan context, all 4 telecom operators have lesser than 10 million subscribers as upper bound and all are above 2.5 million in lower bound. This seems to be a justifiable size of telecommunication supplier with respect Sri Lanka.

Secondly all telecommunication operators maintain demographical data of customer, operational data, transactional data, service-related data, and payment related data. So, this data set also has same type of data. Due to same type of attribute this can be localized to Sri Lanka telecommunication operator context too.

3.4 Project Management Methodology

A hybrid of both Agile and Waterfall was used to manage the project. Hybrid of both agile and waterfall project management methodology is perhaps the most suitable management technique for multidisciplinary research projects.

Agile methodology is an iterative process, where large amount of work is divided into small chunks called sprints. Whereas waterfall model follows a different approach, it emphasizes a sequential movement of a project with a defined set of phases. A strict waterfall method would hardly be a suitable approach for a research project due to the changing nature of requirements in the beginning as the requirements are to be learned and researched and may not even be correct. A strict agile project on the other hand would emphasize to abide by the agile manifesto which specifies to prioritize on individuals and interactions over processes and tools, working software over comprehensive documentation, customer collaboration over contract negotiation, responding to change over following a plan. Not all these 4 principals can be accounted for in a research project

where documentation is utmost important and customer collaboration and contract negotiation being business centric. A hybrid model was therefore chosen to manage the project where the work is divided into sprints where each sprint acts as a mini waterfall. Requirements gathering, designing, implementation and evaluation is practiced for small chunks of work within a sprint. This process will continue iteratively. The author may not be familiar with the concepts and terminologies used in such subjects therefore it is easier to learn and develop incrementally. A product backlog is maintained to identify the tasks and user stories that are required to be completed. The backlog will incrementally grow with time as and when requirements are clarified. These backlog items will be assigned to sprints as tasks or user stories for requirements gathering, designing, implementation and evaluation. The use of sprints and backlogging will show the progress of the research where what is done and what is yet to be done for weather prediction research and development can be identified.

3.5 Anticipated Risks and Mitigations.

1	Lack of domain knowledge	Domain	Get assistance local telecommunication operators (Hutch, Dialog) get the domain knowledge.
2	To find an open source dataset Kegalle dataset.	Development	To obtain a dataset which is a correct representation of telecom data set and could be generalized to local context
3	Identifying the correct predictive model/models	Development	Verify with the standard predictive models used for telecommunication churn prediction and combining these models make an ensemble model.

TABLE 4: RISKS AND MITIGATIONS.

3.6 Data set analysis

In this study use open source data set that published in Kaggle Machine learning repository that is Cell2cell is the 6th largest wireless company in the US, Cell2cell dataset consists of 71,047 signifying whether the customer had left the company two months after observation and 57 attributes.

Data set Characteristics	Multivariate
Attribute Characteristics	Integer, real
Associated Tasks	Classification
Number of Instances	71,047
Number of Attributes	67
Area	Business

TABLE 5:CELL2CELL DATA SET OVERVIEW

Cell2Cell dataset can be divided in to following categories as per each attribute characteristics.

Customer care Service

In here all the attributes are related to customer service 5 numeric and 4 Boolean type variables are found as per below table.

Customer care Service		
Attribute	Description	Type
blckvce	Mean number of blocked voice calls	Numeric
custcare	Mean number of customer care calls	Numeric
dropblk	Mean number of dropped or blocked calls	Numeric
mailres	Responds to mail offers	Boolean
mailflag	Has chosen not to be solicited by mail	Boolean
retcalls	Number of calls previously made to retention team	Numeric
incmiss	Income data is missing	Boolean
income	Income (0=>missing)	Numeric
retcall	Customer has made call to retention team	Boolean

TABLE 6:CUSTOMER CARE SERVICE ATTRIBUTES IN CELL2CELL DATASET

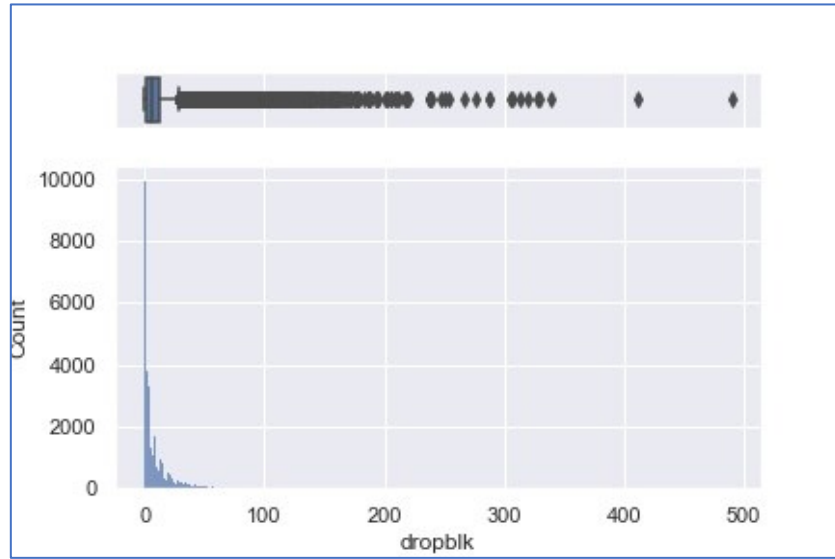


FIGURE 6:MEAN NUMBER OF DROPPED OR BLOCKED CALLS

Like above box lot and histogram, I have examined all numerical and Boolean variable visually. Which explain its distribution in the histogram and box plot show the span of distribution and outliers. In her it is right skew distribution with outlies and long tails

<i>Dropblk</i>	
Mean	10.15
Mode	0.00
Standard Deviation	15.46
Kurtosis	69.20
Skewness	5.69
Range	489.67
Minimum	0.00
Maximum	489.67
Count	71047.00

TABLE 7:DROPLK MEAN NUMBER OF DROPPED OR BLOCKED CALLS

As per the above summary statistics of Dropblk (Mean number of dropped or blocked calls) has mean value of 10.15 and standard deviation of 15.46. but it maximum is 489.67 and it is over 31.015 times of standard deviation from mean value. It has long right tail and under data cleaning it will be processed to remove outliers. In Appendix 01 it will describe all descriptive statistics and visual inspection of all variables under Customer care Service.

Generalization of customer service data:

Every telecommunication operator has same set of customer service data which can be applied same way with their own data set. It will predict the results based on own data.

Customer Demography

In here all the attributes are related to customer service 10 numeric and 26 Boolean type variables are found as per below table.

Customer Demography		
Attribute	Description	Type
months	Months in Service	Numeric
uniqsubs	Number of Unique Subs	Numeric
actvsubs	Number of Active Subs	Numeric
phones	# Handsets Issued	Numeric
models	# Models Issued	Numeric
eqpdays	Number of days of the current equipment	Numeric
age1	Age of first HH member	Numeric
age2	Age of second HH member	Numeric
children	Presence of children in HH	Boolean
prizmrur	Prizm code is rural	Boolean
prizmub	Prizm code is suburban	Boolean
prizmtwn	Prizm code is town	Boolean
refurb	Handset is refurbished	Boolean
webcap	Handset is web capable	Boolean
truck	Subscriber owns a truck	Boolean
rv	Subscriber owns a recreational vehicle	Boolean
occprof	Occupation - professional	Boolean
occcler	Occupation - clerical	Boolean
occcrft	Occupation - crafts	Boolean
occstud	Occupation - student	Boolean
occhmkr	Occupation - homemaker	Boolean
occret	Occupation - retired	Boolean

occself	Occupation - self-employed	Boolean
ownrent	Home ownership is missing	Boolean
marryun	Marital status unknown	Boolean
marryyes	Married	Boolean
mailord	Buys via mail order	Boolean
travel	Has traveled to non-US country	Boolean
pcown	Owens a personal computer	Boolean
credited	Possesses a credit card	Boolean
newcelly	Known to be a new cell phone user	Boolean
newcelln	Known not to be a new cell phone user	Boolean
refer	Number of referrals made by subscriber	Numeric
mcycle	Owens a motorcycle	Boolean
setpre	Missing data on handset price	Boolean
setprc	Handset price (0=>missing)	Numeric

TABLE 8: CUSTOMER DEMOGRAPHY DATA

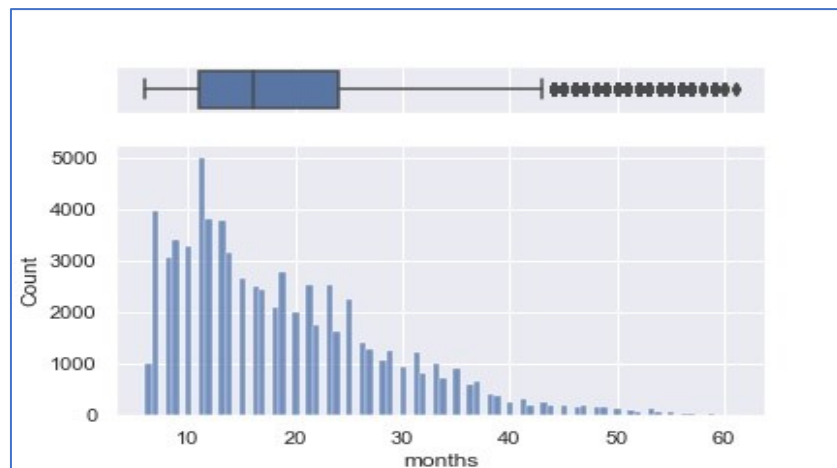


FIGURE 7: MONTH DISTRIBUTION

Like above box lot and histogram, I have examined all numerical and Boolean variable visually. Which explain its distribution in the histogram and box plot show the span of distribution and outliers. In her it is right skew distribution with outlies and long tail.

<i>onths</i>	
Mean	18.75
Standard Deviation	9.79
Kurtosis	0.90
Skewness	1.06
Range	55.00
Minimum	6.00
Maximum	61.00
Count	71047

TABLE 9: MONTHS DISTRIBUTION

As per the above summary statistics of Months has mean value of 18.75 and standard deviation of 9.79 but its maximum is 61.00 and it is over 4 times of standard deviation from mean value. It has long right tail and under data cleaning it will be processed to remove outliers. In Appendix 01 it will describe all descriptive statistics and visual inspection of all variables under Customer demography analysis.

Customer Credit Score

In here all the attributes are related to Customer Credit Score 2 Boolean type variables are found as per below table.

Customer Credit Score		
Attribute	Description	Type
credita	Highest credit rating	Boolean
creditaa	High credit rating	Boolean

TABLE 10: CUSTOMER CREDIT SCORE

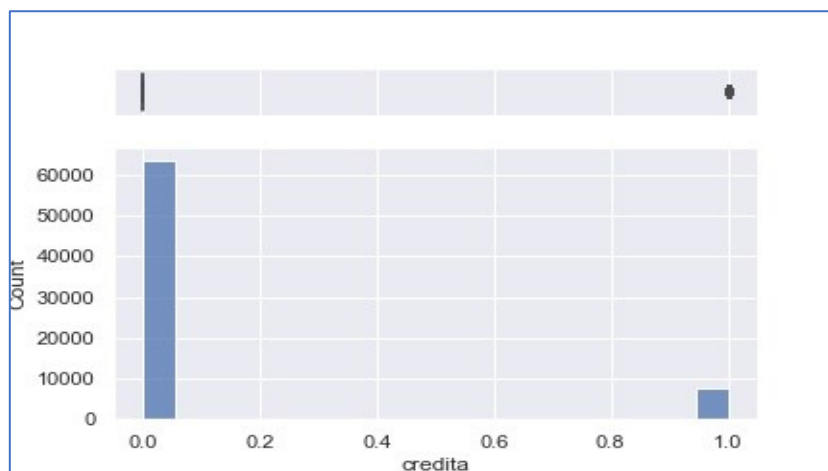


FIGURE 8: HIGHEST CREDIT RATING

Like above box lot and histogram, I have examined two Boolean variables visually. In here as per Histogram occurrence of “0 is mor probable than “1”.

High credit rating	
Mode	0
0-count	61919
1-count	9128
Minimum	0
Maximum	1
Count	71047

TABLE 11: HIGH CREDIT RATING

Mode of the High credit rating is “0” and it has 61,919 occurrence which is over 6 times higher with respect to “1” occurrence 9,128.

Bill & Payment Analysis

In here all the attributes are related to Bill & Payment Analysis3 numeric type variables are found as per below table.

Bill & Payment Analysis		
Attribute	Description	Type
revenue	Mean monthly revenue	Numeric
recchrge	Mean total recurring charge	Numeric
changer	% Change in revenues	Numeric
Customer Usage Pattern (Behaviour pattern)		

TABLE 12: BILL & PAYMENT ANALYSIS IN CELL2CELL DATA SET

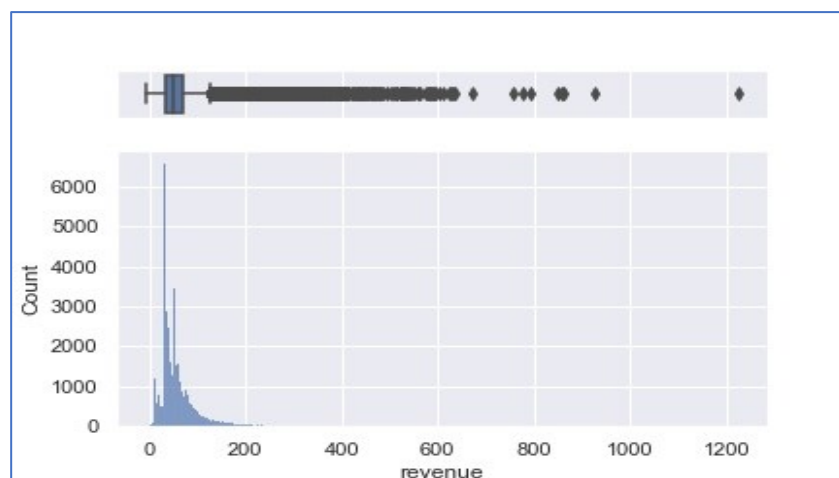


FIGURE 9: MEAN MONTHLY REVENUE.

Like above box lot and histogram, I have examined all numerical variable visually. Which explain its distribution in the histogram and box plot show the span of distribution and outliers. In her it is right skew distribution with outlies and long tail

<i>Revenue</i>	
Mean	58.85
Standard Deviation	44.24
Kurtosis	35.94
Skewness	3.97
Range	1229.55
Minimum	-6.17
Maximum	1223.38
Count	70831

TABLE 13:SUMMARY STATISTIC REVENUE

As per the above summary statistics of Revenue has mean value of 58.85 and standard deviation of 44.24 but it maximum is 1229.55 and it is over 30 times of standard deviation from mean value. It has long right tail and under data cleaning it will be processed to remove outliers. In Appendix 01 it will describe all descriptive statistics and visual inspection of all variables under bill and payment analysis.

Customer Usage Pattern (Behaviors pattern)

In here all the attributes are related to Customer Usage Pattern (Behaviors pattern) 9 numeric type variables are found as per below table.

Customer Usage Pattern (Behavior pattern)		
Attribute	Description	Type
mou	Mean monthly minutes of use	Numeric
overage	Mean overage minutes of use	Numeric
roam	Mean number of roaming calls	Numeric
changem	% Change in minutes of use	Numeric
dropvce	Mean number of dropped voice calls	Numeric
unansvce	Mean number of unanswered voice calls	Numeric
outcalls	Mean number of outbound voice calls	Numeric
peakvce	Mean number of in and out peak voice calls	Numeric
opeakvce	Mean number of in and out off-peak voice calls	Numeric

TABLE 14:CUSTOMER USAGE PATTERN (BEHAVIOR PATTERN)

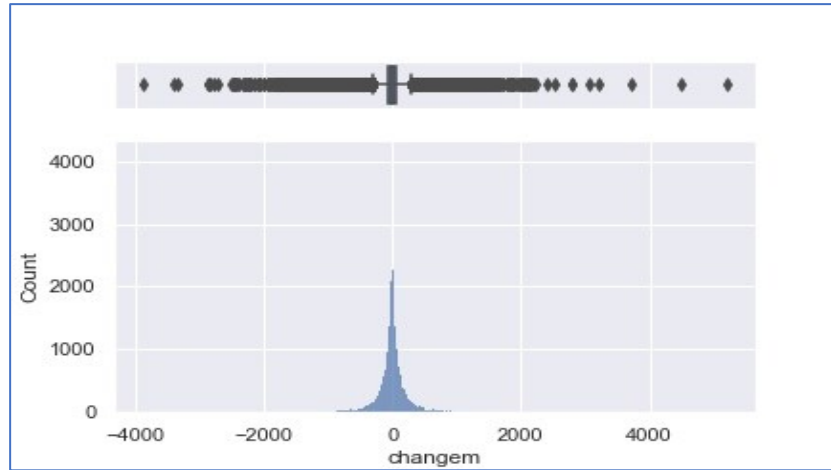


FIGURE 10: CHARHEM DISTRIBUTION

Like above box lot and histogram, I have examined all numerical variable visually. Which explain its distribution in the histogram and box plot show the span of distribution and outliers. It has symmetric distribution.

<i>Changem</i>	
Mean	-10.85
Standard Deviation	255.31
Kurtosis	19.30
Skewness	-0.26
Minimum	-3875.00
Maximum	5192.25
Count	70545

TABLE 15:SUMMARY STATISTICS CHARGEM

As per the above summary statistics of Revenue has mean value of -10.85 and standard deviation of 255.31 but it maximum is 5192.25 and minimum -3875.00 and it is over 12 times of standard deviation from mean value to both sides. It has shown symmetric distribution and under data cleaning it will be processed to remove outliers. In Appendix 01 it will describe all descriptive statistics and visual inspection of all variables under Customer Usage Pattern (Behavior pattern) analysis

Value added services

In here all the attributes are related Value-added services 5 numeric and 3 Boolean type variables are found as per below table.

Value added services		
Attribute	Description	Type
directas	Mean number of directors assisted calls	Numeric
threeway	Mean number of three-way calls	Numeric
mourec	Mean unrounded mou received voice calls	Numeric
callfwdv	Mean number of calls forwarding calls	Numeric
callwait	Mean number of calls waiting calls	Numeric
retcalls	Number of calls previously made to retention team	Numeric
retacct	Number of previous retentions offers accepted	Numeric

TABLE 16: VALUE ADDED SERVICES IN CELL2CELL DATA SET

Read datasets

Data set is transfer to *.csv (comma separated version) and import and read using python panda's library.

Raw data set discrete statistics.

Prior to process/clean the data set it was examining for missing values, outliers, and descriptive statistics of each variable.

Dependent variable (Y) Churn

Churn is a Boolean type of variable. And "1" represent the chum customer and "0" represent retain customers. In here churn divided in two classes and its descriptive statistics as follows.

<i>Churn</i>	
Mode	0
0-count	50438
1- count	20609
Count	71047

TABLE 17: CHURN DEPENDENT VARIABLE

In here majority of customer retain in the network under the "0" count of 50,438 (70.99%) and churned customers in "1" count of 20,609 (29,00 %). Based on this number of records and percentages this data set is dependent variable is good enough to supervise learning with clusters.

Independent Variables (X's)

There are 2 types of variable exist under independent variable named numerical (34) and Boolean (33). in the data cleaning and processing part it is necessary to apply appropriate

techniques to each category. In Appendix 01 has detail description on descriptive summary statistics of used variable for this research.

Application Framework

Based on the literature there are many academic works have been conducted covering many areas on churn prediction, impact on customer satisfaction, customer loyalty and customer retention. Also, they have identified the different accuracy level of different classifiers.

In this research study is focus on the. Naïve Bayes,. Decision Trees, .Support Vector Machines . Based all permutation individual and ensembled classifier performance analysis using statistical method.

3.6.1 Application Framework model

In here researcher has modeled the application flow diagram which will be used for this research work. First step is importing the data set into Python Pandas library then visualizing the data set for each variable/ attribute. Boxplot and histogram method can be used, and it will be presented the distribution of the variable and outliers' availability.

Then Raw data set will be subject to preprocessing with changing the categorical in to numerical (binary0 format and examine the data set missing values and outlier under each variable. The outliers and missing value will be replaced with respective mean imputation. Progress of the visualization can be evaluating with descriptive statistics measurements. Visualization method such as histogram and boxplot can be used as graphical evaluation of processed data set.

This Cell2Cell data set has 57 attributes 55 as independent and 2 as dependent variables with respect to this churn prediction research study. The contribution or correlation of these independent variable and dependent variables may have different correlation and covariance. Due to these 55 variables will be process under feature selection algorithms and select the most appropriate variables which determine the dependent variable or customer churn. Using the independent variable similarities duplicated attributes will be removed from the data set. This feature selection process will minimize the weight of the data set and improve the processing power and accuracy of the processing. It is necessary to do the t avoid the data set biasness toward one direction.

Then preprocess data set will be divided in to 80% as training data set and 20% as testing or validation data set. Each predictive classifier will be trained using this training data set and its progress will be evaluate and validate using testing dataset. based on each classifiers

performance (accuracy) it will determine to accept the results or restart the process from the feature selection onwards. This will be iterative process and once it reaches the satisfactory level of accuracy record the data for further statistical analysis.

Application Framework model

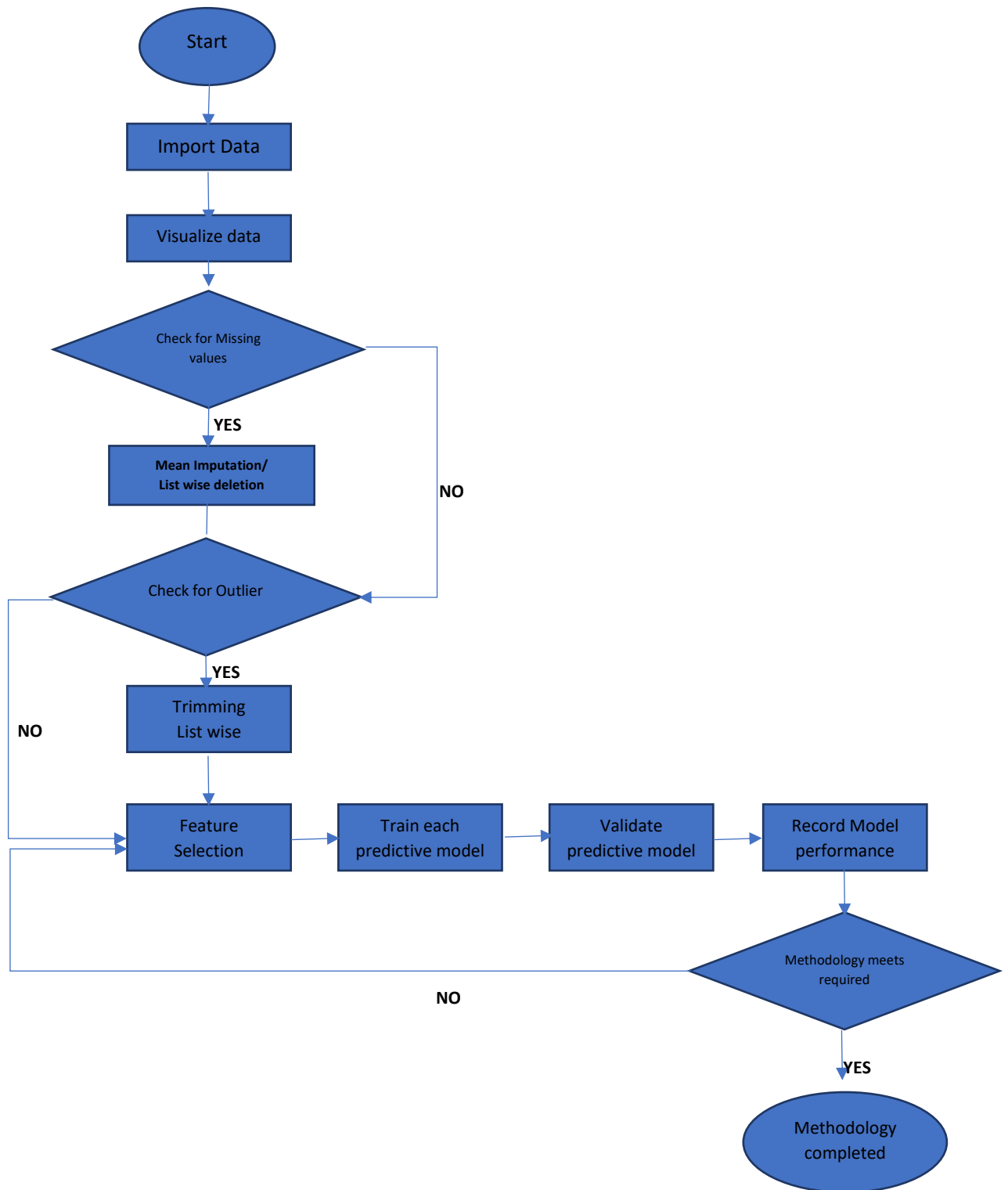


FIGURE 11: APPLICATION FRAMEWORK MODEL

Framework -01(Identify the classifier or two classifiers combination which predict the customer churn more accurately.

Single classifiers accuracy and two classifiers accuracy research framework can be presented as follows.



FIGURE 12: CLASSIFIERS (1,2)

Each Classifier will be evaluated based on below matrix separately.

NB Accuracy		Feature											
		1	2	3	4	5	6	7	8	9	10		
OutLier	0.20												
	0.30												
	0.40												
	0.50												
	0.60												
	0.70												
	0.80												
	0.90												
	1.00												
	1.10												
	1.20												
	1.30												
	1.40												
	1.50												

FIGURE 13: NB CLASSIFIES EVALUATION LIST WISE DELETION

Naive Bayer’s Classifies accuracy

- (Missing value) X1 in Listwise deletion/ mean, mode imputation, Accuracy Y
- (Outlier) X2 in range 0.2 to 1.5 with 0.1 increment, Accuracy Y
- (Feature selection) X3 in range 1to 10 with 1 increment Accuracy Y

Decision Tree Classifies accuracy

(Missing value) X1 in Listwise deletion/ mean, mode imputation,	Accuracy Y
(Outlier) X2 in range 0.2 to 1.5 with 0.1 increment,	Accuracy Y
(Feature selection) X3 in range 1to 10 with 1 increment	Accuracy Y

Support vector Machine Classifies accuracy

(Missing value) X1 in Listwise deletion/ mean, mode imputation,	Accuracy Y
(Outlier) X2 in range 0.2 to 1.5 with 0.1 increment,	Accuracy Y
(Feature selection) X3 in range 1to 10 with 1 increment	Accuracy Y

Na Machine Classifies and Decision Tree Classifies accuracy

(Missing value) X1 in Listwise deletion/ mean, mode imputation,	Accuracy Y
(Outlier) X2 in range 0.2 to 1.5 with 0.1 increment,	Accuracy Y
(Feature selection) X3 in range 1to 10 with 1 increment	Accuracy Y

Decision Tree Classifies and Support vector Machine Classifies accuracy

(Missing value) X1 in Listwise deletion/ mean, mode imputation,	Accuracy Y
(Outlier) X2 in range 0.2 to 1.5 with 0.1 increment,	Accuracy Y
(Feature selection) X3 in range 1to 10 with 1 increment	Accuracy Y

Support vector Machine & Naive Bayer's Classifies accuracy

(Missing value) X1 in Listwise deletion/ mean, mode imputation,	Accuracy Y
(Outlier) X2 in range 0.2 to 1.5 with 0.1 increment,	Accuracy Y
(Feature selection) X3 in range 1to 10 with 1 increment	Accuracy Y

3.7Data set visualization

Raw data set of Cell2Cell has 57 attributes and over 70,000 records and it is necessary to get the insight of all the variables prior to preprocessing or the analysis. This is the very first part and all numerical variable can be visualized using boxplot and histogram and categorical variable can be visualized using histogram. This will provide distribution pattern outliers existence and visual idea of point estimation and interval estimation.

01. Box Plot

All numerical attribute will be visualized using Box plot and output will provide the mean value, Standard deviation, inter quartile range, total range, and outlier information.

02. histogram

All attribute will be visualized using histogram and output will provide the distribution patterns, tailers of distribution (outliers), kurtosis and skewness. Refer Appendix 01 for the visualized Box Plot and histogram.

3.8 Data set preparation

Data preparation is the one of key task in data analysis and predictive modeling as a second step. In this processing part will be subject to different areas such as data cleaning or cleansing, how to manage missing data and outliers, feather selection will be some of them.as per the attribute type all need to transfer into numerical format especially categorical variables transform into binary values.

This is the one of hard part of the entire research work. It consumes more time and correct method should be used to improve the accuracy and not to bias the decision into one direction. In previous step visualization.

3.8.1 Missing values and outliers

This Cell2 Cell data set missing values will be process under two methods as follows. These two methods will transfer the dataset into processed data set.

01. mean imputation for missing value.

02. list wise deletion for missing value

Mean value imputation will not be impact on other variables reading due to missing value in that records. So, there is less impact on the biasness. If the missing value come as MCAR or MAR this does not have impact due to randomness properties. But listwise deletion reduce the number of instance and results can be biased or inaccurate. This Cell2 Cell data set outliers will be process under two methods as follows. These two methods will transfer the dataset into processed data set.

03. Trimming list wise deletion for outliers

Mean value imputation will not be impact on other variables reading due to missing value in that records. So, there is less impact on the biasness. If outliers come as random this imputation does not have impact due to randomness properties. But listwise deletion reduce the number of instance and results can be biased or inaccurate.

3.9 Feature selection

In this step select the most relevant features and building the classification model. All the independent attribute and dependent attribute correlation and covariance can be used to identify the impact and relationship. Based on that most correlated variables are selected. Filter method will be used with less proven to over fitting and much faster compare to wrapper method.

01. Chi-square method
02. Anova

Are some of filter method to select the variables.

3.10 Output data format data recording and evaluation methodology

Following environment variable or status were set before starting the machine learning experiment.

Preconditions:

Missing data: mean imputation

Outliers data: mean imputation

Feature selection: filter method (Chi-square)

Process data set:

80% of dataset (56,838) used as training data set with selected attribute set and dependent label variable and 20% of dataset (14,209) used as testing data set to validate and get the accuracy of prediction models.

Output

Following classifiers will be used and recorded the accuracy of each based on

		Actual Class	
		p (+)	n (-)
Hypothesized class	p (+)	True Positive	False Positive
	n (-)	False Negative	True Negative
Column totals		P	N

FIGURE 14: THE CONFUSION MATRIX

The overall accuracy (OA) is the ratio of the correctly classified instances.

Accuracy of Naïve Bayes classifier (NB)

Accuracy of Decision Trees classifier (DT)

Accuracy of Support Vector Machines classifier (SVM)

When two classifiers (X,Y)are used in Sequentially it will denoted following sequential order ($X_a \gg Y_a$).And when two classifiers (X, Y) are used in parallely it will denoted ($X_a + Y_a$).

All classifiers combination used in the experiment with sequential approach.

Classifiers	Total accuracy in test (X1, X2)				
	1	2	999	1000
01. Classifier 01: (NB) classifier					
02. Classifier 02: (DT) classifier					
03. Classifier 03: (SVM) classifier					
04. Classifier 04: (NB & DT) classifier					
05. Classifier 04: (DT & SVM) classifier					
06. Classifier 05: (SVM & NB) classifier					

TABLE 18: ALL COMBINATION OF CLASSIFIERS

3.11 Data analysis methodology.

Experimental data is recorded for 15 classifiers and 1,000 instances. Based on the statical analysis it will be select the followings.

01. What will best individual classifier out of Naïve bayes, Decision tree and SVM for churn prediction in telco domain with confident interval
02. What will best individual ensembled classifier with combination of two classifiers out of Naïve bayes, Decision tree and SVM for churn prediction in telco domain with confident interval
03. What will best individual ensembled classifier with combination of two classifiers out of Naïve bayes, Decision tree and SVM for churn prediction in telco domain with confident interval
04. What is the best classifier out of all 15 classifiers subjected researcher based on accuracy and confidence interval?
05. What are the correspondence attributes which derive the best classifier?

3.12 Role of the researcher

Researcher expects to broaden mobile telecommunication understanding regarding the customers churn and help them to protect their existing customer base and redefine who are the loyal customers and how to convert average customer to loyal customer. Specially, researcher expects to use the findings of this study to guide Sri Lankan mobile telecommunication industry to sustain their revenue and size of the customer base.

3.13 Validity

Finding of the research will help to understand variables or attribute cause to drive the churn and improve the efficiency and effectiveness of the how to avoid the churn in prediction basis.

3.14 Generalizability

This study used publicly available data set for preparing the proposed model but attributes in the selected data set almost similar to the attributes in relating to churn prediction in Sri Lankan mobile telecommunication industry after preparing the model using the selected data set can apply it into data in relating to Sri Lankan mobile telecommunication operators.

3.15 Installation methodology

This program is developed in client server architecture with administration or configuration console.

01. This can be used to generate the statistical dataset running number of times (user define sample size) for input independent variables recording output. Then output will be produced as excel format in the server end which can be used for analysis of classifier with respect to data set.
02. Using Configuration console initially classifiers are trained using the 80:20 size dataset and then validate with 20% of test dataset. Once configuration side is fixed client end user can elevate new records churn prediction accuracy parameters.

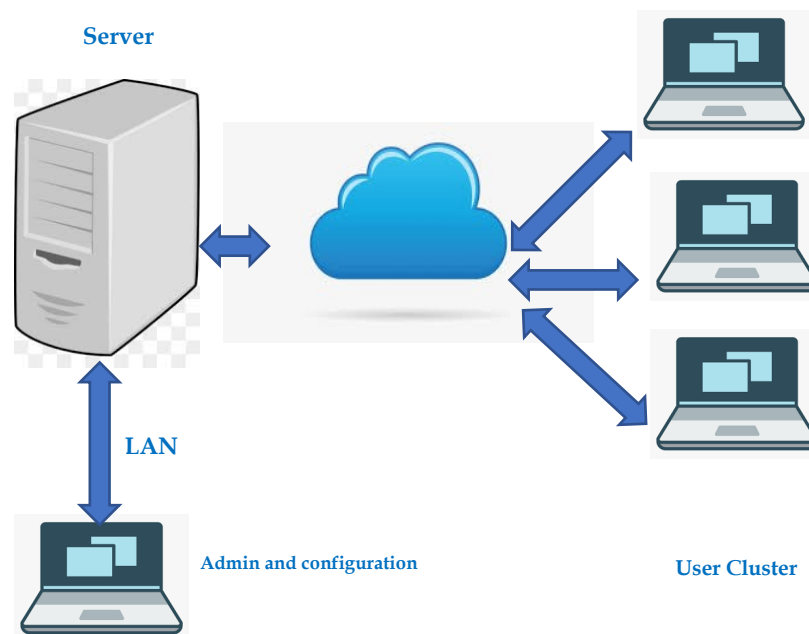


FIGURE 15: INSTALLATION DIAGRAM

3.16 Configuration and generalization of use

It is necessary to train the selected classifiers and evaluate the performance or accuracy using the test data set before it uses. In following window user can select of enter the missing Value, outlier ratio, feature selection variable and necessary classifiers in the configuration console.

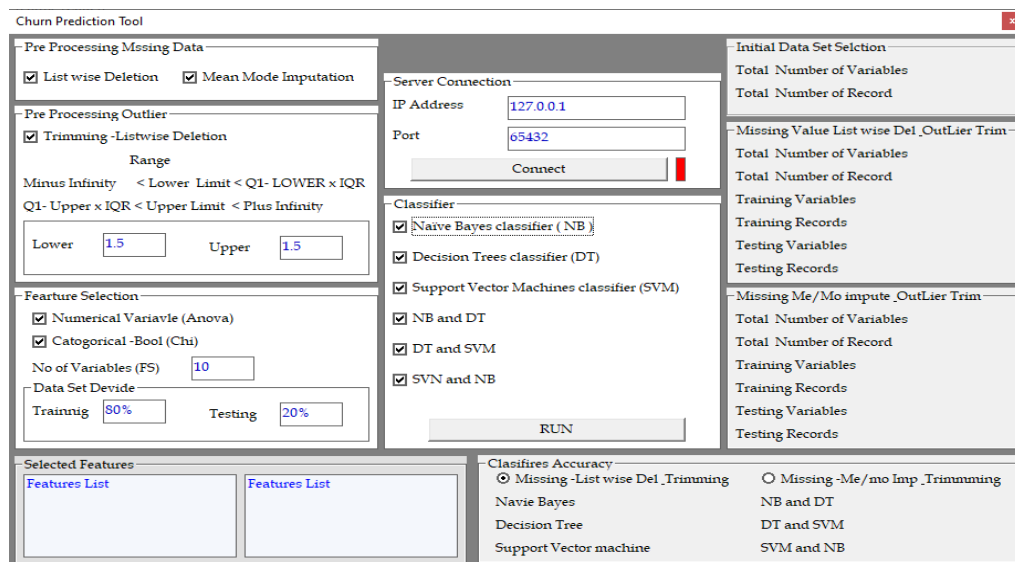


FIGURE 16: CONFIGURATION CONSOLE INITIAL APPEARANCE

After connecting with server where the data set and processing is happening green light start blinking as below.

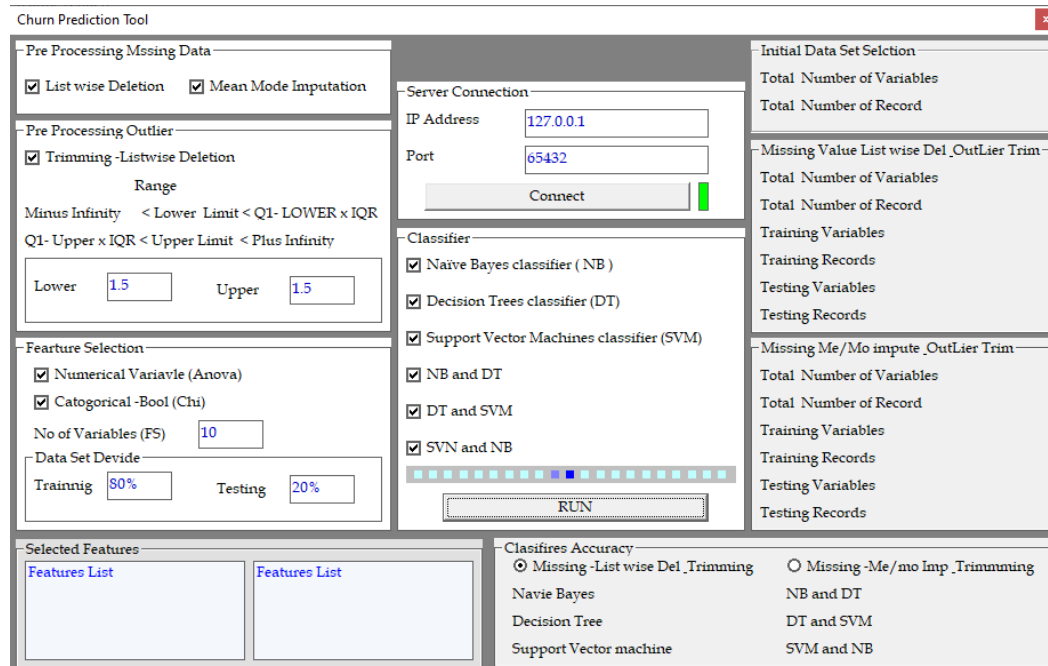


FIGURE 17: CONFIGURATION CONSOLE CONNECT WITH SERVER

Input the independence variables and classifier requirement and send the dataset to sever via TCP connected API an s server end catch it process and return the results. It can be appearing as below in the configuration console.

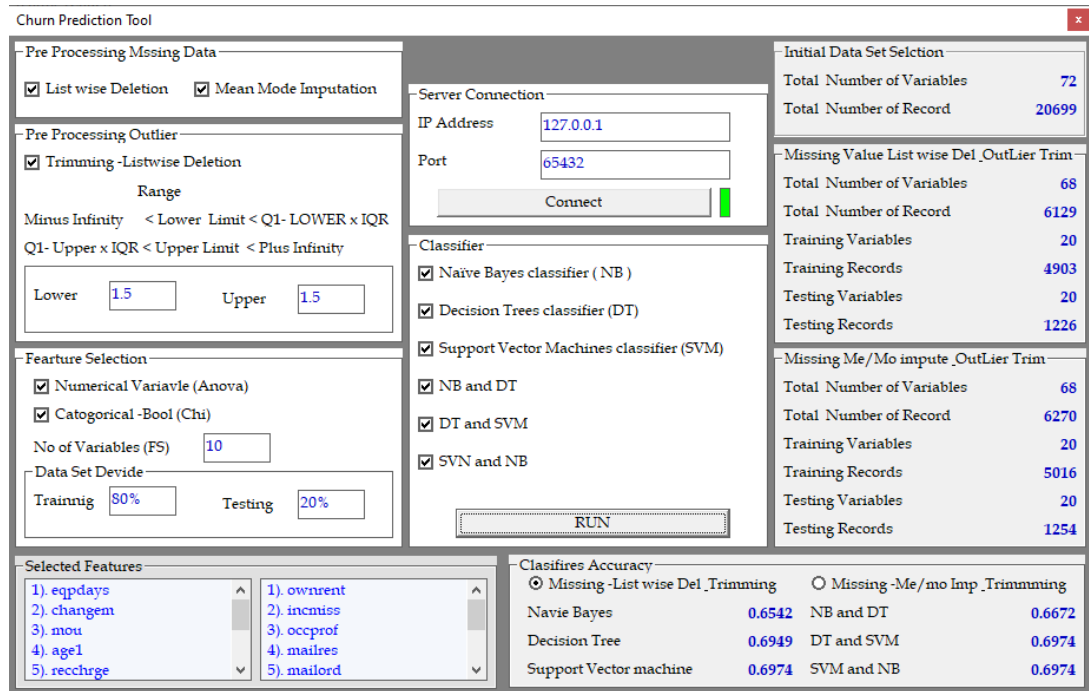


FIGURE 18: CONFIGURATION CONSOLE WITHOUT PUT ACCURACY VALUES

3.17 User interface and generalization of use

Login interface for the general user will be like shown as follows with red color conection

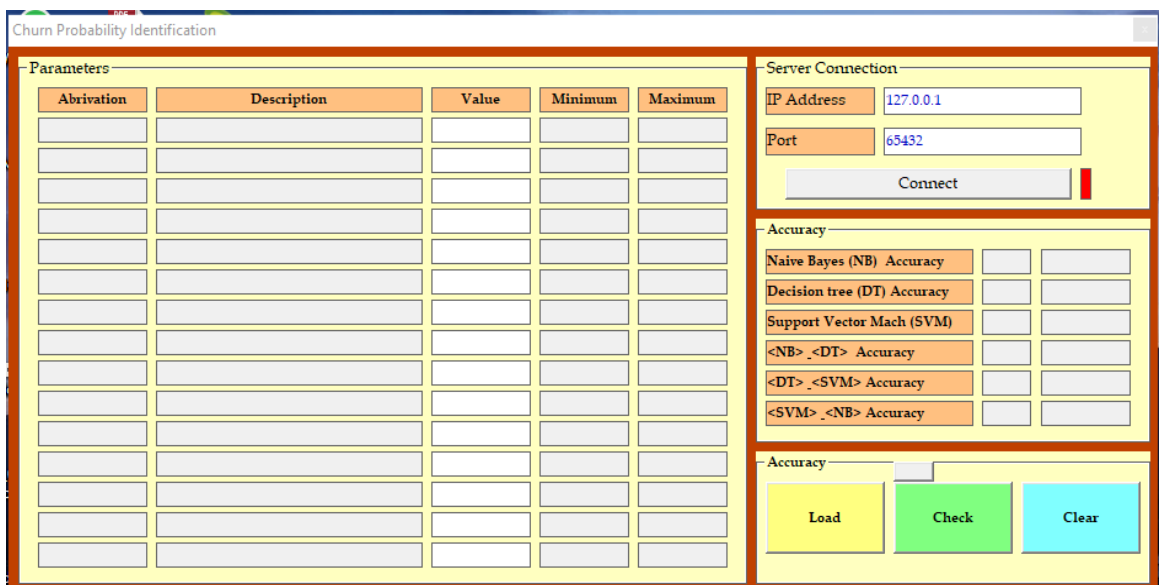


FIGURE 19:: USER INTERFACE DEFAULT VIEW

When User add the IP and port and connect to the server where the data set application is running connectivity will be shown in green color.

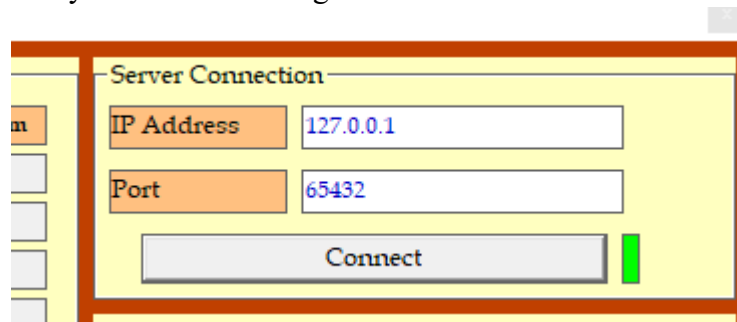


FIGURE 20:: USER INTERFACE AFTER CONNECT

FIGURE 21:TEST

User can send load request to the application server and get the required attribute and based on the minimum maximum value user can get the idea on preproperate range. Minimum and maximum is 1.5 times of Standard deviation from mean.

Then user can fill the attribute and get the prediction as shown in below picture.

Abrivation	Description	Value	Minimum	Maximum
eqpdays	Number of days of the current	250	-5.0	1812.0
changem	Presentage Change in minutes of	-15	-3331.25	5192.25
mou	Mean monthly minutes of use	34	0.0	6336.25
age1	Age of first HH member	37	0.0	99.0
recchrg	Mean total recurring charge	30	-11.29	359.93
months	Months in Service	0	6.0	60.0
age2	Age of second HH member	41	0.0	99.0
ownrent	Home ownership is missing	0	0.0	1.0
incnuss	Income data is missing	0	0.0	1.0
occpfrof	Occupation - professional		0.0	1.0
mailres	Responds to mail offers	0	0.0	1.0
mailord	Buys via mail order	0	0.0	1.0
marryun	Marital status unknown	0	0.0	1.0
travel	Has traveled to non-US country	0	0.0	1.0

FIGURE 22: USER INTERFACE AFTER LOAD VIEW

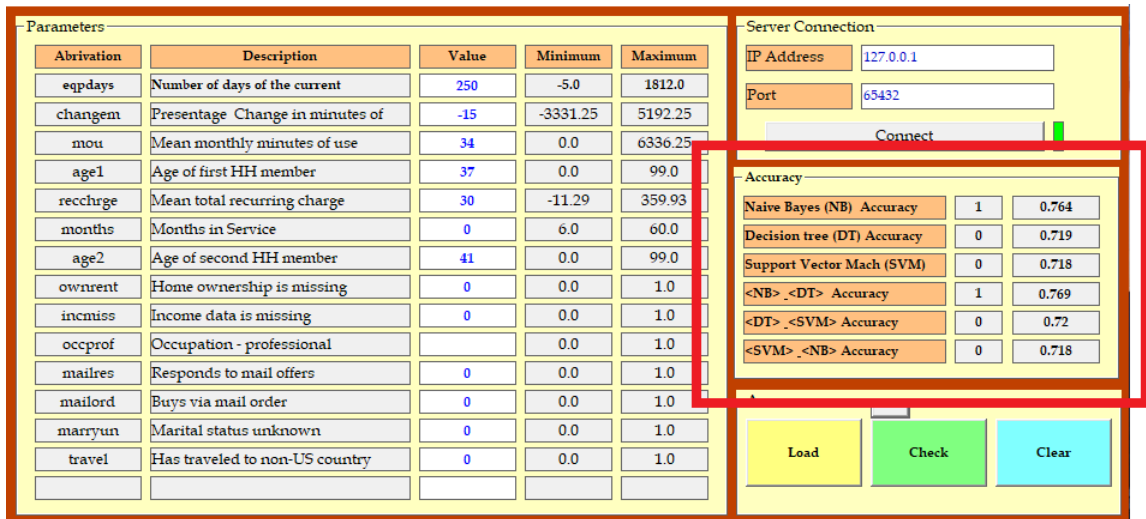


FIGURE 23::: USER INTERFACE GET ACCURACY

3.18 Legal, Ethical, Professional and Social Issues (LEPSI)

3.18.1 Legal

The project does not require the use of any licensed software to run as it is an open-source solution python. This dataset is opensource and is readily available to anybody thereby no personal data related to any-other entity is used nor stored. Furthermore, materials or research papers protected by copyright will not be used. All research papers and articles referenced will be from open source projects

3.18.2 Ethical

In this research is based on open source document and resources and all the research work were conducted as per the social ethical way. This research also subject to all the guidelines of UCSC of university of Colombo Sri Lanka. Ethical consideration upholds following to the University guidelines of university of Colombo and UCSC, disclosing resources and using Harvard reference system

3.18.3 Social

The project may have social implications because predictions will be used for various decision making like marketing and promotion in telecommunication domain.

04. Chapter – Results and Analysis

4.1 Chapter -introduction

This chapter presenting results of the research and its analysis in a descriptive approach. There are 3 different classifiers and its combinations is considered. So, result and analysis will be in a flow of main 6 categories based on the classifies.

Each classifies were tested in with following conditions

01. Missing value – List wise deletion (constantan way)
02. Outliers – list wise deletion based on span from the mean in terms of standers deviation. (variable X SD)
03. Feature selection based on number of features (variable No of features use)

It is necessary to analysis 3 subcategory under each and finally need to summarize all. In this research output is the accuracy of each classifiers and input are the above two variables.

4.2 Row data Set analysis

As it was introduced in the chapter 3 methodology row data set had 67 variables and it can divide in in 34 numerical and 33 categorical variables. Row data set can be statistically analyzed to get high level understanding.

4.2.1 Numerical Variables. Descriptive Statistics.

Numerical variable	Mean	Standard Deviation	Kurtosis	Skewness	Range	Minimum	Maximum	Count
revenue	58.85	44.24	35.94	3.97	1,229.55	(6.17)	1,223.38	70,831
mou	525.73	530.13	8.00	2.18	7,667.75	0.00	7,667.75	70,831
recchrge	46.88	23.91	9.20	1.68	411.28	(11.29)	399.99	70,831
directas	0.89	2.20	483.61	12.06	159.39	0.00	159.39	70,831
overage	40.10	96.35	149.24	7.91	4,320.75	0.00	4,320.75	70,831
roam	1.22	9.08	5,192.06	55.93	1,112.45	0.00	1,112.45	70,831
changem	(10.85)	255.31	19.30	(0.26)	9,067.25	(3,875.00)	5,192.25	70,545
changer	(1.21)	38.77	364.42	6.48	3,591.22	(1,107.74)	2,483.48	70,545
dropvce	6.01	9.01	40.40	4.49	221.67	0.00	221.67	71,047
blckvce	4.07	10.67	164.93	9.59	384.33	0.00	384.33	71,047

unansvce	28.36	38.90	38.63	4.31	848.67	0.00	848.67	71,047
custcare	1.87	5.16	788.55	16.75	365.67	0.00	365.67	71,047
threeway	0.30	1.16	577.93	17.31	66.00	0.00	66.00	71,047
mourec	114.94	166.30	16.54	3.09	3,287.25	0.00	3,287.25	71,047
outcalls	25.40	35.15	21.42	3.46	644.33	0.00	644.33	71,047
incalls	8.18	16.52	61.54	5.62	519.33	0.00	519.33	71,047
peakvce	90.58	104.91	20.41	3.24	2,090.67	0.00	2,090.67	71,047
opeakvce	67.82	93.33	22.67	3.53	1,572.67	0.00	1,572.67	71,047
dropblk	10.15	15.46	69.20	5.69	489.67	0.00	489.67	71,047
callfwdv	0.01	0.56	10,372.08	92.82	81.33	0.00	81.33	71,047
callwait	1.85	5.56	220.16	11.02	212.67	0.00	212.67	71,047
months	18.75	9.79	0.90	1.06	55.00	6.00	61.00	71,047
uniqsubs	1.53	1.13	12,272.96	72.55	195.00	1.00	196.00	71,047
actvsubs	1.35	0.66	533.55	8.68	53.00	0.00	53.00	71,047
phones	1.81	1.34	21.46	3.33	27.00	1.00	28.00	71,046
models	1.56	0.91	9.79	2.40	15.00	1.00	16.00	71,046
eqpdays	380.27	254.29	1.62	1.09	1,828.00	(5.00)	1,823.00	71,046
age1	31.38	22.08	(1.06)	(0.25)	99.00	0.00	99.00	69,803
age2	21.16	23.92	(1.11)	0.54	99.00	0.00	99.00	69,803
retcalls	0.04	0.21	48.45	6.31	4.00	0.00	4.00	71,047
retacct	0.02	0.14	97.94	8.93	4.00	0.00	4.00	71,047
refer	0.05	0.29	3,169.41	32.47	35.00	0.00	35.00	71,047
income	4.33	3.14	(1.32)	(0.18)	9.00	0.00	9.00	71,047
setpre	35.80	57.04	2.91	1.72	499.99	0.00	499.99	71,047

TABLE 19: NUMERICAL VARIABLES. DESCRIPTIVE STATISTICS.

4.2.2 Categorical Binary variables

Variable	Binary 1	Binary 0	Count	1- %	0-%
churn	20,609	50,438	71,047	29.01%	70.99%
children	17,221	53,826	71,047	24.24%	75.76%
credita	7,420	63,627	71,047	10.44%	89.56%
creditaa	9,128	61,919	71,047	12.85%	87.15%

prizmrur	3,392	67,655	71,047	4.77%	95.23%
prizmub	22,814	48,233	71,047	32.11%	67.89%
prizmtwn	10,545	60,502	71,047	14.84%	85.16%
refurb	9,919	61,128	71,047	13.96%	86.04%
webcap	64,142	6,905	71,047	90.28%	9.72%
truck	13,301	57,746	71,047	18.72%	81.28%
rv	5,769	65,278	71,047	8.12%	91.88%
occprof	12,355	58,692	71,047	17.39%	82.61%
occcler	1,425	69,622	71,047	2.01%	97.99%
occcrft	2,106	68,941	71,047	2.96%	97.04%
occestud	538	70,509	71,047	0.76%	99.24%
occhmkr	224	70,823	71,047	0.32%	99.68%
occret	1,031	70,016	71,047	1.45%	98.55%
occselc	1,267	69,780	71,047	1.78%	98.22%
ownrent	23,582	47,465	71,047	33.19%	66.81%
marryun	27,340	43,707	71,047	38.48%	61.52%
marryes	25,959	45,088	71,047	36.54%	63.46%
mailord	25,717	45,330	71,047	36.20%	63.80%
mailres	26,799	44,248	71,047	37.72%	62.28%
mailflag	1,024	70,023	71,047	1.44%	98.56%
travel	4,084	66,963	71,047	5.75%	94.25%
pcown	13,173	57,874	71,047	18.54%	81.46%
credited	48,058	22,989	71,047	67.64%	32.36%
newcelly	13,708	57,339	71,047	19.29%	80.71%
newcelln	9,860	61,187	71,047	13.88%	86.12%
inmiss	17,750	53,297	71,047	24.98%	75.02%
mcycle	956	70,091	71,047	1.35%	98.65%
setprem	40,249	30,798	71,047	56.65%	43.35%
retcall	2,418	68,629	71,047	3.40%	96.60%

TABLE 20: CATEGORICAL BINARY VARIABLES. DESCRIPTIVE STATISTICS.

It is necessary to understand the dataset using statistical methods before it is used for prediction model. This preliminary study of the dataset cause to decide high-level picture of the data set. Descriptive statistics shows the insight on data set.

Understanding the data set missing value also a part of data cleaning and t will help to select correct method of data cleaning and how effective leach category can use.

4.2.3 Missing values in Numerical variables

Numerical Variables			
Variable	No of missing value	Variable	No of missing value
revenue	216	opeakvce	0
mou	216	dropblk	0
recchrge	216	callfwdv	0
directas	216	callwait	0
overage	216	months	0
roam	216	uniqsubs	0
changem	502	actvsbs	0
changer	502	phones	1
dropvce	0	models	1
blckvce	0	eqpdays	1
unansvce	0	age1	1244
custcare	0	age2	1244
threeway	0	retcalls	0
mourec	0	retacct	0
outcalls	0	refer	0
incalls	0	income	0
peakvce	0	setpre	0

TABLE 21: MISSING VALUES IN NUMERICAL VARIABLES

4.2.4 Missing values in Categorical (Boolean) variable

Boolean Variables			
Variable	No of missing value	Variable	No of missing value
children	0	ocself	0
credita	0	ownrent	0
credita	0	marryun	0
prizmrur	0	marryes	0
prizmub	0	mailord	0
prizmtwn	0	mailres	0
refurb	0	mailflag	0
webcap	0	travel	0
truck	0	pcown	0
rv	0	credited	0
occpof	0	newcelly	0
occler	0	newcelln	0
occrft	0	incmiss	0
ocstud	0	mcycle	0
occhmkr	0	setpre	0
occret	0	retcall	0
churn	0		

TABLE 22: CATEGORICAL (BOOLEAN) VARIABLE MISSING VALUES

4.3 Naïve Bayes Results analysis

Naïve Bayes classifier behaviors is presented and analyzed in here with respect to its accuracy in Cell2Cell Data set.

4.3.1 Naïve Bayes and Missing data processing

Cell2Cell dataset is subjected to preprocessing and prior to fit the prediction model. Data cleaning process was included missing data management and outlier processing. This processing caused to accuracy of the model and processing method were considered as variable.

X1- missing value Listwise deletion

Under the preprocessing or data cleaning process it is necessary to focus on the missing value in the subjected data set. There is different method of treating the missing values and based on the consideration such as dataset size, variable category, and the nature of the missingness such as MCAR, MAR, MNAR.

In here research has used listwise deletion based on the literature and larger size of the data set which will be derived a process dataset sizable for research work.

There are 3 independent variables in the research.

X1- missing value processing (categorical, nominal)

- 01. List wise deletion
- 02. Mean mode imputation.

X2- Outlier processing (categorical, nominal) Keep constant value mean mode imputation with 1.5 X SD span

X3- Feature selection Keep constant at 10 features.

Dependent variable Y -accuracy of Naïve Bayes classifier output

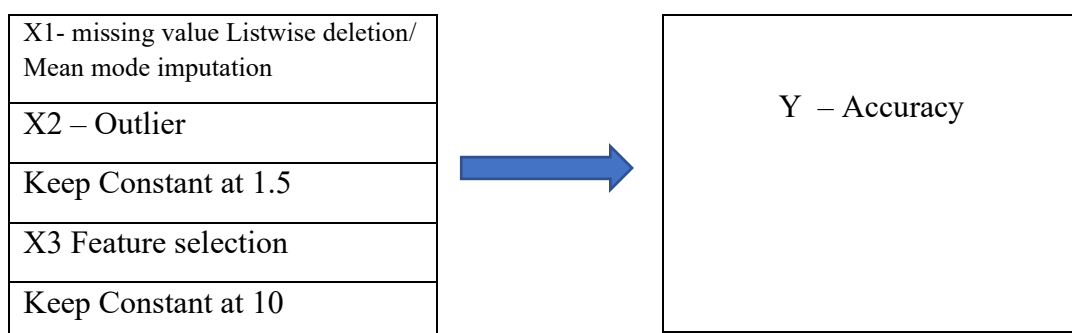


FIGURE 24: MODEL OF TESTING NB IN MISSING VALUES

Below graph and tables is showing the variation of missing value processing method

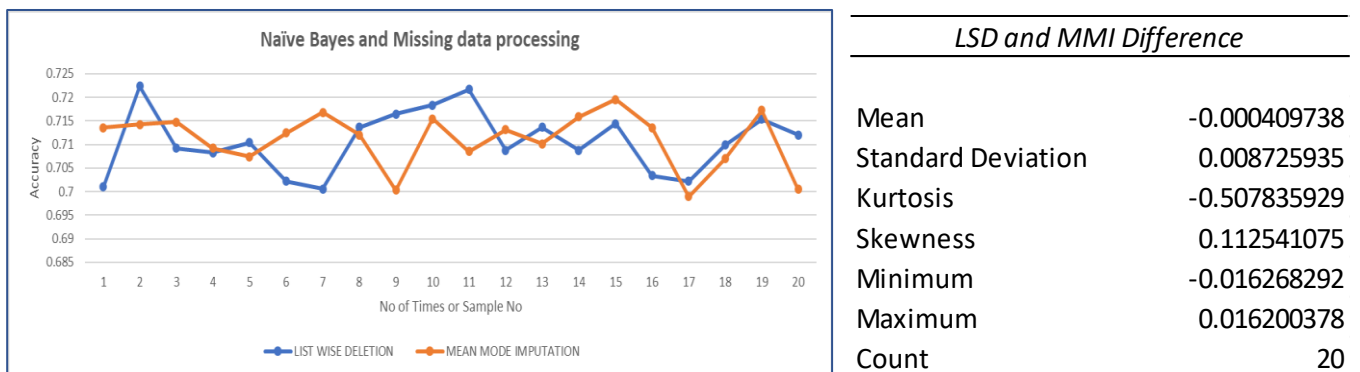


FIGURE 25: NB MISSING DATA PROCESSING

As per above graph there is random and closely following two curves were appearing for List wise deletion and Mean mode imputation method. When accuracy difference of List wise deletion and Mean mode imputation method is considered it has varied with Mean of (-0.000409) and SD of 0.0008725 with respect to 20 set sample.

To get the better understanding of the difference of List wise deletion and Mean mode imputation method is consider it can show as below with 200 data samples.

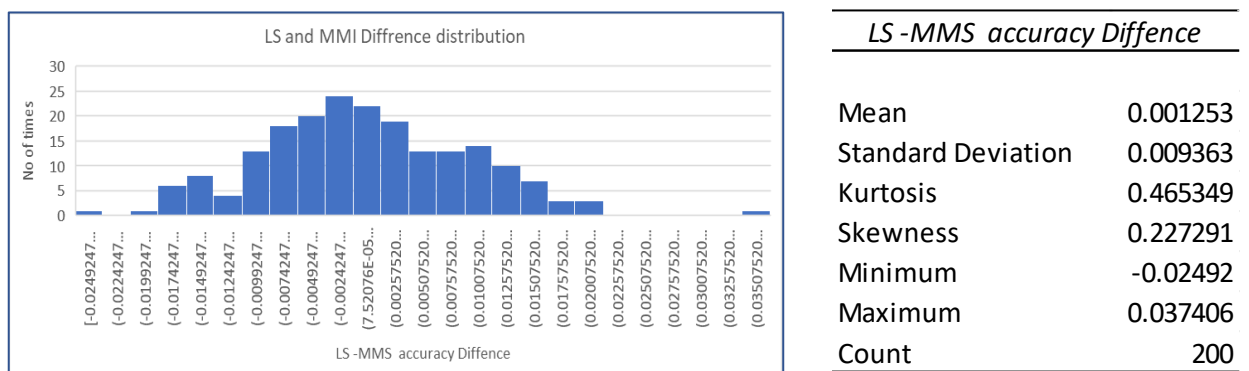


FIGURE 26: NB ACCURACY LWD AND MMI DIFFERENCE DISTRIBUTION

When accuracy difference of List wise deletion and Mean mode imputation method is considered it has nearly Gaussian distribution with varied with Mean of 0.001253 and SD of 0.009363 with respect to 200 data samples.

Analysis Naïve Bayes accuracy against list wise deletion and Mean mode imputation

As per the result difference of List wise deletion and Mean mode imputation method variable has closely follow a nearly Gaussian distribution with varied with Mean of (0.001253) and SD of 0.009363. it implies of application of List wise deletion and Mean mode imputation method has 0.0001 accuracy difference can be impacted with 0.009363 with stander deviation. Based on this result these two methods Significantly smaller impact on churn prediction in telecommunication domain.

4.3.2 Naïve Bayes and Outlier data processing

Under the preprocessing or data cleaning process it is necessary to focus on the outlier value in the subjected data set. There is different method of treating the outlier values and based on the consideration such as dataset size, variable category, nature of experiment (time based) and relationship to independent and dependent variable

In here research has used listwise deletion – Trimming based on the literature and larger size of the data set which will be derived a process dataset sizable for research work.

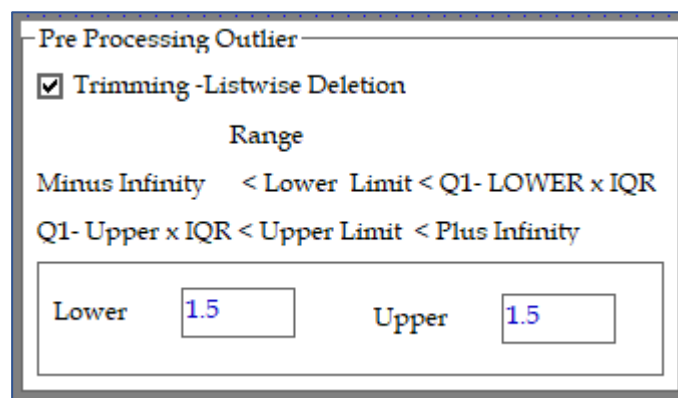


FIGURE 27: OUTLIER TRIMMING RANGE

When we listwise deletion – Trimming is considered it is necessary to define what is the focus region as acceptable data records and considered as outlier. It has defined the span for both side from mean value with variable factor of SD for each numerical variable.

There are 3 independent variables in the research.

X1- missing value processing (categorical, nominal)

03. List wise deletion

04. Mean mode imputation.

X2- Outlier processing (categorical, nominal) vary from 0.2 to 1.5-time SD as per each variable own mean and SD.

X3- Feature selection Keep constant at 10 features.

Dependent variable Y -accuracy of Naïve buyer’s classifier output

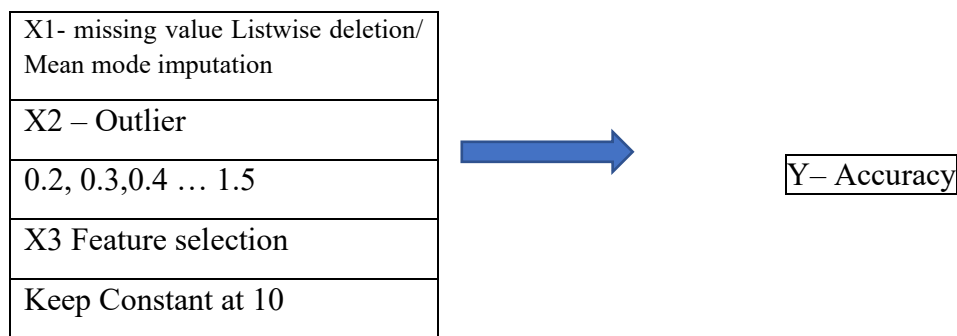


FIGURE 28:MODEL OF TESTING NB IN OUTLIER

Below graph and tables is showing the variation of Outlier processing method

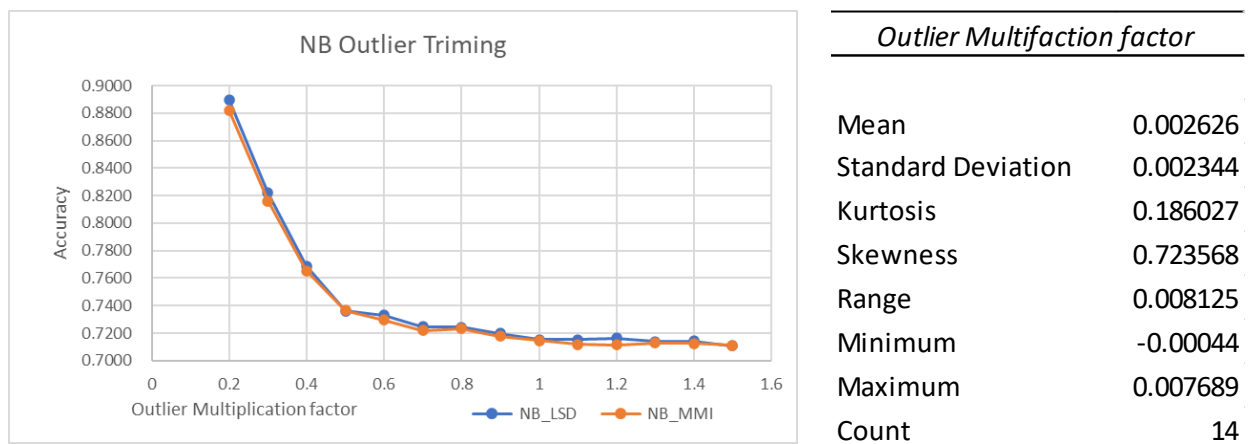


FIGURE 29:NB OUTLIER DATA PROCESSING

As per above graph there is closely following two curves were appearing for List wise deletion and Mean mode imputation method. When accuracy difference of List wise deletion and Mean mode imputation method is considered it has varied with Mean of 0.002626 and SD of 0.002347 with respect to 14 set sample. And with trimming factor it

shows closely negative exponential distribution with constant value as per the shape of the curve.

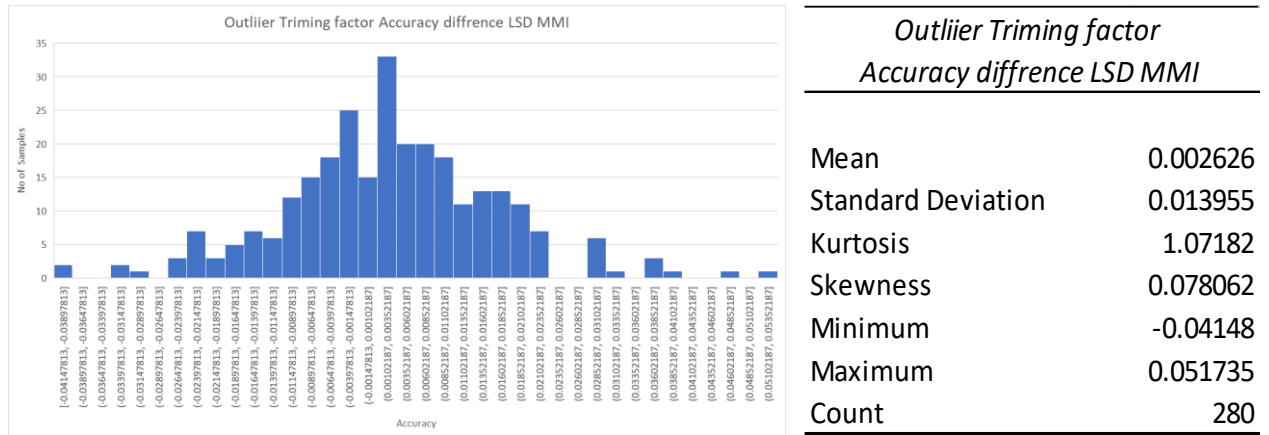


FIGURE 30: NB ACCURACY AND OUTLIER FACTOR

When accuracy difference of List wise deletion and Mean mode imputation method is considered it has nearly Gaussian distribution with varied with Mean of 0.002626 and SD of 0.013955 with respect to 280 data samples.

Analysis Naïve Bayes accuracy against list wise deletion and Mean mode imputation

As per the result difference of List wise deletion and Mean mode imputation method variable has closely follow a nearly Gaussian distribution with varied with Mean of 0.002626 and SD of 0.019355 it implies of application of List wise deletion and Mean mode imputation method has 0.002 accuracy difference can be impacted with 0.014 with stander deviation. Based on this result these two methods Significantly smaller impact on churn prediction in telecommunication domain.

4.3.3 Naïve Bayes and feature selection

Under the preprocessing and improving model performance it is necessary to focus on the feature selection in the data set. In this case there are two main type of main variable types of namely categorical nominal variable and continuous numerical variable.

In here research has used listwise deletion based on the literature and larger size of the data set which will be derived a process dataset sizable for research work.

There are 3 independent variables in the research.

X1- missing value processing (categorical, nominal)

01. List wise deletion
02. Mean mode imputation.

X2- Outlier processing (categorical, nominal) Keep constant value mean mode imputation with 1.5 X SD span

X3- Feature selection (1-10)

Dependent variable Y -accuracy of Naïve Bayes classifier output

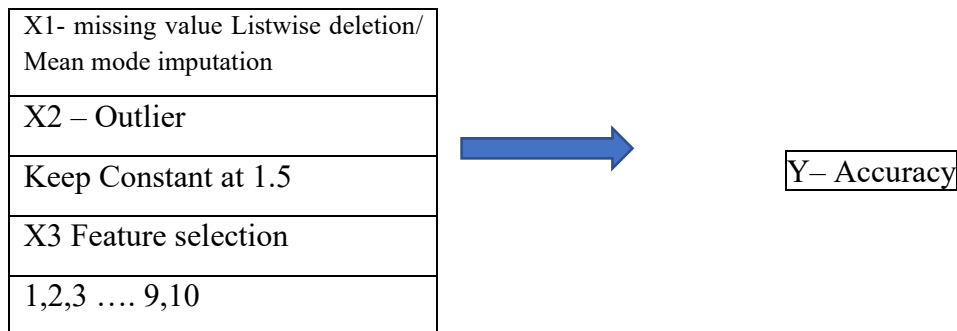


FIGURE 31: MODEL OF TESTING NB IN FEATURE SELECTION

Below graph and tables is showing the how feature selection has impact on Naïve Bayes accuracy based on stational method (Chi-square method, Anova) under the filtering methodology

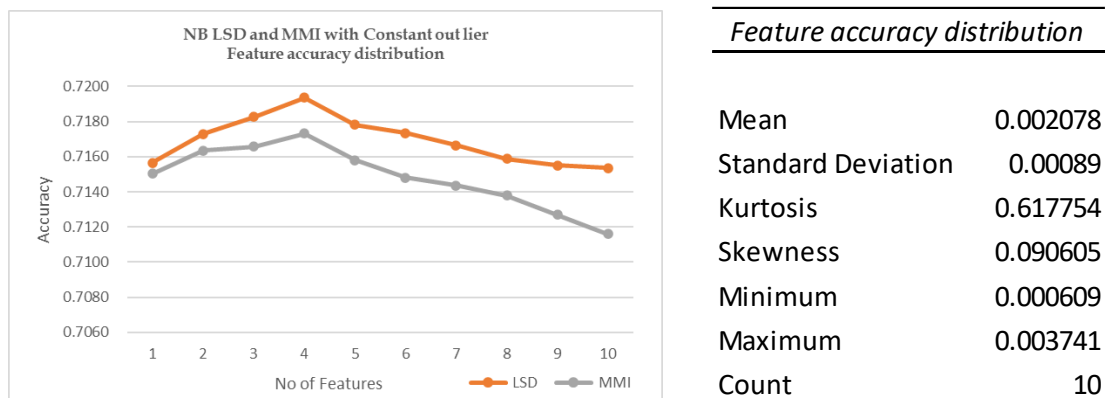


FIGURE 32: NB FEATURE SELECTION

As per the above graph it has mean value of 0.002078 and SD of 0.00089 and once shape of the graph is considered it has maximum at the 4 feature and dropping the accuracy once it accommodated more features.

4.3.4 Naïve Bayes Stabilization

Under the preprocessing or data cleaning process it is necessary to focus on the missing value and outlier management in the subjected data set. There is different method of treating the missing values and based on the consideration such as dataset size, variable category, and the nature of the missingness such as MCAR, MAR, MNAR. In the outlier it can be used Mean mode imputation or trimming (list wise deletion)

Then this data set has 67 feature and it should be process under feature selection. Filter method was used sperate statistical parameter (Anova and chi-square)

In here research has process all three different variable movement and dependent variable accuracy against it.

There are 3 independent variables in the research.

X1- missing value processing (categorical, nominal)

01. List wise deletion (keep constant)
02. Mean mode imputation.

X2- Outlier processing (categorical, nominal) with (0.2~1.5) X SD span

X3- Feature selection (1~10)

Dependent variable Y -accuracy of Naïve bayes classifier output

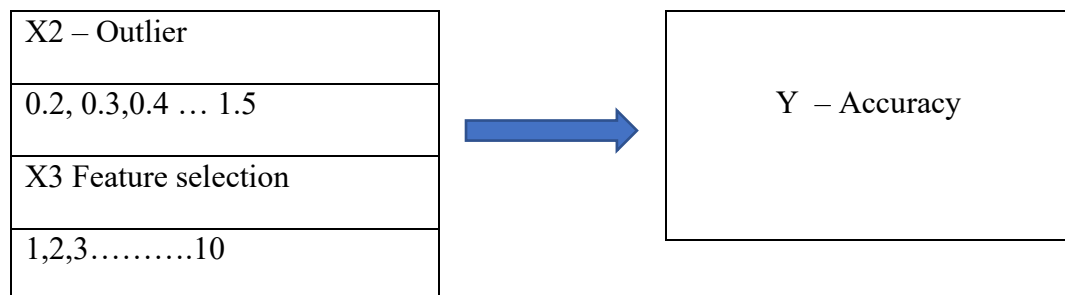
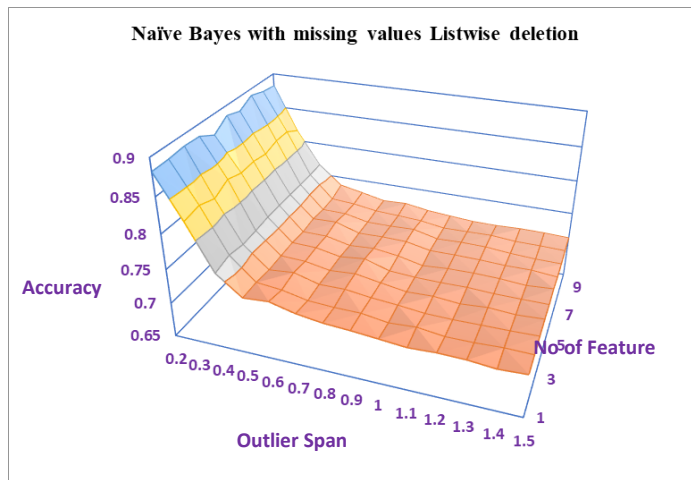


FIGURE 33: MODEL OF TESTING NB ACCURACY WITH 3 VARIABLES

Below graph and tables is showing the variation of missing value processing method



Naïve Bayes and Listwise deletion

Mean	0.742805
Standard Deviation	0.049448
Kurtosis	2.89964
Skewness	2.021323
Range	0.181526
Minimum	0.710631
Maximum	0.892157
Count	140

FIGURE 34: NB ACCURACY DISTRIBUTION

As per above graph accuracy plane is gradually dropping and settled. When accuracy distribution is considered with respect to three independent variables it has varied with Mean of 0.7428 (74.28%) and SD of 0.040448 (4.98%) with respect to 140 sample data set. These each sample is derived as average value of 20 time. So total samples subjected for this derivation is $20 \times 140 = 2,800$ no of test for accuracy.

4.4 Decision Tree Results Analysis

4.4.1 Decision Tree and Missing data processing

Cell2Cell dataset is subjected to preprocessing and prior to fit the prediction model. Data cleaning process was included missing data management and outlier processing. This processing caused to accuracy of the model and processing method were considered as variable.

X1- missing value Listwise deletion

Under the preprocessing or data cleaning process it is necessary to focus on the missing value in the subjected data set. There is different method of treating the missing values and based on the consideration such as dataset size, variable category, and the nature of the missingness such as MCAR, MAR, MNAR.

In here research has used listwise deletion based on the literature and larger size of the data set which will be derived a process dataset sizable for research work.

There are 3 independent variables in the research.

X1- missing value processing (categorical, nominal)

01. List wise deletion
02. Mean mode imputation.

X2- Outlier processing (categorical, nominal) Keep constant value mean mode imputation with 1.5 X SD span

X3- Feature selection Keep constant at 10 features.

Dependent variable Y -accuracy of Decision Tree classifier output

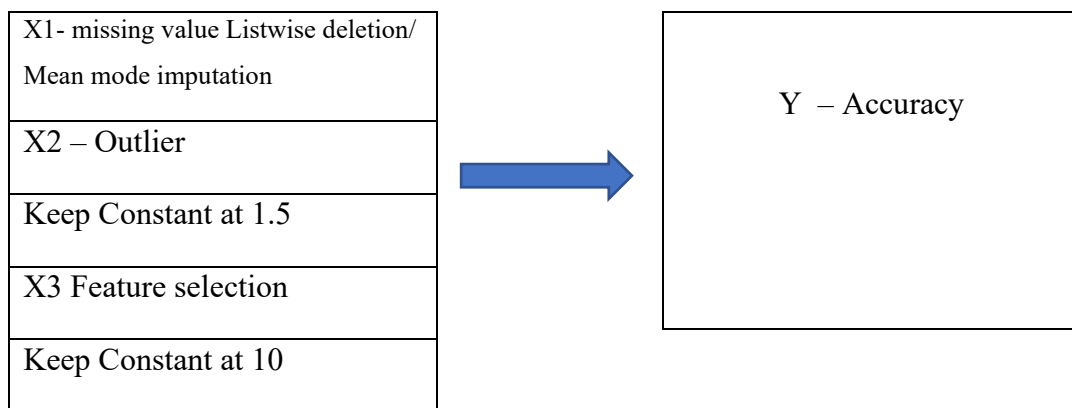


FIGURE 35: MODEL OF TESTING DT IN MISSING VALUES

Below graph and tables is showing the variation of missing value processing method

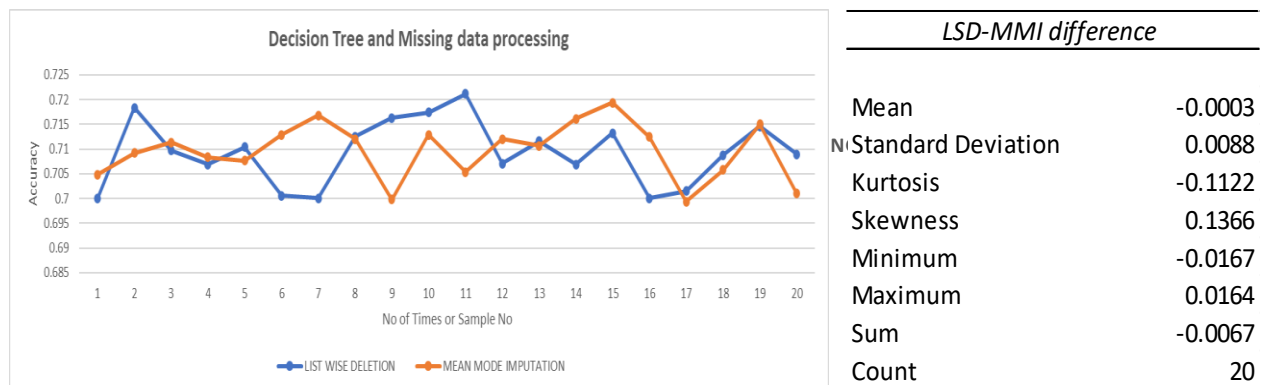


FIGURE 37 : DT MISSING DATA PROCESSING

As per above graph there is random and closely following two curves were appearing for List wise deletion and Mean mode imputation method. When accuracy difference of List wise deletion and Mean mode imputation method is considered it has varied with Mean of (-0.0003) and SD of 0.0088 with respect to 20 set sample.

To get the better understanding of the difference of List wise deletion and Mean mode imputation method is consider it can show as below with 200 data samples.

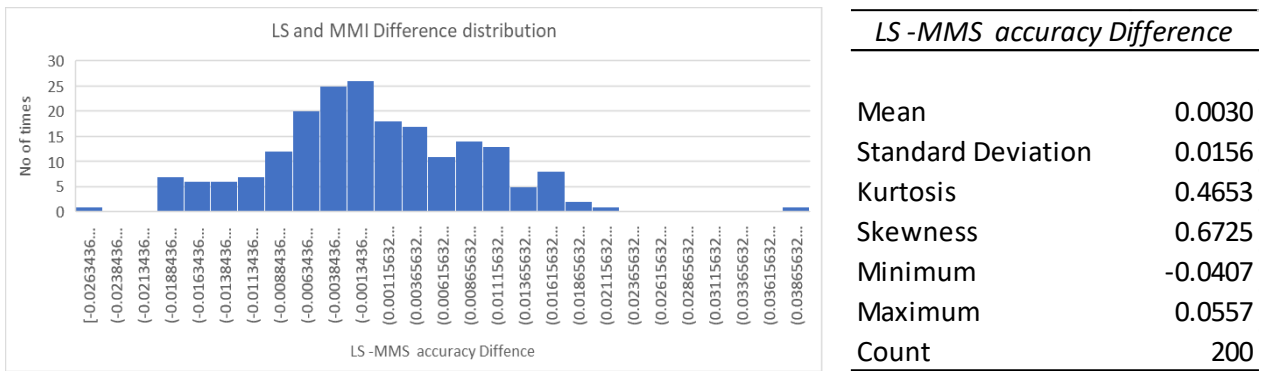


FIGURE 38: DT ACCURACY LWD AND MMI DIFFERENCE DISTRIBUTION

When accuracy difference of List wise deletion and Mean mode imputation method is considered it has nearly Gaussian distribution with varied with Mean of 0.0030 and SD of 0.0156 with respect to 200 data samples.

Analysis Naïve Bayes accuracy against list wise deletion and Mean mode imputation

As per the result difference of List wise deletion and Mean mode imputation method variable has closely follow a nearly Gaussian distribution with varied with Mean of 0.0030 and SD of 0.0156 it implies of application of List wise deletion and Mean mode imputation method has 0.003 accuracy difference can be impacted with 0.00156 with stander deviation. Based on this result these two methods Significantly smaller impact on churn prediction in telecommunication domain.

4.4.2 Decision Tree and Outlier data processing

Under the preprocessing or data cleaning process it is necessary to focus on the outlier value in the subjected data set. There is different method of treating the outlier values and based on the consideration such as dataset size, variable category, nature of experiment (time based) and relationship to independent and dependent variable

In here research has used listwise deletion – Trimming based on the literature and larger size of the data set which will be derived a process dataset sizable for research work.

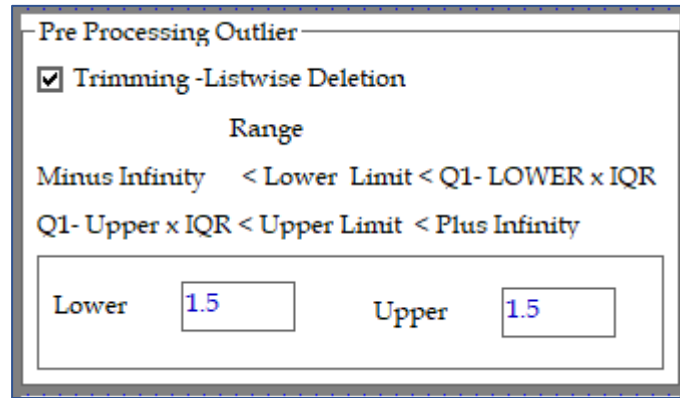


FIGURE 39:OUTLIER TRIMMING RANGE

When we listwise deletion – Trimming is considered it is necessary to define what is the focus region as acceptable data records and considered as outlier. It has defined the span for both side from mean value with variable factor of SD for each numerical variable.

There are 3 independent variables in the research.

X1- missing value processing (categorical, nominal)

- 01. List wise deletion
- 02. Mean mode imputation.

X2- Outlier processing (categorical, nominal) vary from 0.2 to 1.5-time SD as per each variable own mean and SD.

X3- Feature selection Keep constant at 10 features.

Dependent variable Y -accuracy of Decision tree classifier output

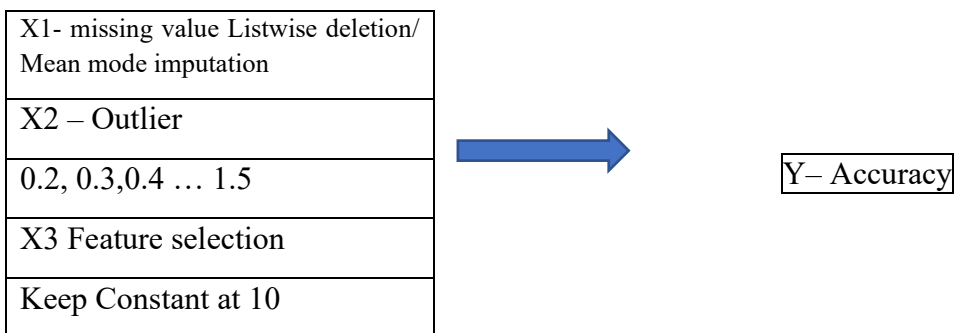
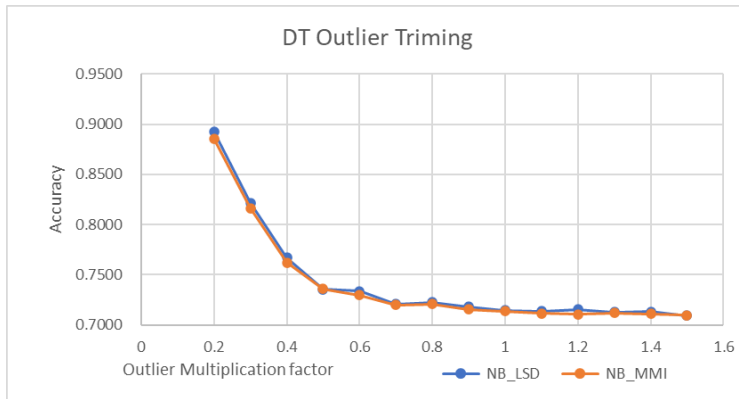


FIGURE 40: MODEL OF TESTING DT IN OUTLIER

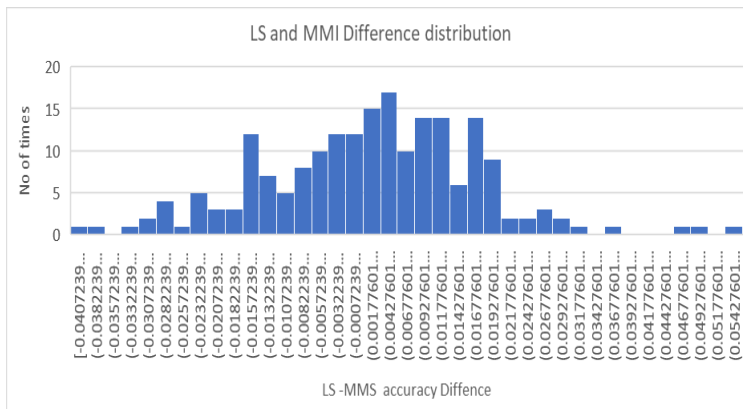
Below graph and tables is showing the variation of Outlier processing method



<i>Outlier Multifaction factor</i>	
Mean	0.0026
Standard Deviation	0.002202
Kurtosis	-0.39067
Skewness	0.575167
Range	0.007701
Minimum	(0.0003)
Maximum	0.0074
Count	14

FIGURE 41: DT OUTLIER DATA PROCESSING

As per above graph there is closely following two curves were appearing for List wise deletion and Mean mode imputation method. When accuracy difference of List wise deletion and Mean mode imputation method is considered it has varied with Mean of 0.0026 and SD of 0.0022 with respect to 14 set sample. And with trimming factor it shows closely negative exponential distribution with constant value as per the shape of the curve.



<i>LS -MMS accuracy Difference</i>	
Mean	0.002951
Standard Deviation	0.015601
Kurtosis	0.465349
Skewness	0.672457
Minimum	-0.04072
Maximum	0.055656
Count	200

FIGURE 42: DT ACCURACY AND OUTLIER FACTOR

When accuracy difference of List wise deletion and Mean mode imputation method is considered it has nearly Gaussian distribution with varied with Mean of 0.0002951 and SD of 0.01560 with respect to 200 data samples.

Analysis Decision tree accuracy against list wise deletion and Mean mode imputation

As per the result difference of List wise deletion and Mean mode imputation method variable has closely follow a nearly Gaussian distribution with varied with Mean of 0.0002951 and SD of 0.01560 it implies of application of List wise deletion and Mean mode imputation method has 0.0002 accuracy difference can be impacted with 0.015 with stander

deviation. Based on this result these two methods Significantly smaller impact on churn prediction in telecommunication domain.

4.4.3 Decision Tree and feature selection

Under the preprocessing and improving model performance it is necessary to focus on the feature selection in the data set. In this case there are two main type of main variable types of namely categorical nominal variable and continuous numerical variable.

In here research has used listwise deletion based on the literature and larger size of the data set which will be derived a process dataset sizable for research work.

There are 3 independent variables in the research.

X1- missing value processing (categorical, nominal)

01. List wise deletion
02. Mean mode imputation.

X2- Outlier processing (categorical, nominal) Keep constant value mean mode imputation with 1.5 X SD span

X3- Feature selection (1-10)

Dependent variable Y -accuracy of Decision tree classifier output

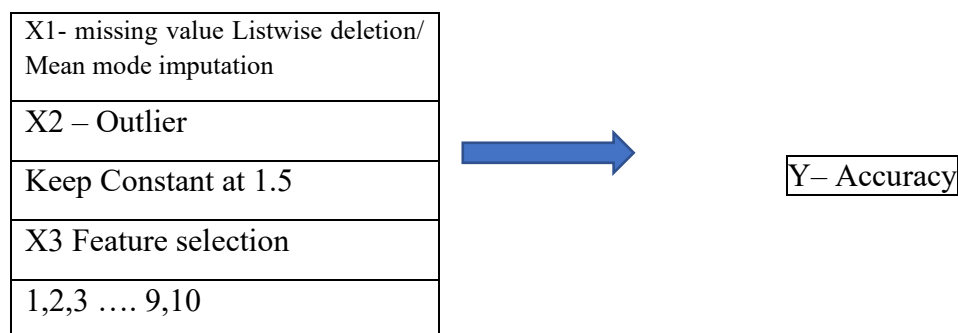
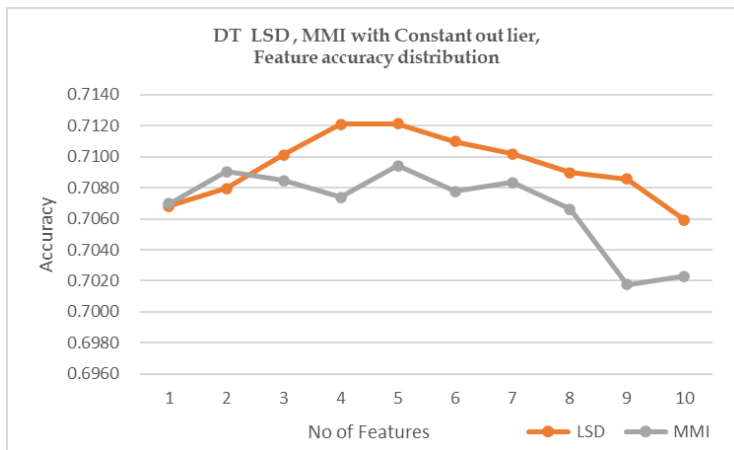


FIGURE 43: FIGURE 26: MODEL OF TESTING DT ACCURACY

accuracy based on stactical method (Chi-square method, Anova) under the filtering methodology.



<i>Feature accuracy distribution</i>	
Mean	0.00208
Standard Deviation	0.00089
Kurtosis	0.61775
Skewness	0.09061
Minimum	0.00061
Maximum	0.00374
Count	10

FIGURE 44: DT FEATURE SELECTION

As per the above graph it has mean value of 0.00208 and SD of 0.00089 and once shape of the graph is considered it has maximum at the 4 feature and dropping the accuracy once it accommodated more features.

4.4.4 Decision Tree Stabilization

Under the preprocessing or data cleaning process it is necessary to focus on the missing value and outlier management in the subjected data set. There is different method of treating the missing values and based on the consideration such as dataset size, variable category, and the nature of the missingness such as MCAR, MAR, MNAR. In the outlier it can be used Mean mode imputation or trimming (list wise deletion)

Then this data set has 67 feature and it should be process under feature selection. Filter method was used sperate statistical parameter (Anova and chi-square)

In here research has process all three different variable movement and dependent variable accuracy against it.

There are 3 independent variables in the research.

X1- missing value processing (categorical, nominal)

01. List wise deletion (keep constant)
02. Mean mode imputation.

X2- Outlier processing (categorical, nominal) with (0.2~1.5) X SD span

X3- Feature selection (1~10)

Dependent variable Y -accuracy of Naïve bayes classifier output

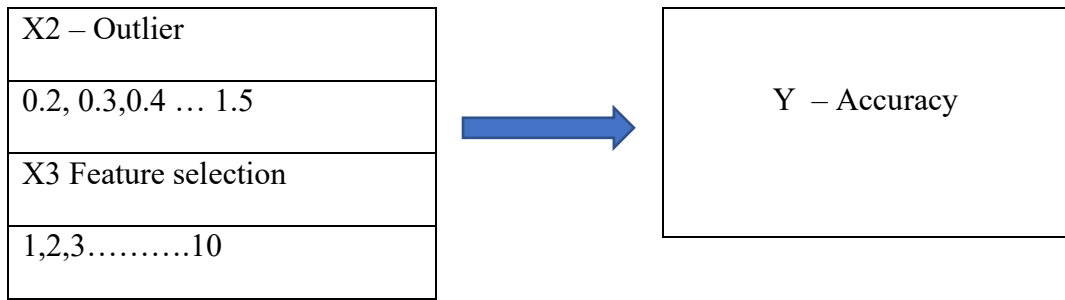


FIGURE 45: MODEL OF TESTINGDT ACCURACY WITH 3 VARIABLES

Below graph and tables is showing the variation of missing value processing method

<i>Decision Tree with Listwise deletion</i>	
Mean	0.74179
Standard Deviation	0.04991
Kurtosis	2.99769
Skewness	2.01838
Range	0.18403
Minimum	0.70931
Maximum	0.89333
Count	140

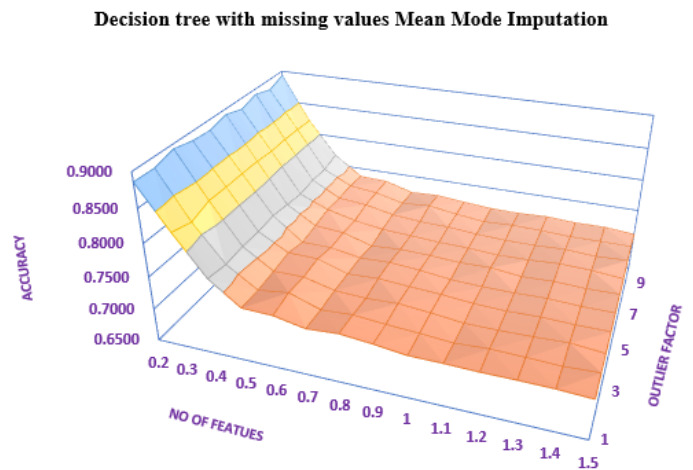


FIGURE 46: SUPPORT VECTOR MACHINE RESULTS ANALYSIS

4.5 Support Vector Machine Results Analysis

4.5.1 Support Vector Machine and Missing data processing

Cell2Cell dataset is subjected to preprocessing and prior to fit the prediction model. Data cleaning process was included missing data management and outlier processing. This processing caused to accuracy of the model and processing method were considered as variable.

X1- missing value Listwise deletion

Under the preprocessing or data cleaning process it is necessary to focus on the missing value in the subjected data set. There is different method of treating the missing values and

based on the consideration such as dataset size, variable category, and the nature of the missingness such as MCAR, MAR, MNAR.

In here research has used listwise deletion based on the literature and larger size of the data set which will be derived a process dataset sizable for research work.

There are 3 independent variables in the research.

X1- missing value processing (categorical, nominal)

01. List wise deletion
02. Mean mode imputation.

X2- Outlier processing (categorical, nominal) Keep constant value mean mode imputation with 1.5 X SD span

X3- Feature selection Keep constant at 10 features.

Dependent variable Y -accuracy of Support vector classifier output

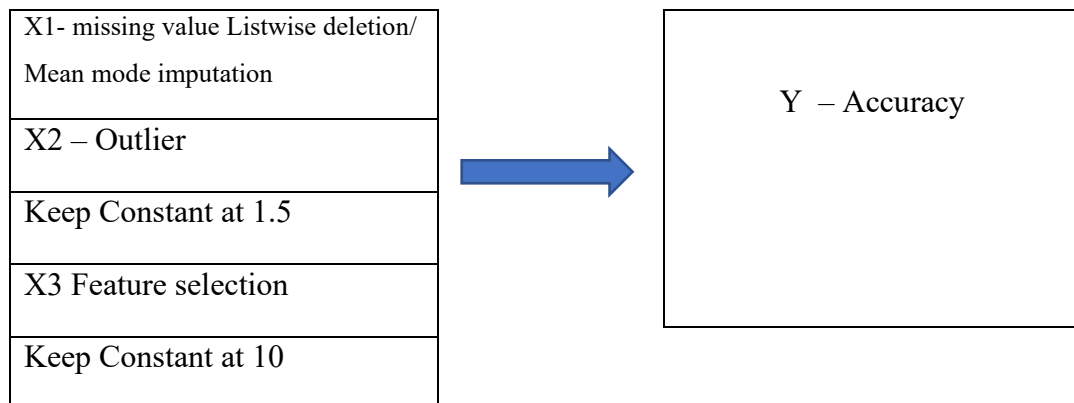
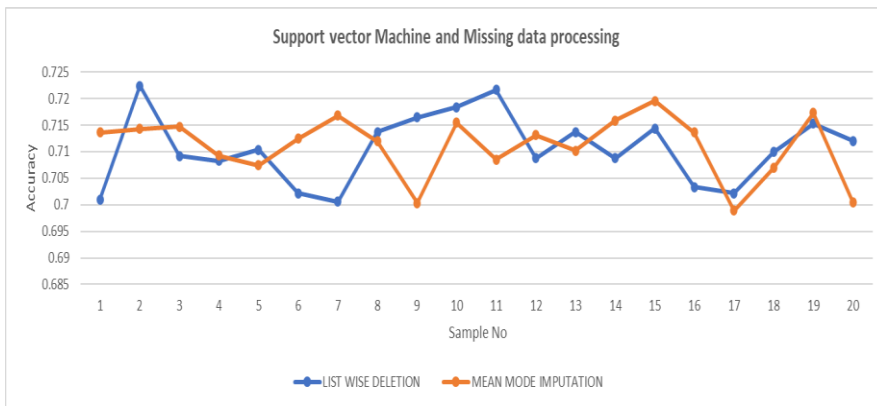


FIGURE 47: MODEL OF TESTING SVM IN MISSING VALUES

Below graph and tables is showing the variation of missing value processing method

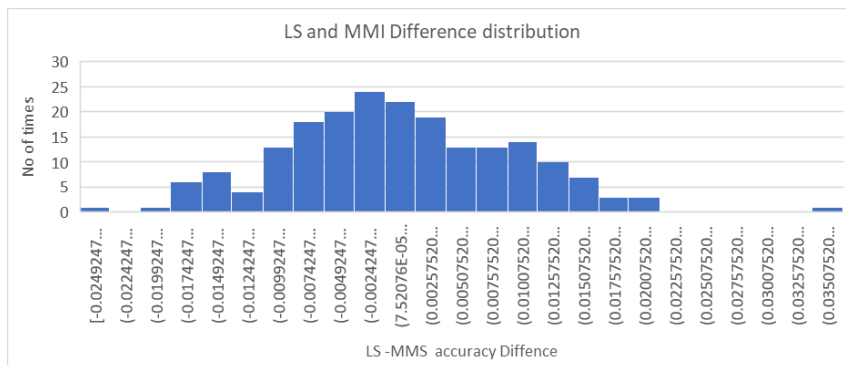


<i>LSD-MMI difference</i>	
Mean	-0.0004
Standard Deviation	0.0085
Kurtosis	-0.5078
Skewness	0.1039
Minimum	-0.0163
Maximum	0.0162
Sum	-0.0082
Count	20

FIGURE 48: SVM MISSING DATA PROCESSING

As per above graph there is random and closely following two curves were appearing for List wise deletion and Mean mode imputation method. When accuracy difference of List wise deletion and Mean mode imputation method is considered it has varied with Mean of (-0.0004) and SD of 0.0085 with respect to 20 set sample.

To get the better understanding of the difference of List wise deletion and Mean mode imputation method is consider it can show as below with 200 data samples.



<i>LS -MMS accuracy Difference</i>	
Mean	0.0013
Standard Deviation	0.0093
Kurtosis	0.4653
Skewness	0.4653
Minimum	-0.0249
Maximum	0.0374
Count	200

FIGURE 49: SVM ACCURACY LWD AND MMI DIFFERENCE DISTRIBUTION

When accuracy difference of List wise deletion and Mean mode imputation method is considered it has nearly Gaussian distribution with varied with Mean of 0.0013 and SD of 0.0093 with respect to 200 data samples.

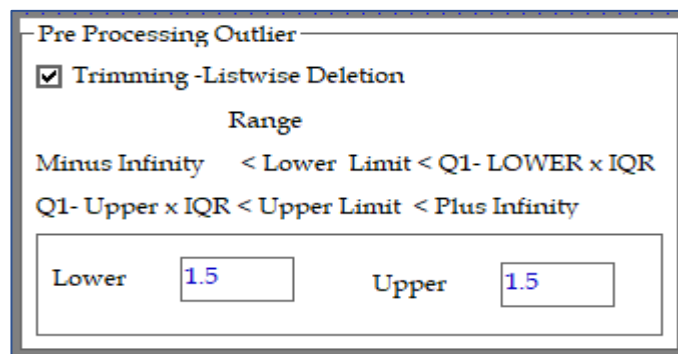
Analysis Naïve Bayes accuracy against list wise deletion and Mean mode imputation

As per the result difference of List wise deletion and Mean mode imputation method variable has closely follow a nearly Gaussian distribution with varied with Mean of 0.0013 and SD of 0.0093 it implies of application of List wise deletion and Mean mode imputation method has 0.001 accuracy difference can be impacted with 0.009 with stander deviation. Based on this result these two methods Significantly smaller impact on churn prediction in telecommunication domain.

4.5.2 Support Vector Machine and Outlier data processing

Under the preprocessing or data cleaning process it is necessary to focus on the outlier value in the subjected data set. There is different method of treating the outlier values and based on the consideration such as dataset size, variable category, nature of experiment (time based) and relationship to independent and dependent variable

In here research has used listwise deletion – Trimming based on the literature and larger size of the data set which will be derived a process dataset sizable for research work.



The image shows a software dialog box titled "Pre Processing Outlier". It contains a checked checkbox for "Trimming -Listwise Deletion". Below this, the text "Range" is centered. The range is defined by the inequalities: "Minus Infinity < Lower Limit < Q1- LOWER x IQR" and "Q1- Upper x IQR < Upper Limit < Plus Infinity". At the bottom, there are two input fields: "Lower" with a value of "1.5" and "Upper" with a value of "1.5".

FIGURE 50: OUTLIER TRIMMING RANGE

When we listwise deletion – Trimming is considered it is necessary to define what is the focus region as acceptable data records and considered as outlier. It has defined the span for both side from mean value with variable factor of SD for each numerical variable.

There are 3 independent variables in the research.

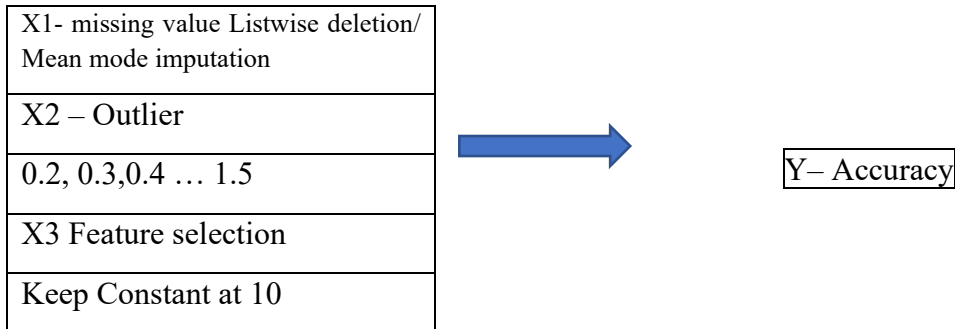
X1- missing value processing (categorical, nominal)

01. List wise deletion
02. Mean mode imputation.

X2- Outlier processing (categorical, nominal) vary from 0.2 to 1.5-time SD as per each variable own mean and SD.

X3- Feature selection Keep constant at 10 features.

Dependent variable Y -accuracy of Support vector classifier output



Below graph and tables is showing the variation of Outlier processing method

FIGURE 51: MODEL OF TESTING SVM IN OUTLIER

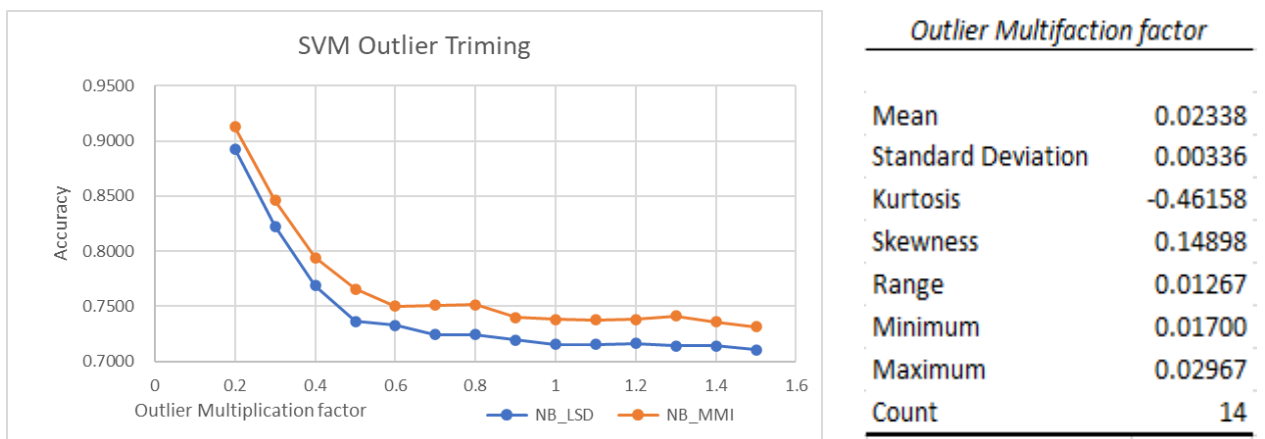


FIGURE 52: SVM OUTLIER DATA PROCESSING

As per above graph there is closely following two curves were appearing for List wise deletion and Mean mode imputation method. When accuracy difference of List wise deletion and Mean mode imputation method is considered it has varied with Mean of 0.023338 and SD of 0.00336 with respect to 14 set sample. And with trimming factor it shows closely negative exponential distribution with constant value as per the shape of the curve.

<i>Outlier Trimming factor</i>	
<i>Accuracy difference LSD MMI</i>	
Mean	0.00263
Standard Deviation	0.01393
Kurtosis	1.07182
Skewness	0.07806
Minimum	-0.04148
Maximum	0.05173
Count	280

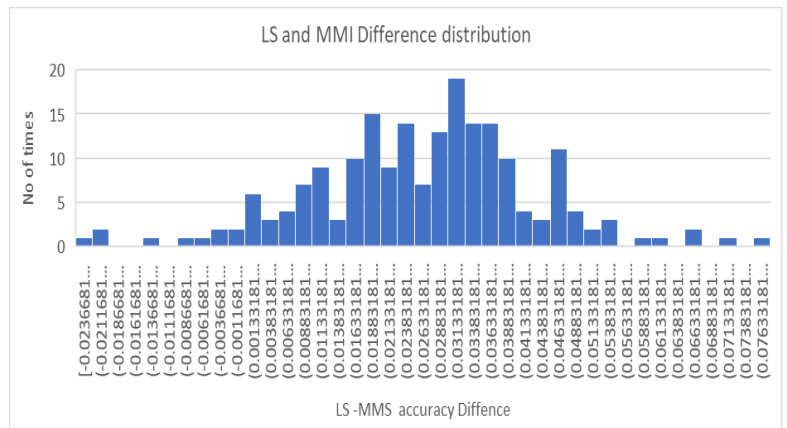


FIGURE 53: SVM ACCURACY AND OUTLIER FACTOR

When accuracy difference of List wise deletion and Mean mode imputation method is considered it has nearly Gaussian distribution with varied with Mean of 0.002631 and SD of 0.01393 with respect to 200 data samples.

Analysis Support Vector Machine accuracy against list wise deletion and Mean mode imputation

As per the result difference of List wise deletion and Mean mode imputation method variable has closely follow a nearly Gaussian distribution with varied with Mean of 0.002631 and SD of 0.01393 it implies of application of List wise deletion and Mean mode imputation method has 0.0002 accuracy difference can be impacted with 0.013 with stander deviation. Based on this result these two methods Significantly smaller impact on churn prediction in telecommunication domain.

4.5.3 Support Vector Machine and feature selection

Under the preprocessing and improving model performance it is necessary to focus on the feature selection in the data set. In this case there are two main type of main variable types of namely categorical nominal variable and continuous numerical variable.

In here research has used listwise deletion based on the literature and larger size of the data set which will be derived a process dataset sizable for research work.

There are 3 independent variables in the research.

X1- missing value processing (categorical, nominal)

01. List wise deletion

02. Mean mode imputation.

X2- Outlier processing (categorical, nominal) Keep constant value mean mode imputation with 1.5 X SD span

X3- Feature selection (1-10)

Dependent variable Y -accuracy of Decision tree classifier output

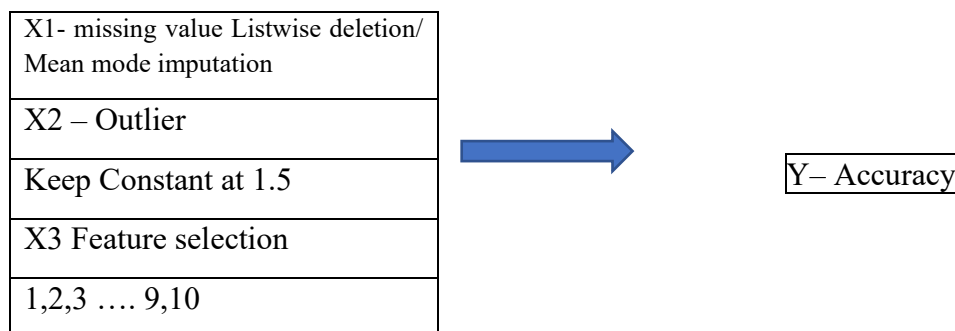
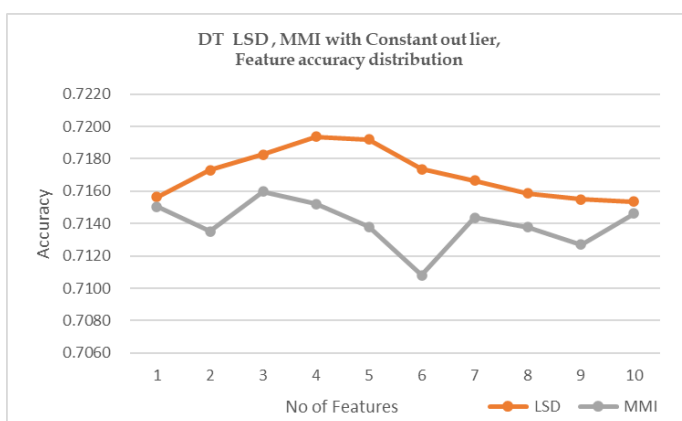


FIGURE 54: FIGURE 26: MODEL OF TESTING SVM ACCURACY

Below graph and tables is showing the how feature selection has impact on Naïve Bayes accuracy based on stactical method (Chi-square method, Anova) under the filtering methodology.



<i>Feature accuracy distribution</i>	
Mean	0.00208
Standard Deviation	0.00089
Kurtosis	0.61775
Skewness	0.09061
Minimum	0.00061
Maximum	0.00374
Count	10

FIGURE 55: SVM FEATURE SELECTION

As per the above graph it has mean value of 0.002078 and SD of 0.00089 and once shape of the graph is considered it has maximum at the 4 feature and dropping the accuracy once it accommodated more features.

4.5.4 Support Vector Machine Stabilization

Under the preprocessing or data cleaning process it is necessary to focus on the missing value and outlier management in the subjected data set. There is different method of treating the missing values and based on the consideration such as dataset size, variable category, and the nature of the missingness such as MCAR, MAR, MNAR. In the outlier it can be used Mean mode imputation or trimming (list wise deletion)

Then this data set has 67 feature and it should be process under feature selection. Filter method was used sperate statistical parameter (Anova and chi-square)

In here research has process all three different variable movement and dependent variable accuracy against it.

There are 3 independent variables in the research.

X1- missing value processing (categorical, nominal)

01. List wise deletion (keep constant)
02. Mean mode imputation.

X2- Outlier processing (categorical, nominal) with (0.2~1.5) X SD span

X3- Feature selection (1~10)

Dependent variable Y -accuracy of Naïve bayes classifier output

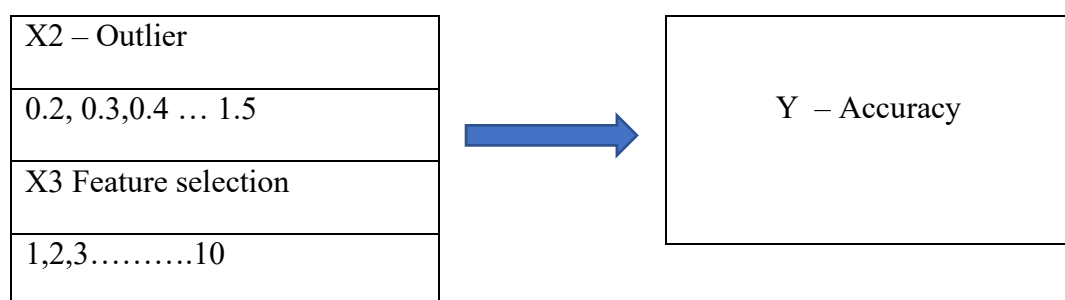


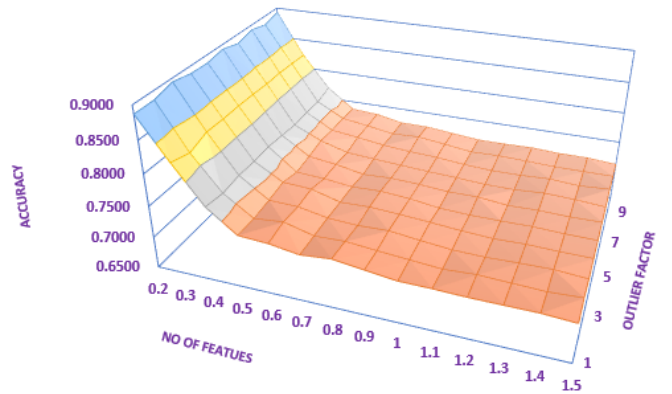
FIGURE 56: MODEL OF TESTING SVM ACCURACY WITH 3 VARIABLES

Below graph and tables is showing the variation of missing value processing method

SVM with Listwise deletion

Mean	0.74304
Standard Deviation	0.04996
Kurtosis	3.00611
Skewness	2.02129
Range	0.18407
Minimum	0.71063
Maximum	0.89471
Count	140

Support Vector Machine with missing values Mean Mode Imputation



4.6 Ensembled classifiers Analysis

Ensemble classifiers can be made combining two classifiers together. In this research study as per the methodology section two combine classifiers performance are considered with respect to contributed single classifiers.

4.6.1 Naïve Bayer and Decision tree Ensembled classifiers

Naïve bayes and Decision tree classifiers ensembled and form an ensembled classifiers and it is necessary to identify what will the progress of the ensembled classifiers with respect to in base classifiers name Naïve bayes and decision tree.

X1(missing value): is varied form Listwise deletion and mean mode imputation

X2 outlier Factor: 0.2, 0.3, 0.4, 1.5

Feature selection: 10

With above for initiation input variable were considered as independent variable and accuracy was considered as output variable.

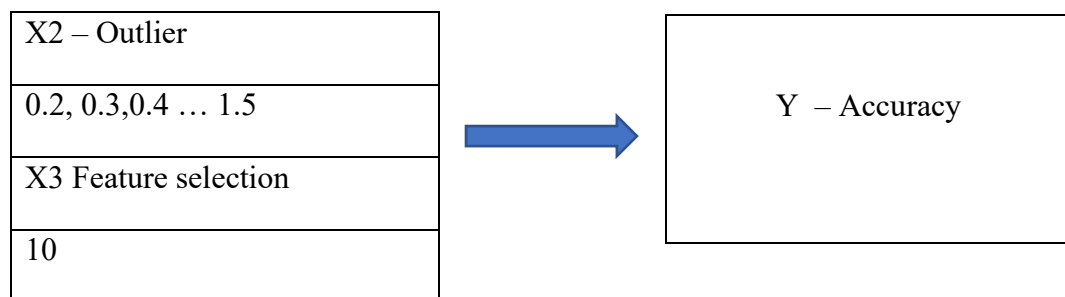


FIGURE 57: BASE DIAGRAM FOR NB, DT ENS CLASSIFIERS

Accuracy was measured for Naïve bayes, Decision tree, and combine classifies by ruining 20time for each individual input and get the average value of it. 2800 runs will be average in to140 recodes per each classifier accuracy. Then difference of accuracy in-between ensembled and base were calculated and recode the discrete statistics as follows.

	List Wise Deletion		Mean Mode Imputation	
	DIFF : NB & DT - NB	DIFF : NB & DT - DT	DIFF : NB & DT - NB	DIFF : NB & DT - DT
Mean	0.1431	0.1441	0.1434	0.1443
Standard Deviation	0.0493	0.0499	0.0493	0.0505
Kurtosis	2.8996	2.9977	2.7960	2.9802
Skewness	-1.9996	-2.0184	-1.9849	-2.0260
Range	0.1815	0.1840	0.1821	0.1870
Minimum	-0.0063	-0.0075	-0.0065	-0.0102
Maximum	0.1753	0.1766	0.1755	0.1768
Count (mean 20 sample)	140	140	140	140

TABLE 23: ENS, NB, DT COMPARISON

As per above table and its reading can be summarized as below.

Ensembled Classifier (NB &DT) and NB output accuracy difference mean value is 0.1431 and Standard deviation is 0.0493 in the List Wise Deletion and subject to the above describe inputs and conditions. And it has 2.8996 kurtosis and Skewness is (-1.9996). based on these two reading it has deviated from the normal distribution.

Ensembled Classifier (NB &DT) and DT output accuracy difference mean value is 0.1441 and Standard deviation is 0.0499 in the List Wise Deletion and subject to the above describe inputs and conditions. And it has 2.9977 kurtosis and Skewness is (-2.0184). based on these two reading it has deviated from the normal distribution.

Ensembled Classifier (NB &DT) and NB output accuracy difference mean value is 0.1434 and Standard deviation is 0.0493 in the Mean Mode Imputation and subject to the above describe inputs and conditions. And it has 2.7960 kurtosis and Skewness is (-1.9849). based on these two reading it has deviated from the normal distribution.

Ensembled Classifier (NB &DT) and DT output accuracy difference mean value is 0.1443 and Standard deviation is 0.0505 in the Mean Mode Imputation and subject to the above describe inputs and conditions. And it has 2.9802 kurtosis and Skewness is (-2.0260). based on these two reading it has deviated from the normal distribution.

Below set of graphs show the distribution of accuracy differences for NB, DT, NB & DT classifiers.

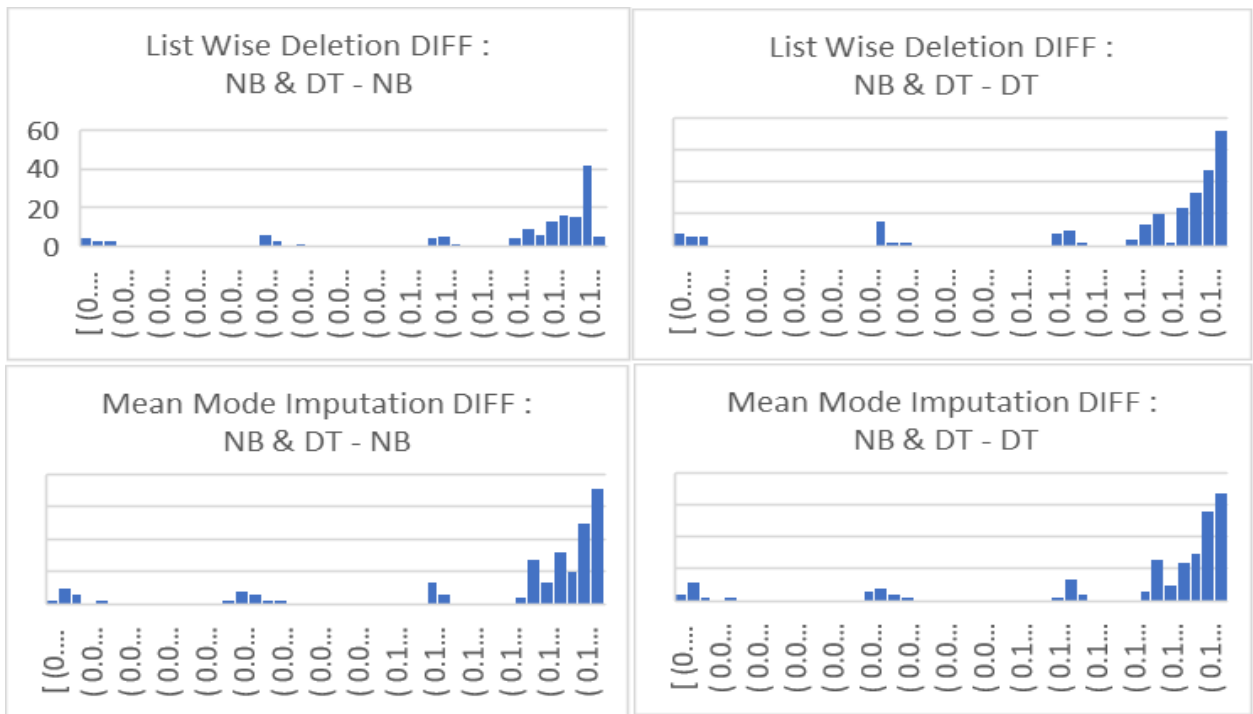


FIGURE 58: GRAPHICAL REPRESENTATION OF ACCURACY DIFFERENCE DISTRIBUTION NB&DT, NB, DT WITH MISSING VALUE

4.6.2 Decision tree and Support vector machine Ensembled classifiers

Decision tree and support vector machines classifiers ensembled and form an ensembled classifiers and it is necessary to identify what will the progress of the ensembled classifiers with respect to in base classifiers name decision tree.

X1(missing value): is varied form Listwise deletion and mean mode imputation

X2 outlier Factor: 0.2, 0.3, 0.4, 1.5

Feature selection: 10

With above for initiation input variable were considered as independent variable and accuracy was considered as output variable.

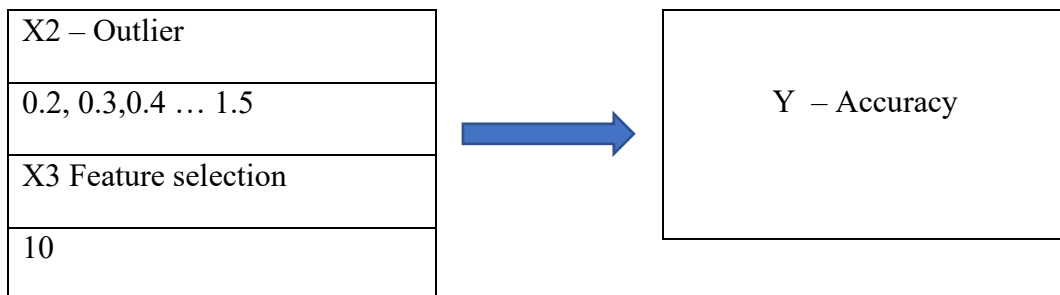


FIGURE 59:BASE DIAGRAM FOR DT, SVM ENS CLASSIFIERS

Accuracy was measured for Naïve bayes, Decision tree, and combine classifies by ruining 20time for each individual input and get the average value of it. 2800 runs will be average in to140 recodes per each classifier accuracy. Then difference of accuracy in-between ensembled and base were calculated and recode the discrete statistics as follows.

	List Wise Deletion		Mean Mode Imputation	
	DIFF : DT & SVM - DT	DIFF : DT & SVM - SVM	DIFF : DT & SVM - DT	DIFF : DT & SVM - SVM
Mean	0.1455	0.1443	0.1474	0.1461
Standard Deviation	0.0499	0.0500	0.0505	0.0506
Kurtosis	2.9977	3.0067	2.9802	2.9506
Skewness	-2.0184	-2.0215	-2.0260	-2.0178
Range	0.1840	0.1841	0.1870	0.1863
Minimum	-0.0061	-0.0073	-0.0071	-0.0077
Maximum	0.1779	0.1768	0.1798	0.1786
Count (mean 20 sample)	140	140	140	140

TABLE 24: ENS, DT, SVM COMPARISON

As per above table and it is reading can be summarized as below.

Ensembled Classifier (DT &SVM) and DT output accuracy difference mean value is 0.1455 and Standard deviation is 0.0499 in the List Wise Deletion and subject to the above describe inputs and conditions. And it has 2.9977 kurtosis and Skewness is (-2.0184). based on these two reading it has deviated from the normal distribution.

Ensembled Classifier (DT &SVM) and SVM output accuracy difference mean value is 0.1443 and Standard deviation is 0.0500 in the List Wise Deletion and subject to the above describe inputs and conditions. And it has 3.0067 kurtosis and Skewness is (-2.0215). based on these two reading it has deviated from the normal distribution.

Ensembled Classifier (DT &SVM) and DT output accuracy difference mean value is 0.1461 and Standard deviation is 0.0505 in the Mean Mode Imputation and subject to the above describe inputs and conditions. And it has 2.9802 kurtosis and Skewness is (-2-0260). based on these two reading it has deviated from the normal distribution.

Ensembled Classifier (DT &SVM) and SVM output accuracy difference mean value is 0.1443 and Standard deviation is 0.0506 in the Mean Mode Imputation and subject to the above describe inputs and conditions. And it has 2.506 kurtosis and Skewness is (-2.0178). based on these two reading it has deviated from the normal distribution.

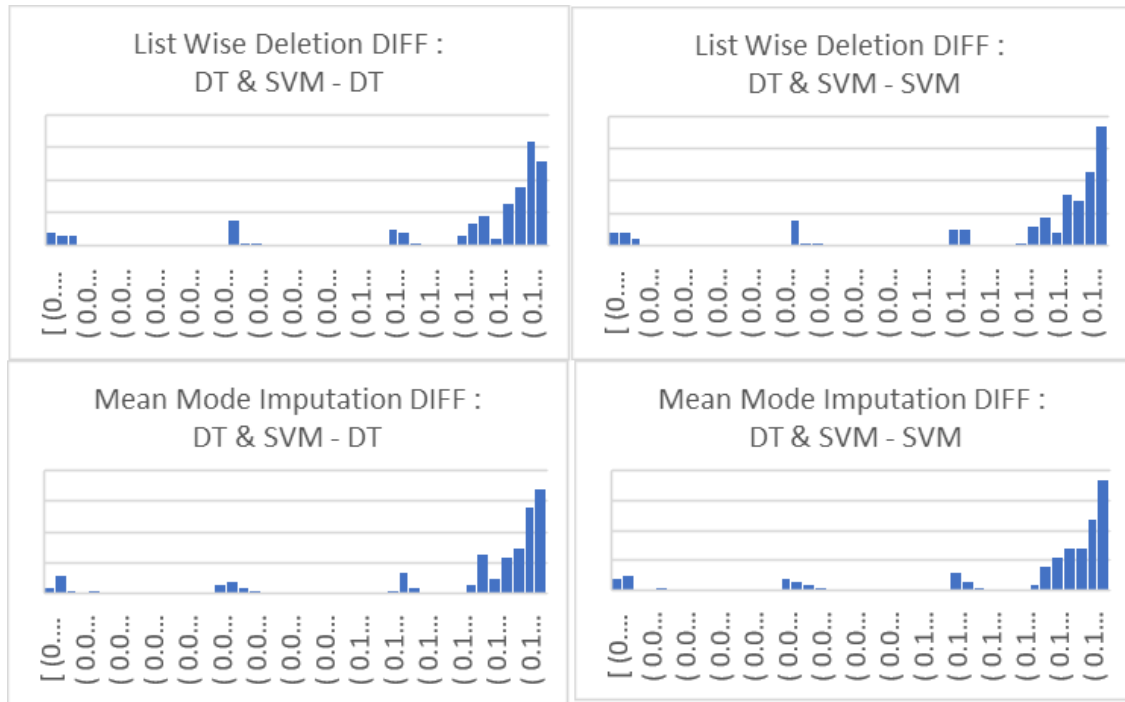


FIGURE 60:ACCURACY DIFFERENCE DISTRIBUTION DT& SVN, DT, SVM WITH MISSING VALUE

Below set of graphs show the distribution of accuracy differences for DT, SVM, NB & SVM classifiers.

4.6.3 Support vector machine and Naïve bayes Ensembled classifiers

Naïve bayes and Support vector machine classifiers ensembled and form an ensembled classifiers and it is necessary to identify what will the progress of the ensembled classifiers with respect to in base classifiers name Naïve bayes and Support vector machine.

X1(missing value): is varied form Listwise deletion and mean mode imputation

X2 outlier Factor: 0.2, 0.3, 0.4, 1.5

Feature selection: 10

With above for initiation input variable were considered as independent variable and accuracy was considered as output variable.

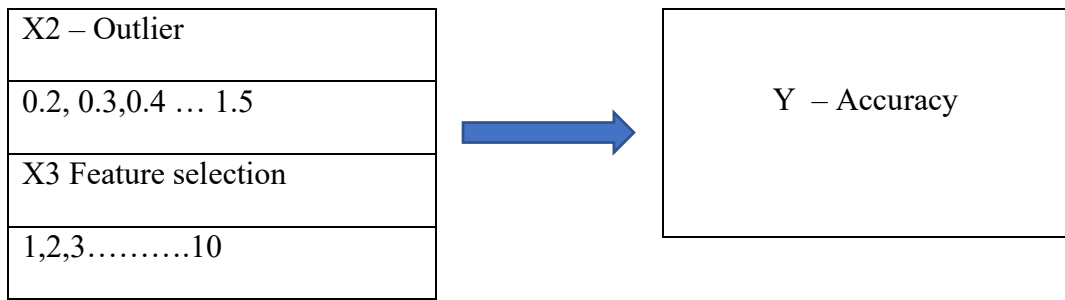


FIGURE 61: BASE DIAGRAM FOR NB, SVM ENS CLASSIFIERS

Accuracy was measured for Naïve bayes, support vector machine, and combine classifiers by running 20 times for each individual input and get the average value of it. 2800 runs will be average in to 140 records per each classifier accuracy. Then difference of accuracy in-between ensembled and base were calculated and recode the discrete statistics as follows.

	List Wise Deletion		Mean Mode Imputation	
	DIFF: SVM & NB - SVM	DIFF: SVM & NB - NB	DIFF: SVM & NB - SVM	DIFF: SVM & NB - NB
Mean	0.1442	0.1444	0.1472	0.1475
Standard Deviation	0.0500	0.0493	0.0506	0.0494
Kurtosis	3.0120	2.9112	2.9551	2.8062
Skewness	-2.0224	-2.0021	-2.0190	-1.9876
Range	0.1838	0.1818	0.1864	0.1820
Minimum	-0.0071	-0.0051	-0.0067	-0.0023
Maximum	0.1767	0.1767	0.1797	0.1797
Count (mean 20 sample)	140	140	140	140

TABLE 25: ENS, NB, SVM COMPARISON

As per above table and its reading can be summarized as below.

Ensembled Classifier (NB & SVM) and NB output accuracy difference mean value is 0.1442 and Standard deviation is 0.0500 in the List Wise Deletion and subject to the above describe inputs and conditions. And it has 3.0120 kurtosis and Skewness is (-2.0224). based on these two readings it has deviated from the normal distribution.

Ensembled Classifier (NB & SVM) and SVM output accuracy difference mean value is 0.1444 and Standard deviation is 0.0493 in the List Wise Deletion and subject to the above describe inputs and conditions. And it has 2.9112 kurtosis and Skewness is (-2.0021). based on these two readings it has deviated from the normal distribution.

Ensembled Classifier (NB &SVM) and NB output accuracy difference mean value is 0.1472 and Standard deviation is 0.0506 in the Mean Mode Imputation and subject to the above describe inputs and conditions. And it has 2.9551 kurtosis and Skewness is (-2.0190). based on these two reading it has deviated from the normal distribution.

Ensembled Classifier (NB &SVM) and SVM output accuracy difference mean value is 0.1494 and Standard deviation is 0.0494 in the Mean Mode Imputation and subject to the above describe inputs and conditions. And it has 2.8062 kurtosis and Skewness is (-1.9876). based on these two reading it has deviated from the normal distribution.

Below set of graphs show the distribution of accuracy differences for NB, SVM, NB & SVM classifiers.

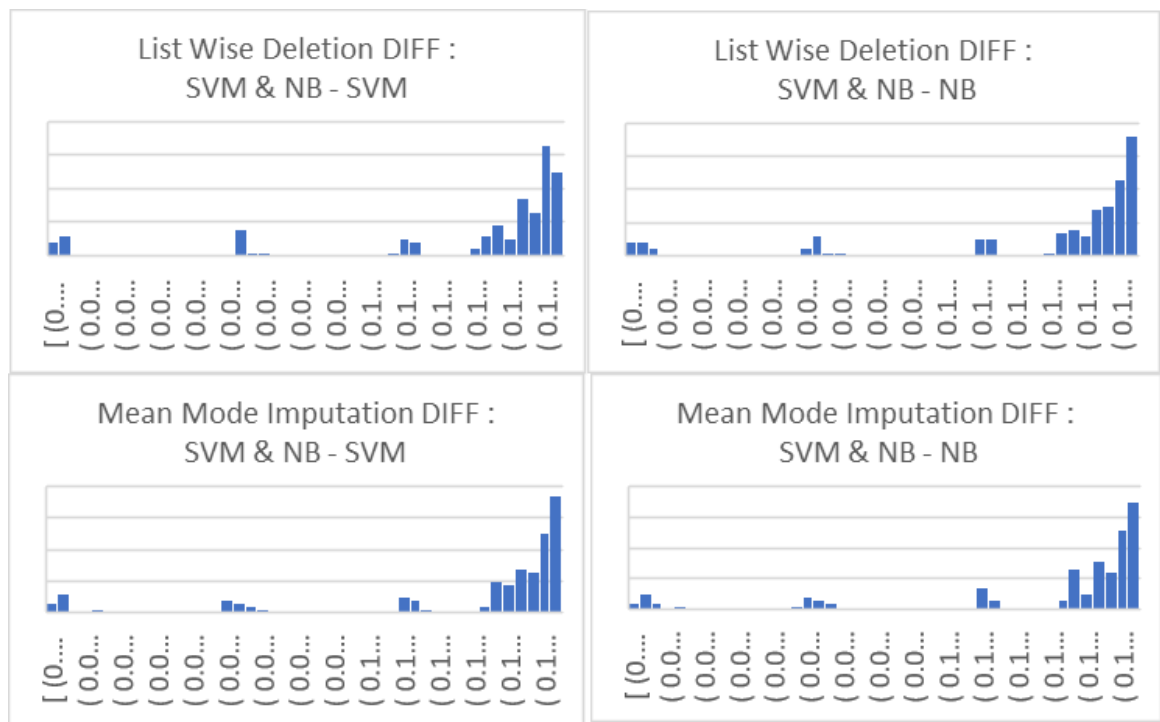


FIGURE 62:: ACCURACY DIFFERENCE DISTRIBUTION NB&SVM, NB, SVM WITH MISSING VALUE

4.7 Accuracy differences analysis

In this research work is focus on what is the best classifier which can predict the customer churn with highest accuracy. In here all reading are observed as the average value of 20 sample without changing the input conditions or parameters. All test has 280 records which has linked with 5600 samples.

4.7.1 Single Classifiers performance evaluation

AUC-ROC method

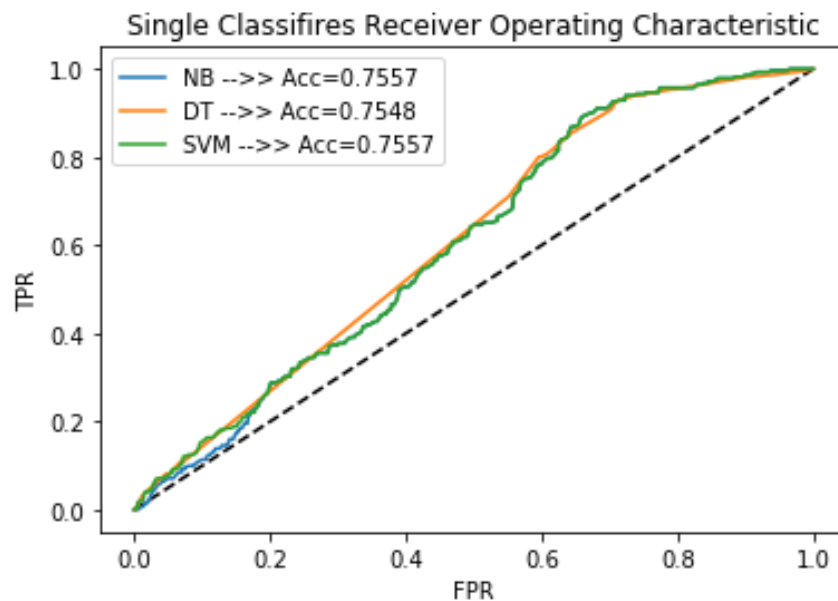


FIGURE 63:SINGLE CLASSIFIERS ROC CURVE

According to Figure 63 shows that AUC of the Naïve Bayes classifier, Decision Tree and Support vector machine as single classifiers AUC is higher than the 0.500 with. When $0.5 < AUC < 1$, there is a high chance that the classifier will be able to distinguish the positive class values from negative class values. It can explain as classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives. In this instance all three classifiers close accuracy but area under the AUC ROC curve is different. For Naïve Bayes it is 75.57% and Decision Tree 75.48% and Support vector machine 75.57%.

It is necessary to further analysis to decide on what is the best performing classifier with large set of output data.

Discrete statistic method

In this research there are 3 different classifiers used and there were 3 input independent variables with $2 \times 14 \times 10 = 280$ input condition and each has derived based on average value of 20 time run and gathered samples.

Input Conditions

Missing value	List Wise Deletion	Mean Mode Imputation
Out Lier Ration	0.20	1.50
No of Features	1	10

Number of Occurance Accuracy		
	No of times Maximum	No of times Minimum
DT	13	245
NB	234	35
SVM	33	
Grand Total	280	280

TABLE 26: INDIVIDUAL CLASSIFIERS COMPARISON

As per above table Naïve bayes classifier has better accuracy than the other two classifiers. based on number of times it has get the maximum accuracy of 234 out of 280. This is a point figure and what is the statistical behavior is presented under each classifier individual performance evaluation.

Graphical presentation can be shown in below.

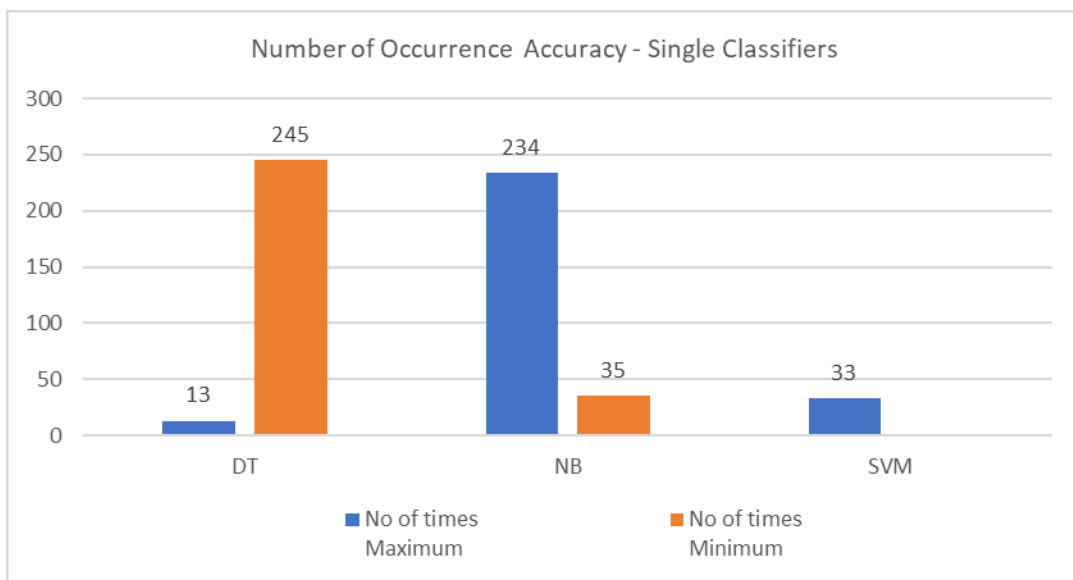


FIGURE 64: ACCURACY NUMBER OF TIME MAXIMUM AND MINIMUM

With respect to descriptive statistics analysis shown in the below table SVM has positive mean value and small standard deviation with respect to accuracy differences. SVM-NB

has mean value of 0.00037 and standard deviation of 0.00125. SVM-DT has mean value of 0.00132 and standard deviation of 0.00083

<i>Classifier</i>	<i>NB-DT</i>	<i>SVM-NB</i>	<i>SVM-DT</i>
Mean	0.00095	0.00037	0.00132
Standard Deviation	0.00146	0.00125	0.00083
Kurtosis	7.42916	16.82210	1.18087
Skewness	-2.20806	3.99172	-0.24673
Range	0.01059	0.00808	0.00500
Minimum	-0.00692	0.00000	-0.00133
Maximum	0.00367	0.00808	0.00367
Count	280	280	280

FIGURE 65: SINGLE CLASSIFIER ACCURACY DIFFERENCE COMPARISON

4.7.2 Ensembled Classifiers performance evaluation

AUC-ROC method

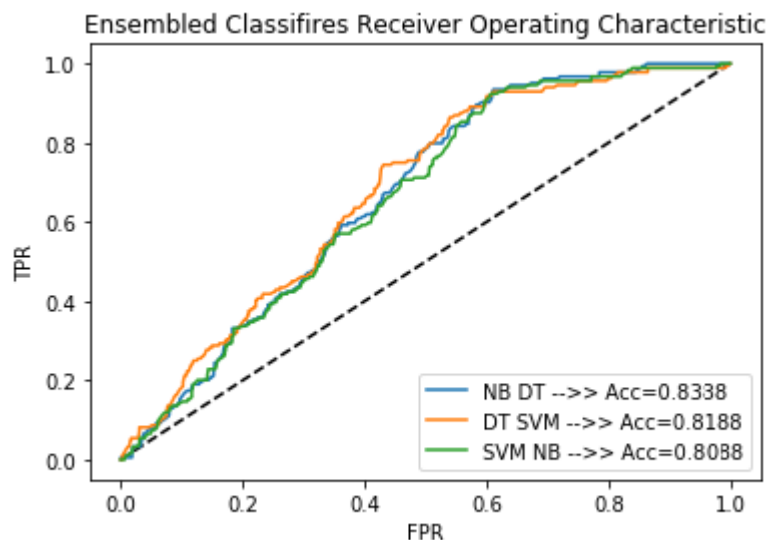


FIGURE 66: ENSEMBLED CLASSIFIER ROC CURVE

According to Figure 66 shows that AUC of the Naïve Bayes classifier, Decision Tree and Support vector machine were combined as ensembled classifiers AUC is higher than the 0.500 with. When $0.5 < AUC < 1$, there is a high chance that the classifier will be able to distinguish the positive class values from negative class values. It can explain as classifier is able to detect more numbers of True positives and True negatives than False negatives

and False positives. AUC ROC curve is different. For Naïve Bayes and Decision Tree ensemble classifier is 83.38%. Decision Tree and Support vector machine ensemble classifier is 81.88%. Support vector machine and Naïve Bayes ensemble classifiers 80.038%.

With respect to single classifiers all ensemble classifiers has better performance. It is necessary to further analysis to decide on what is the best performing classifier with large set of output data.

Discrete statistic method

In this research there are 3 different ensemble classifiers used and there were 3 input independent variables with $2 \times 14 \times 10 = 280$ input condition and each has derived based on average value of 20 time run and gathered samples.

Input Conditions

Missing value	List Wise Deletion	Mean Mode Imputation
Out Lier Ration	0.20	1.50
No of Features	1	10

Number of Occurance Accuracy		
	No of times Maximum	No of times Minimum
NB & DT	22	90
DT & SVM	90	22
SVM & NB	0	0
All 3 are Same	168	168
Grand Total	280	280

TABLE 27: ENSEMBLED CLASSIFIERS COMPARISON

As per above table DT& SVM classifier has better accuracy than the other two classifiers. based on number of times it has get the maximum accuracy of 90 out of 280. This is a point figure and what is the statistical behavior is presented under each classifier individual performance evaluation.

Graphical presentation can be shown in below.

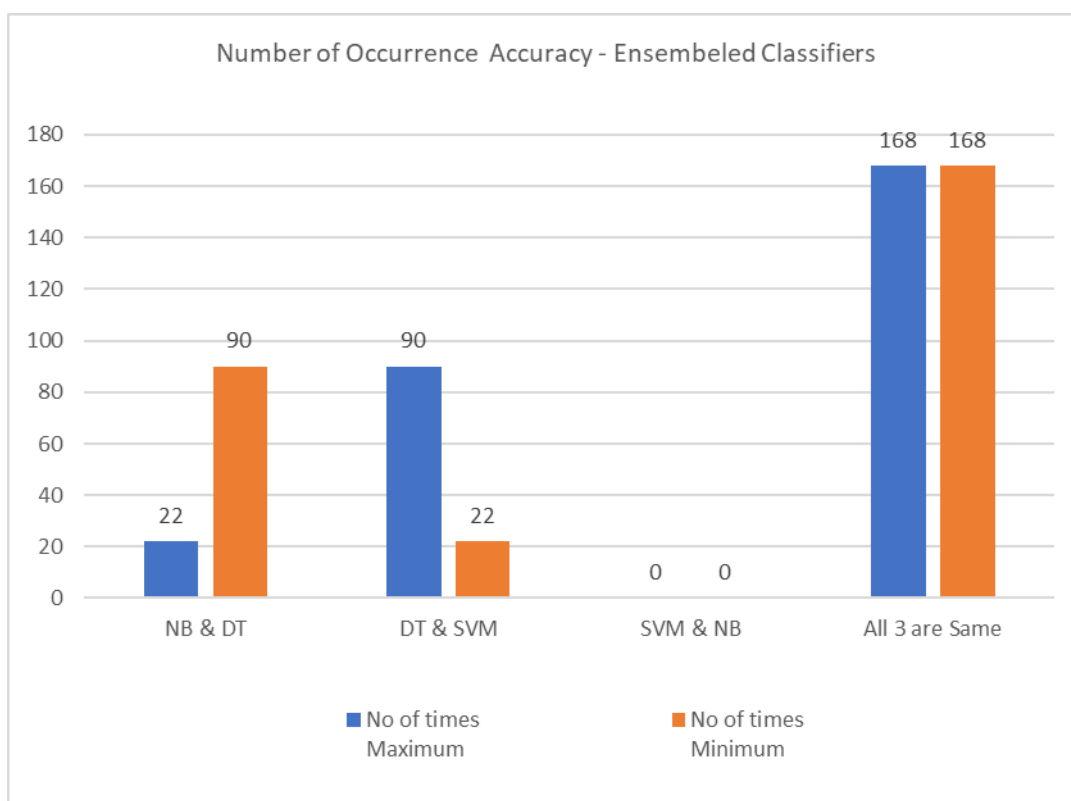


FIGURE 67: ACCURACY NUMBER OF TIME MAXIMUM AND MINIMUM

With respect to descriptive statistics analysis shown in the below table SVM has positive mean value and small standard deviation with respect to accuracy differences. SVM-NB has mean value of 0.00037 and standard deviation of 0.00125. SVM-DT has mean value of 0.00132 and standard deviation of 0.00083

Classifier	DT & SVM - NB & DT	SVM & NB - NB & DT	DT & SVM - SVM & NB
Mean	0.00118	0.00060	0.00099
Standard Deviation	0.00064	0.00034	0.00071
Kurtosis	16.65517	8.63517	5.16677
Skewness	3.99914	1.40366	1.39540
Range	0.00439	0.00670	0.00945
Minimum	-0.00036	-0.00086	-0.00099
Maximum	0.00404	0.00585	0.00846
Count	280	280	280

FIGURE 68: ENSEMBLED CLASSIFIER ACCURACY DIFFERENCE COMPARISON

05. Chapter – Conclusion and recommendations

5.1 Conclusion

The worldwide telecommunication system has standardized by international Telecommunication Standardization (ITU) and all logs and customer data categories are in the same format. Cell2Cell data also sets CRM and Billing data provided by standardized telecommunication systems and this research work is built on based on that model. However, due to ITU standardization, all telecommunication operators' systems in Sri Lanka can generate the same dataset with local information and developed models can predict local customer churn behaviors.

The data set has 34 numerical variables and 33 Boolean variables (attributes), and it also can be categorized into main five categories. Under the feature selection methods in the research, it has evaluated the similarities (covariance) among the variables and what is the relationship with the dependent variable. Based on that it can be concluded that contributions out of all variables less than 10 have significant relationships irrespective of classifiers subjected in this research.

Data set has over 71,047 records which was a considerably larger data set which was positively contributed based on size for the evaluation and analysis process to proceed in a forward way without cross validation. This input data set shows almost similar behavior for missing value processing list wise deletion and mean mode imputation under the pre data processing due to its size.

Naïve Bayes classifier has shown individual overall accuracy performance around 74.28% irrespective of the independent variable constraints such as missing value processing, outlier processing and feature selection and this accuracy figure has derived from as a mean value of experimental result of 5600 time runs. It means 25.72% of the prediction will be in-accurate and when telecommunication operators are used this classifier needs to consider effort vs outcome based on this accuracy rate.

Decision tree classifier has shown individual overall accuracy performance around 74.19% irrespective of the independent variable constraints such as missing value processing, outlier processing and feature selection and this accuracy figure has derived from as a mean value of experimental result of 5600 time runs. It means 25.81% of the prediction

will be in-accurate and when telecommunication operators are used this classifier needs to consider effort vs outcome based on this accuracy rate.

Support vector machine classifier has shown individual overall accuracy performance around 74.03% irrespective of the independent variable constraints such as missing value processing, outlier processing and feature selection and this accuracy figure has derived from as a mean value of experimental result of 5600 time runs. It means 25.97% of the prediction will be in-accurate and when telecommunication operators are used this classifier needs to consider effort vs outcome based on this accuracy rate.

There are three ensembled classifiers used by combining Naïve Bayes (NB), Decision Tree (DT) and Support vector machine (SVM) Classifiers. Ensembled classifier concerned DT & SVM classifier has the lesser performance out of three ensembled classifiers.

Ensemble Classifier (NB &DT) has performance difference mean value is 0.14 and Standard deviation is 0.04 with respect to NB and DT classifiers. It shows that Ensemble Classifier (NB &DT) is better than it's building classifiers.

Ensemble Classifier (DT & SVM) has performance difference mean value is 0.14 and Standard deviation is 0.04 with respect to DT and SVM classifiers. It shows that Ensemble Classifier (DT & SVM) is better than it's building classifiers.

Ensemble Classifier (SVM & NB) has performance difference mean value is 0.14 and Standard deviation is 0.05 with respect to SVM and NB classifiers. It shows that Ensemble Classifier (SVM & NB) is better than it's building classifiers.

When each ensembled classifies and it is constructed than the base classifiers, ensemble Classifiers have better accuracy. It can prove with all the case differences that it has a positive mean value. As per results it will be more appropriate to use ensembled classifiers than it based individual classifiers.

Above all concluded points are highly dependent with the dataset used for this research work. But telecommunication operators are used and maintain the same set of attributes that can be generalized to Sri Lanka telecommunication operators in that contest as mentioned above.

5.2 Future Research Areas

5.2.1 Accuracy evaluation methodology

There is no algorithm is over all best in all domains (Almana, 2014) it can applied.. Different learning algorithms performances of the classifiers are compared on this target data set (Gutkin, 2008). What are the parameters or attribute of defining a better classifier need to be assess here? Appropriate properties are generality confidence level and accuracy of prediction. Confusion denotes the four results which classifier is applied on a set of instances.

		Actual Class	
		p (+)	n (-)
Hypothesized class	p (+)	True Positive	False Positive
	n (-)	False Negative	True Negative
Column totals		P	N

FIGURE 69:THE CONFUSION MATRIX

Based on a confusion matrix, the most used for evaluation metrics are overall accuracy, true positive rate, and false positive rate. The overall accuracy (OA) is the ratio of the correctly classified instances.

$$OA = \frac{TP + TN}{P + N}$$

01. Researcher is proposing that this research result is totally based on the Overall accuracy and it can be evaluated with the other parameters in the confusion matrix such as TP rate FP rate.

It classifies positive instances with weak evidence so it achieve high TP rate but also consist with many FP mistakes (Fawcett, 2006) (Svendsen, 2013). When the classifier performance is plotted on ROC graph with value of θ varied from 0 to 1, a ROC curve will be plotted on it. It demonstrates the trade-off between TP rate and FP rate.

02. Researcher is proposing that ROC curve will give the better interpretation on trade-off between TP rate and FP rate. But as per the user requirements it should be evaluated

5.2.2 Ensembled classifier enhancement

In this research all the ensembled classifier formation was based on voting method. It means serial usage of the classifiers and try to improve the prediction accuracy. But there is some other method of formation of ensemble classifiers such as

01. Averaging Based Ensemble Methods
02. Bootstrap Aggregating

References

- Ahmed, I. N. M. U. A. S. M. A. N. & R. U., 2010. *A mediation of customer satisfaction relationship between service quality and repurchase intentions for the telecom sector in Pakistan: A case study of university students.*, s.l.: s.n.
- Ahn, H. P. H. S. & S. L. Y., 2006. Customer churn analysis: churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry.. *Telecommunications Policy*, , pp. 30, 552-568.
- AHN, J. H. S. & L. Y. ..., 2006. Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry.. *Telecommunications Policy*, , pp. 10, 552-568.
- Ali, J. A. I. R. K. Y. A. S. N. & A. H., 2010. Determinants of consumer retention in cellular industry in Pakistan.. *African Journal of Business Management*, , pp. 4(12), 2402–2408..
- Almana, A. A. M. & A. R., 2014. A survey on data mining techniques in customer churn analysis for telecom industry.. *International Journal of Engineering Research and Applications*, pp. 4(5), 165-171..
- Almossawi, M., 2012. Customer satisfaction in the mobile telecom industry in Bahrain: Antecedents and consequences. . *International Journal of Marketing Studies*, 4(6), , pp. 139-156. .
- Ascarza, E. I. R. a. S. M., 2016. The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment.. *Journal of Marketing Research*, 53(1):46–60.
- Banker, R. P. G. & S. D., 2000. An empirical investigation of an incentive plan that includes nonfinancial performance measures.. *The Accounting Review*., pp. 75(1), 65-92.
- Barrett, J., 2003. *US mobile market intelligence.* .. s.l.:Dallas, TX: Parks Associates.
- Berson, A. S. S. & T. K., 2000. Building Data Mining Applications for CRM. . In: New York, NY: McGraw-Hill.: s.n.
- Birke, D. & S. P., 2006. Network effects and the choice of mobile phone operator.. *Journal of Evolutionary Economics*, pp. 16(1-2), 65-84..
- Bolton, R. K. P. & B. M., 2000. Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the Academy of Marketing Science*, , pp. 28(1), 95-108..
- Bolton, R. L. K. & V. P., 2004. The theoretical underpinnings of customer asset management: a framework and propositions for future research.. *Journal of the Academy of Marketing Science*, pp. 32(3), 271-292.
- Boohene, R. & A. G., 2011. Analysis of the determinants of customer loyalty of telecommunication industry in Ghana: The case of Vodafone (Ghana).. *International Business Research*, pp. 4(1), 229-240.

- Brown, K., 2004. Holding onto customers. *Wireless Week*, . In: s.l.:s.n., p. 15(6)..
- Burdiek, M., 1993. Strategic network management. *Cellular Business*, . In: s.l.:s.n., p. 10(1)..
- Butt, H., 2011. Measuring customer satisfaction w.r.t restaurant industry in Bahawalpur.. *European Journal of Business Management*, , pp. 3(5), 54-64.
- Cabral, L., 2000. Stretching firm and brand reputation.. *The Rand Journal of Economics, Mount Morris*, , pp. 31(4), 658-674.
- Cai, G. D. O. O. a. S., 2018. A Hybrid Churn Prediction Model in Mobile Telecommunication Industry. *International Journal of e-Education, e-Business, e-Management and e-Learning*, , Vol. 4, No. 1(February 2014).
- Chandrashek, M. R. K. T. S. & G. R., 2007. Satisfaction strength and customer loyalty.. *Journal of Marketing Research*, , pp. 44, 153-163.
- Chandrashek, M. R. K. T. S. & G. R., 2007. Satisfaction strength and customer loyalty.. *Journal of Marketing Research*, , pp. 44, 153-163.
- CHIANG, D. W. Y. L. S. & L. C., 2003. Goal-oriented sequential pattern for network banking and churn analysis.. In: *Expert systems with applications*. s.l.:s.n., pp. 25, 293-302..
- Chitra, P. U. H. M. V. K. D. & D. D., 2013. Prediction of subscriber churn using social network analysis.. *Bell Labs Technical Journal*, 17(4), 63–75..
- Chou, C. & C. S., 2006. *Factors affecting China Mobile Customer satisfaction*. Retrieved, s.l.: s.n.
- Clement Kirui, L. H. W. C. a. H. K., 2013. Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining IJCSI. *International Journal of Computer Science Issues*, , Vol. 10, (Issue 2, No 1, March 201).
- Colgate, M. & H. R., 2001. An investigation into the switching process in retail banking services.. *International Journal of Bank Marketing*, , pp. 19(5), 201-212.
- Coussement, K. & V. d. P. D., 2008. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers.. *Expert Systems with Applications*, , Volume 36(3), Part 2, 6127–6134..
- DATTA, P. M. B. M. D. R. & L. B., 2001. Automated cellular modelling and prediction on a large scale.. 485-502(Issues on the application of data mining.).
- Donnavieve, N. & S. K., 2002. Flow and internet shopping behaviour: a conceptual model and research proposition. .. *Journal of business research*, , pp. 57(10), 199-208.
- Dover, H. & M. B., 2006. Asymmetric effects of dynamic usage behaviour on duration in subscription-based online service.. *Journal of Interactive Marketing*, , pp. 20, 5-15.
- Dover, H. & M. B., 2006. Asymmetric effects of dynamic usage behaviour on duration in subscription-based online service.. *Journal of Interactive Marketing*, , pp. 20, 5-15.

- Eshghi, A. K. S. & G. H., 2008. Service quality and customer satisfaction: An empirical investigation in Indian mobile telecommunications services. . *Marketing Management Journal, fall* , pp. 119-144.
- Foxall, G. & G. R., 1997. *Consumer psychology for marketing.*, New York: Routledge: s.n.
- Ganesh, J. R. K. & A. M., 2000. Understanding the customer base of service providers: an examination of the differences between switchers and stayers.. *Journal of Marketing* , pp. 64(3), 65-87..
- Garrette, B. & A. K., 2010. Challenges in marketing socially useful goods to the poor.. *California Management Review*, pp. 52(4), 29-47..
- Gerpott, T. R. W. & S. A., 2001. Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market.. *Telecommunications Policy* , pp. 25(4), 249-269.
- Gurjeet Kaur, R. & M. S., 2012. Exploring customer switching intentions through relationship marketing paradigm.. *International Journal of Bank Marketing* , pp. 30(4), 280-302. <http://dx.doi.org/10.1108/02652321211236914>.
- Gutkin, M., 2008. *Feature selection methods for classification of gene expression profiles.* Tel-Aviv University, s.l.: s.n.
- HAHM, J. C. W. & Y. J. W., 1997. *A Strategic Approach to Customer*, s.l.: s.n.
- Hansemark, O. & A. M. ., 2004. Customer satisfaction and Retention: The experience of Individual Employees.. *Managing Service Quality*, p. 14(40).
- Haque, A. R. S. & R. M., 2010. Factors determinants the choice of mobile service providers: Structural Equation Modeling approach on Bangladeshi consumers.. *Business and Economics Research Journal*, pp. 1(3), 17-34.
- Hargreaves, C. A., 2019. A Machine Learning Algorithm for Churn Reduction & Revenue Maximization: An Application in the Telecommunication Industry. *International Journal of Future Computer and Communication*, Vol. 8, No. 4, (December 2019).
- Hashmi, N. B. N. & I. M., 2013. Customer churn prediction in telecommunication: A decade review and classification.. *IJCSI International Journal of Computer Science* , Issue Issues, 10(5.2), pp. 271-282.
- Hashmi, N. B. N. & I. M., 2013. Customer churn prediction in telecommunication: A decade review and classification.. *IJCSI International Journal of Computer Science Issues*, 10(5.2), , pp. 271-282..
- Herrmann, A. X. L. M. K. & H. F., 2007. The influence of price fairness on customer satisfaction: an empirical test in the context of automobile purchases.. *Journal of Product & Brand Management* , pp. 16(1), 49–5.
- HSIEH, N., 2004. An Integrated Data Mining and Behavioural Scoring Model for Analysing Bank Customers. . In: *Expert systems with applications*,. s.l.:s.n., pp. 27, 623-633..

- HSU, S. H., 2008. Developing an index for online customer satisfaction: Adaptation of American Customer Satisfaction Index. In: . *Expert Systems with Applications*,. s.l.:s.n., pp. 34, 3033-3042.
- Huan, B. K. M. & B. B., 2012. Customer churn prediction in telecommunications. *Expert Systems with Applications*,. In: s.l.:s.n., pp. 39, 1414–1425.
- HUNG, S. Y. D. C. & W. H., 2006. Applying data mining to telecom churn management.. *Expert Systems with Applications*, , pp. 31, 515-524.
- Hunt, E. B. M. J. & S. P. J., 1966. Experiments in induction. Academic Press.
- Hwang, H. J. T. & S. E., 2004. . *An LTV model and customer segmentation based on customer value*:, s.l.: case study on the wireless telecommunication industry. *Expert Systems*.
- J. Friedman, T. H. a. R. T., 2008. The Elements of Statistical Learning Data Mining, Inference and Prediction. . In: s. edition, ed. s.l.:Springer, Stanford, California USA, .
- Jahanzeb, S. & J. S., 2007. Churn management in the telecom industry of Pakistan: A comparative study of Ufone and Telenor.. *Journal of Database Marketing & Customer Strategy Management*, pp. , 14, 120–129. i:10.1057/palgrave.dbm.3250043.
- JENAMANI, M. M. P. K. J. & G. S., 2003. *A Stochastic model of e-customer behaviour. Electronic commerce research and applications*. s.l.:s.n.
- Karatzoglou, D. M. a. K. H. ..., 2013. Support vector machines in. *R. Journal of Statistical Software*.
- Keramati, A. & A. S., 2011. Churn analysis for an Iranian mobile operator.. *Telecommunications Policy*, 35(2011), 344–356. doi:10.1016/j.telpol.2011.02.009.
- Keramati, A. & A. S., 2011. Churn analysis for an Iranian mobile operator. *Telecommunications Policy*, 35(2011), doi:10.1016/j., Issue telpol. 2011.02.009, p. 344–356..
- Kim, M. & J. D., 2004. The effects of customer satisfaction and switching barriers on customer loyalty in Korean mobile telecommunication services.. *Telecommunications Policy*, pp. 28(2), 145-159..
- KITAYAMA, M. M. R. & I. Y., 2002. *Data mining for customer load profile analysis. Transmission and Distribution Conference and Exhibition 2002*.. s.l., Asia Pacific, IEEE/PES.
- Kollmann, T., 2000. The price/acceptance function: Perspectives of a pricing policy in European telecommunication markets.. *European Journal of Innovation Management*, pp. 3(1), 7-14.
- Kon, M., 2004. *Stop customer churn before it starts*. , s.l.: Harvard Management Update, 9(7), 3-5.
- Kotler, P. & K. K., 2009. *Marketing Management (13th Ed.)*., London: Prentice Hall.: s.n.

- Leelakulthanit, O. & H. B., 2011. Factors that impact customers' satisfaction: Evidence from the Thailand mobile cellular network industry.. *International Journal of Management and Marketing Research*,, pp. 4(2), 67-76..
- Li-Shang Yang, C. C., 2006. *Knowledge Discovery on Customer Churn Prediction Proceedings of the 10th WSEAS Interbational Conference on APPLIED MATHEMATICS*,. Dallas, Texas, USA, November 1-3, 2006, s.n.
- LIU, D. & S. Y., 2004.) Integrating AHP and Data mining for Product Recommendation Based on Customer Lifetime Value. . In: *Information & Management*,. s.l.:s.n., pp. 42, 387-400.
- Lu, 2003. *Modeling customer lifetime value using survival analysis: An application in the telecommunications industry. Data Mining Techniques*,. s.l.:SUGI 28..
- MADNICK, S. & Z. H. .., 2006. Improving data quality through effective use of data semantics. . In: *Data & Knowledge Engineering*. s.l.:s.n., pp. , 59, 460-475.
- Malhotra, A. & M. C. ..., 2013. Exploring switching behavior of US mobile service customers.. *Journal of Services Marketing*,, pp. 27(1), 13-24.
- Malik, M. G. M. & I. H., 2012. Impact of Brand Image, Service Quality and price on customer satisfaction in Pakistan Telecommunication sector. . *International Journal of Business and Social Science*, , pp. 3(23), 123-129.
- MA, M. L. Z. & C. J., 2008. Phase-type distribution of customer relationship with Markovian response and marketing expenditure decision on the customer lifetime value.. *European Journal of Operational Research*, pp. 187, 313-326..
- Mattison, R., 2006. *Telecom churn management: The golden opportunity*.. s.l.:APDG Publishing..
- MEYER-BASE, A. & W. R., 1998. Transformation Radial Basis Neural Network for Relevant Feature Selection.. *Pattern Recognition Letters*, , pp. 19, 1301-1306.
- Min, D. & W. L., 2009. Switching factors of mobile customers in Korea.. *Journal of Service Science*,, pp. 1, 105-120.
- Mirowski, S. C. F. H. a. M. M., 2008. Support vector ma-chines. <http://www.cs.nyu.edu/~yann/2010f-G22-2565-001/diglib/lecture03a-svm-2010.pdf>. In: s.l.:s.n.
- Mitchell, T. M., 1997. *Machine Learning. McGraw-Hill Science/Engineering/Math*.. s.l.:s.n.
- Mittal, V. K. P. & T. M., 1999. Attribute-level performance, satisfaction, and behavioral intentions over time: A consumption-system approach.. *Journal of Marketing*.
- Motley, B. L., 2003. How to thrill your customer?. *Journal of Marketing*, , p. 35(50)..
- Munnukka, J., 2008. Customers' purchase intentions as a reflection of price perception.. *Journal of Product & Brand Management*, pp. 17(3), 188-196.
- NBRI, 1982. (*National Business Research Institute, Inc.*). Retrieved from http://www.nbrii.com/Create_Surveys/Survey_Definitions.html, s.l.: s.n.

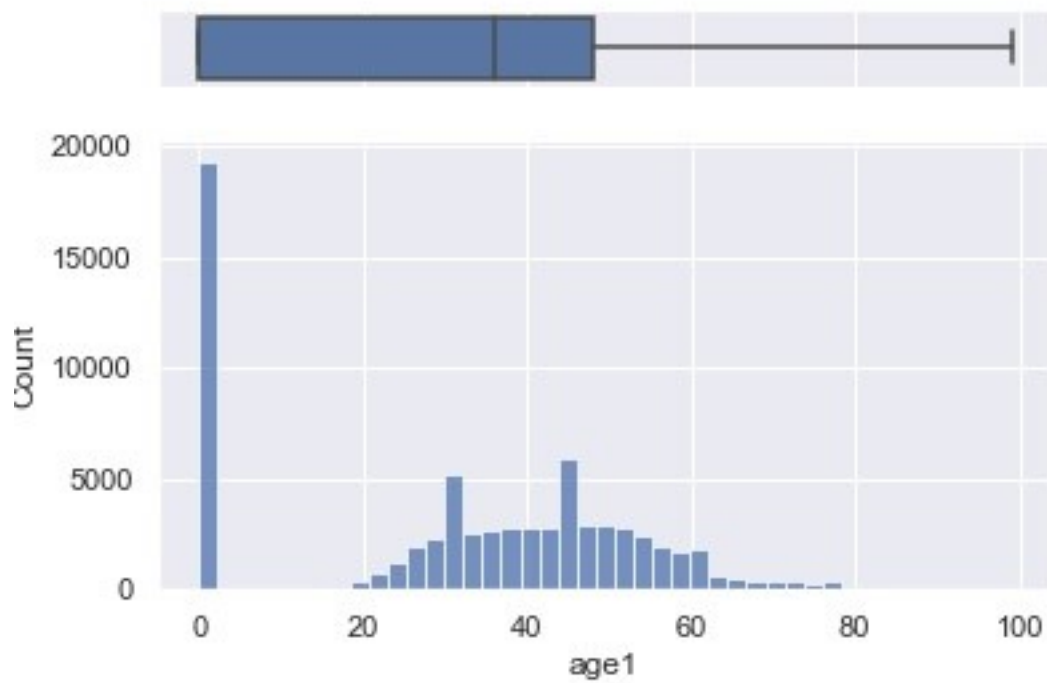
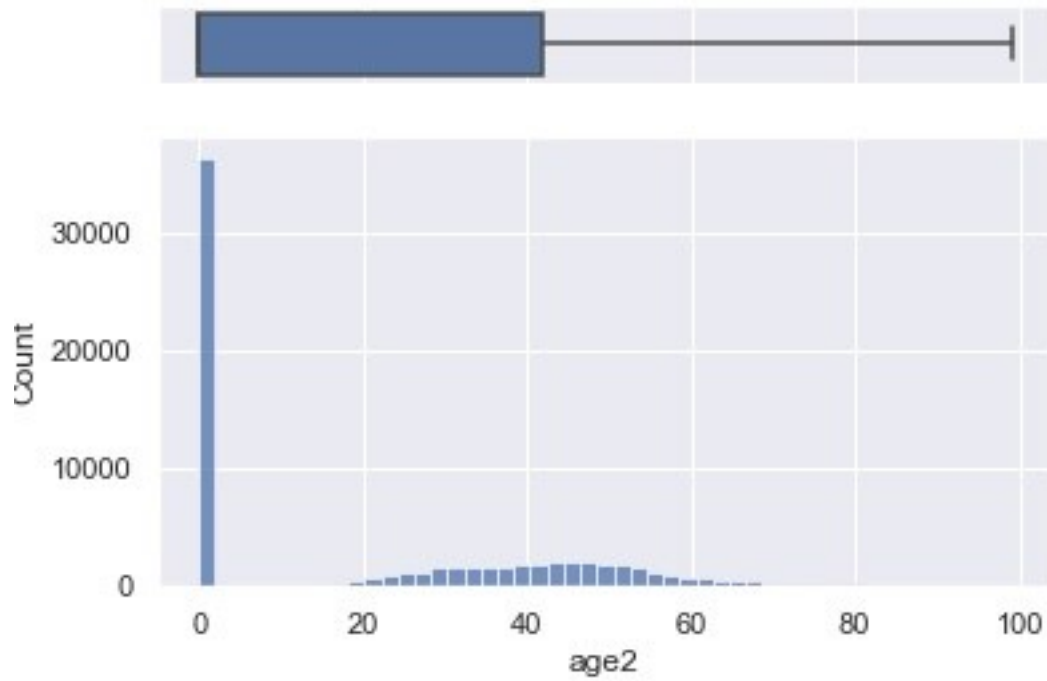
- Ng, K. & L. H., 2000. Customer retention via data mining. *Artificial Intelligence Review*. In: s.l.:s.n., pp. 14(6), 569-590..
- Nguyen, N. & L. G., 2001. Corporate image and corporate reputation in customers' retention decisions in services.. *Journal of Retailing and Consumer Services*, , pp. 8, 227-236.
- Nimako, S. A. F. & D. F., 2010. Overall Customer Satisfaction in Ghana's Mobile Telecommunication Networks: Implications for Management and Policy.. *ATDF Journal*, , pp. 7(3/4), 35-49.
- Oladapo, J. O. O. & A., A. O., 2018. Predictive Analytics for Increased Loyalty and Customer Retention in Telecommunication Industry. *International Journal of Computer Applications (0975 – 8887)*, Volume 179 – No.32,(April 2018).
- Oliver, R., 1997. *Satisfaction: A behavioral perception on the consumer.*, New York: McGraw Hill: s.n.
- Olsen, L. & J. M., 2003. Service equity, satisfaction, and loyalty: From transaction-specific to cumulative evaluations.. *Journal of Service Research*, pp. 5, 184-197.
- Paulrajan, R. a. R. H., 2011. Service quality and customers preference of cellular mobile service providers.. *International Journal of Management*, , Issue doi: <http://dx.doi.org/10.4067/S0718-27242011000100004>, pp. 6(1), 38-45. .
- Pawar, R. J. J. U. T., 2011. Churn Prediction in Telecommunication Using Data Mining Technology (IJACSA). *International Journal of Advanced Computer Science and Applications*, Vol. 2, No.2(February 2011).
- Peng, H. L. F. a. D. C., 2005. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy.. *IEEE T Pattern Anal.* , pp. 27(8) 1226-1238.
- PRINZIE, A. & V. D. P. D., 2004. Investigating purchasing-sequence patterns for financial services using Markov, MTG and MTGg models. .. *European journal of operational research*, , pp. 170, 710-734.
- Quinlan, J. R., 1986. Induction of Decision Trees. *Machine Learning* , 1, In: s.l.:s.n., pp. 81-106..
- Rahman, H. ..., 2014. Factors affecting customer satisfaction in mobile telecommunication industry in Bangladesh.. *Business, Management and Education*, , pp. 12(1), 74-93.
- Rahman, S. H. A. & A. M., 2011. Choice criteria for mobile telecom operator: Empirical investigation among Malaysian customers.. *International Management Review*, pp. 7(1), 50-57.
- RYALS, L., 2002. Are your Customers Worth More Than Money?. *Journal Of Retailing and Consumer Services*,, pp. 9, 241-251.
- Sattar, M. & S. B. ..., 2012. Customer Satisfaction Affects the Customer Loyalty: Evidence from Telecommunication Sector in Pakistan.. *Asian Journal of Business Management*, , pp. 4(3), 252-259.

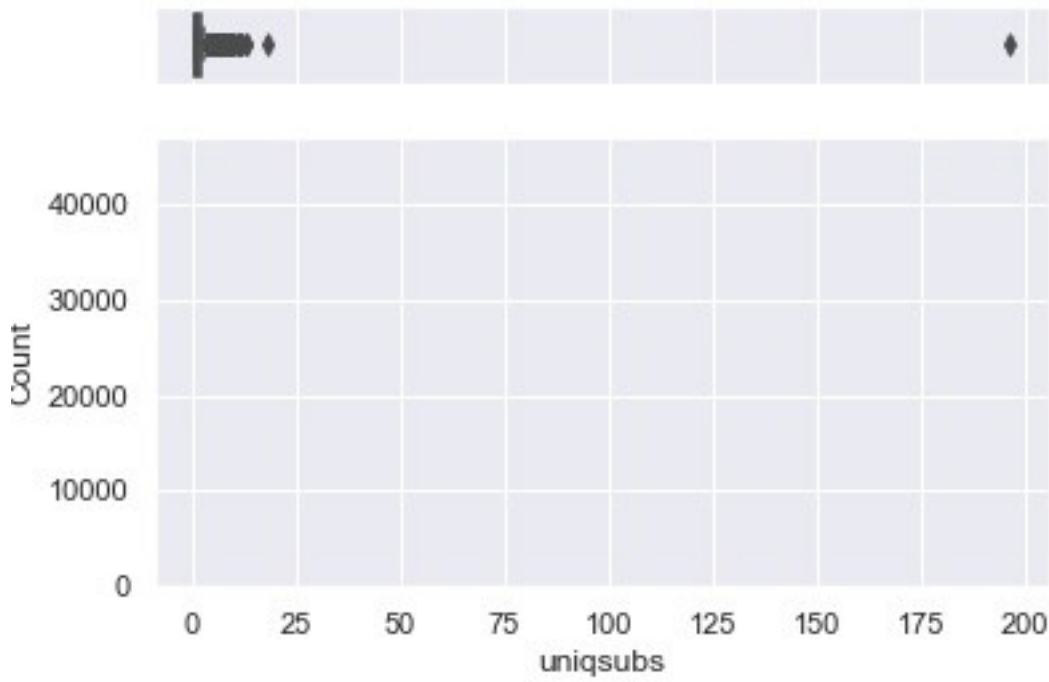
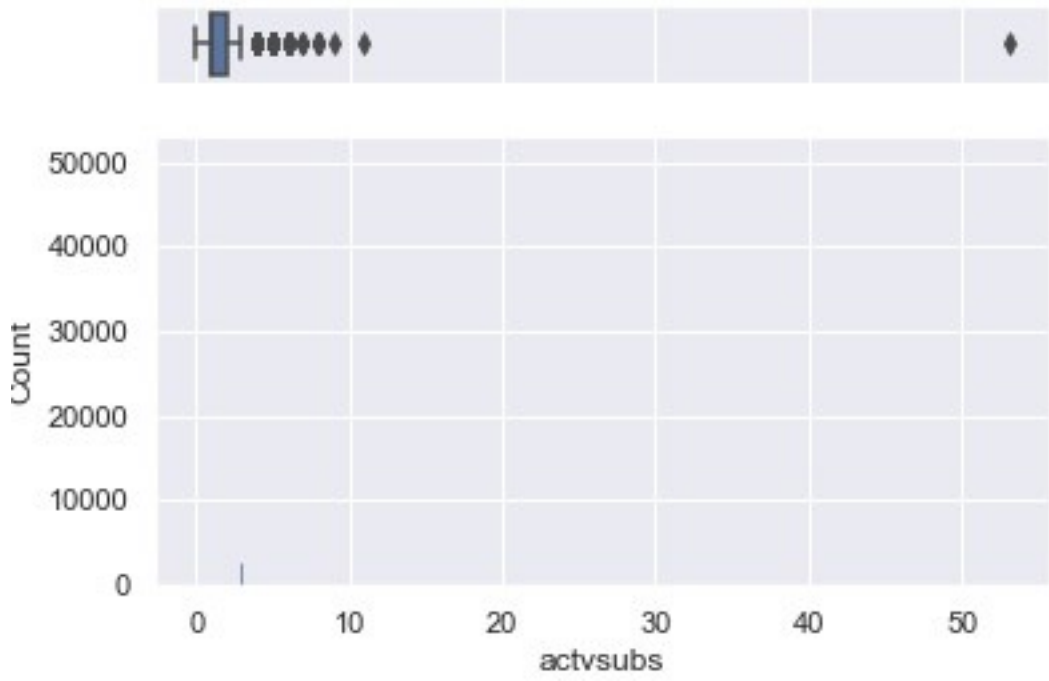
- Seo, D. R. C. & B. Y., 2008. Two-level model of customer retention in the US mobile telecommunications service market.. *Telecommunications Policy*, pp. 32(3–4), 182–196.
- Seo, D. R. C. & B. Y., 2008. Two-level model of customer retention in the US mobile telecommunications service market.. *Telecommunications Policy*, pp. 32(3–4), 182–196.
- Siddiqui, K., 2011. Personality influences customer switching.. *International Journal of Contemporary Research in Business*, , pp. 2(10), 363-372..
- Sindhu M E and Vijaya, 2015. Predicting Churners in Telecommunication Using Variants of Support Vector Machine. *American Journal of Engineering Research (AJER)*, Issue e-ISSN: 2320-0847 p-ISSN : 2320-0936 Volume-4, Issue-3,, pp. pp-11-18.
- SLOTNICK, S. A. & S. M. J., 2005. Manufacturing lead-time rules: Customer retention versus tardiness costs.. *European journal of operational research*, pp. 163, 825-856.
- Söderlund, , M. & R. S., 2008. Revisiting the smiling service worker and customer satisfaction.. *International Journal of Service Industry Management*, pp. 19(5), 552-574. .
- Solomon, M., 1996. *Consumer Behavior*. (2nd ed.) ed. New York: : Allyn & Bacon..
- Study, S. F. S., 2018. *Machine-Learning Techniques for Customer Retention: A Comparative*, s.l.: s.n.
- SUN, Z. B. G. & M. R., 2004. Object detection using feature subset selection.. In: *Pattern Recognition*, . s.l.:s.n., pp. 37, 2165-2176.
- Svendsen, G. & P. N., 2013. The effect of brand on churn in the telecommunications sector.. *European Journal of Marketing*, pp. 47(8), 1177-1189..
- T. Vafeiadis, K. D. G. S. K. C., 2015. *A comparison of machine learning techniques for customer churn prediction Simulation Modelling Practice and Theory* , s.l.: s.n.
- TAO, Y. H. & Y. C. R., 2003. Simple database marketing tools in customer analysis and retention.. *International journal of information management*, pp. 23,291-301.
- TAO, Y. H. & Y. C. R., 2003.) Simple database marketing tools in customer analysis and retention.. *International journal of information management*, , pp. 23,291-301.
- Trubik, E. & S. M., 2000. Developing a model of customer defection in the Australian banking industry.. *Managerial Auditing Journal*, 15(5), pp. 199-208.
- Turki, A., 2010. *Customer service retention: A behavioural perspective of the UK mobile market*, , Durham heses,Durham University: s.n.
- V. Umayaparvathi, K. I., 2012. 2012 Applications of Data Mining Techniques in Telecom Churn Prediction,. *International Journal of Computer Applications* (, 0975 – 8887) Volume 42– (No.20, March 2012).
- Verbeke, W. D. K. M. D. H. J. & B. B., 2008. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *Expert Systems with Applications*. In: s.l.:s.n., pp. 218(1), 211–229.
- VERHOEF, P. C. & D. B., 2001. Predicting Customer Potential Value an Application in the Insurance Industry.. In: *Decision Support Systems*. s.l.:s.n., pp. 32, 189-199. 115.

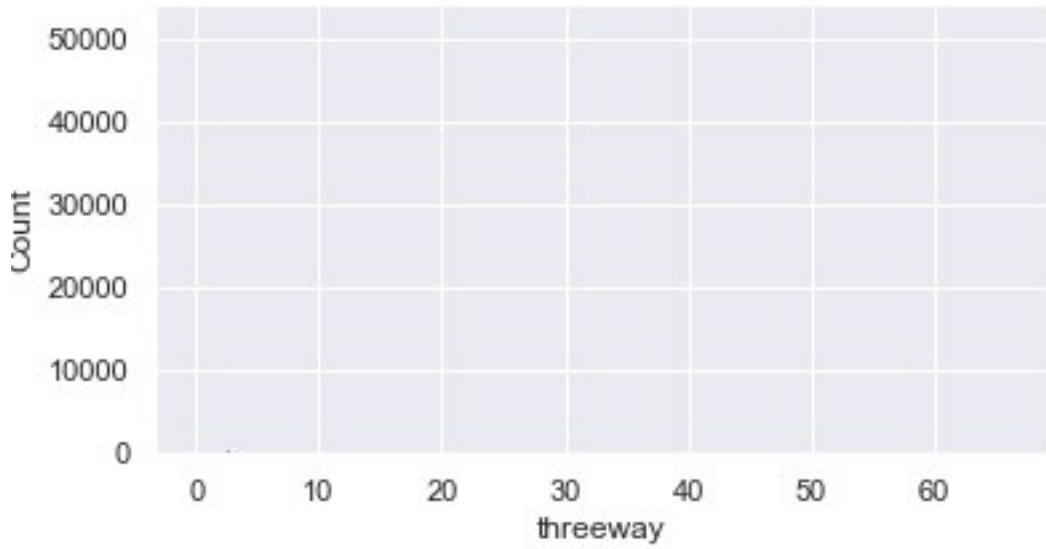
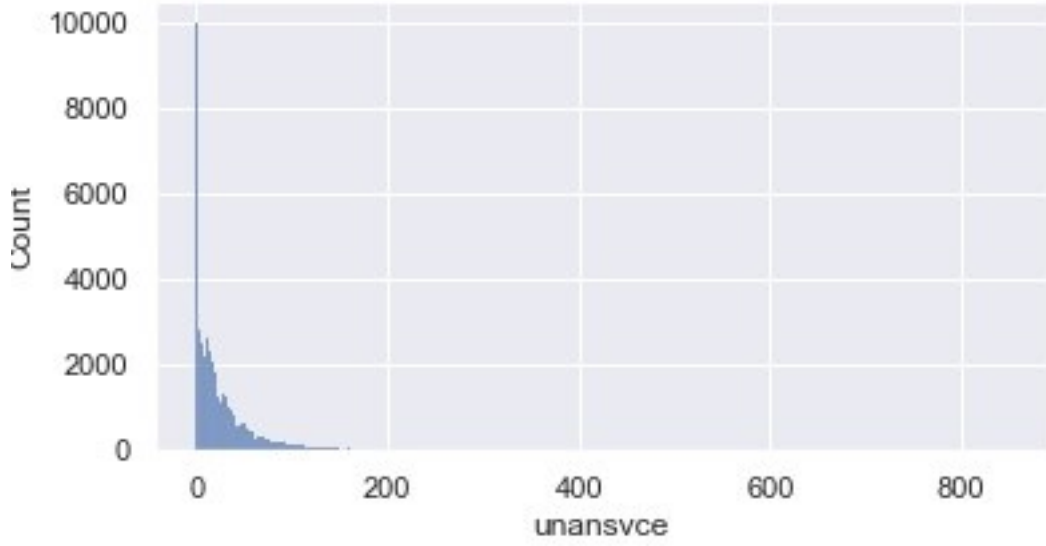
- Weerahandi, S. & M. S., 1995. Using survey data to predict adoption and switching for services.. *Journal of Marketing Research*, , pp. 32(1), 85-96.
- Wei, C. & C. I., 2002. Turning telecommunications to churn prediction: a data mining approach. *Expert Systems with Applications*. In: s.l.:s.n., pp. 23, 103-112.
- Wells, W. & P. D., 1996. *Consumer Behavior*. Hoboken: s.l.:John Wiley and Sons..
- Wen-yeh, H. H. S. & A. J., 2004. Effect of brand name on consumers risk perceptions in online shopping.. *Journal of consumer behavior*, , pp. 4(1), 40-50.
- Westbrook, R. A. N. J. W. & T. J. R., 1978. Satisfaction/dissatisfaction in the purchase decision process,. *Journal of Marketing*, 42(October), 54-60.
- Yang, L. S. & C. C., 2006. *Knowledge discovery on customer churn prediction*. *Proceedings of the 10th WSEAS International Conference on Applied Mathematics*. , Dallas, Texas, USA., s.n.
- YAN, L. W. R. & D. R., 2004. *Predicting customer behaviour in telecommunications*., s.l.: IEE Intelligent Systems, 19, 50-58.
- Zeithaml, V., 2000. Service quality, profitability and the economic worth of customers: What we know and what we need to learn. *Journal of the Academy of Marketing Science*, pp. 28(1), 67-85..
- Zorn, S. J. W. & B. S., 2010. Attitudinal perspectives for predicting churn.. *Journal of Research in Interactive Marketing*, , pp. 4(2), 157-169.

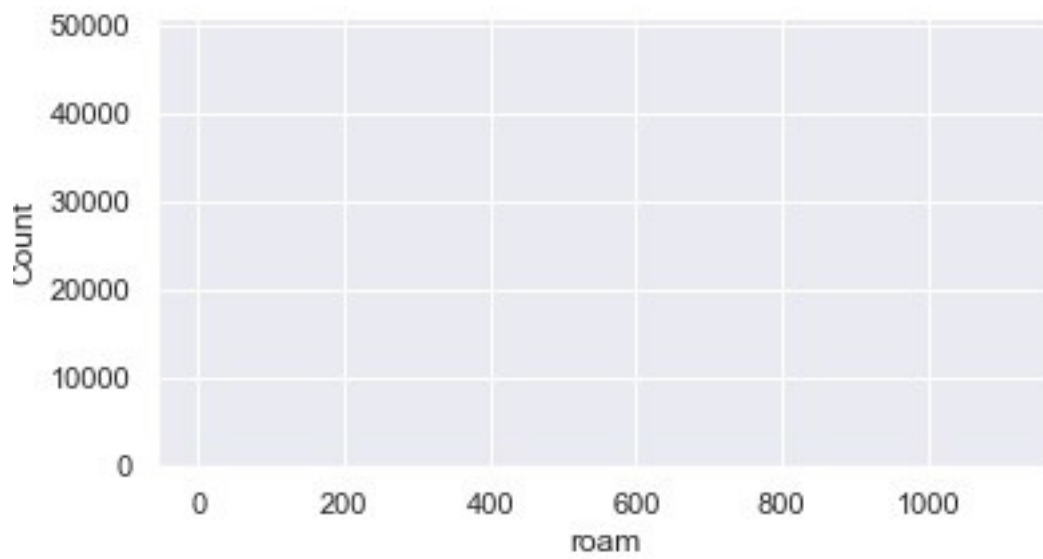
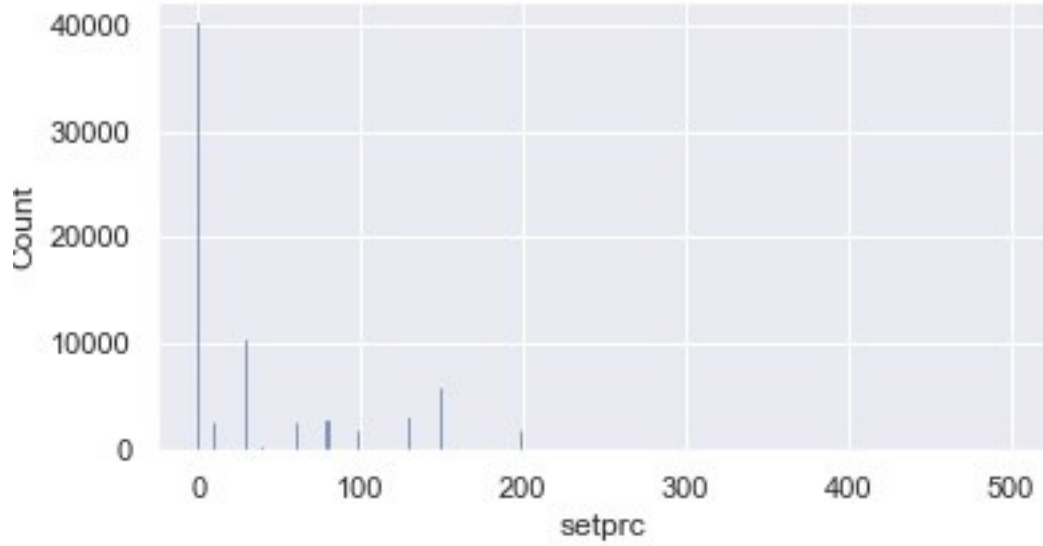
Appendix 01

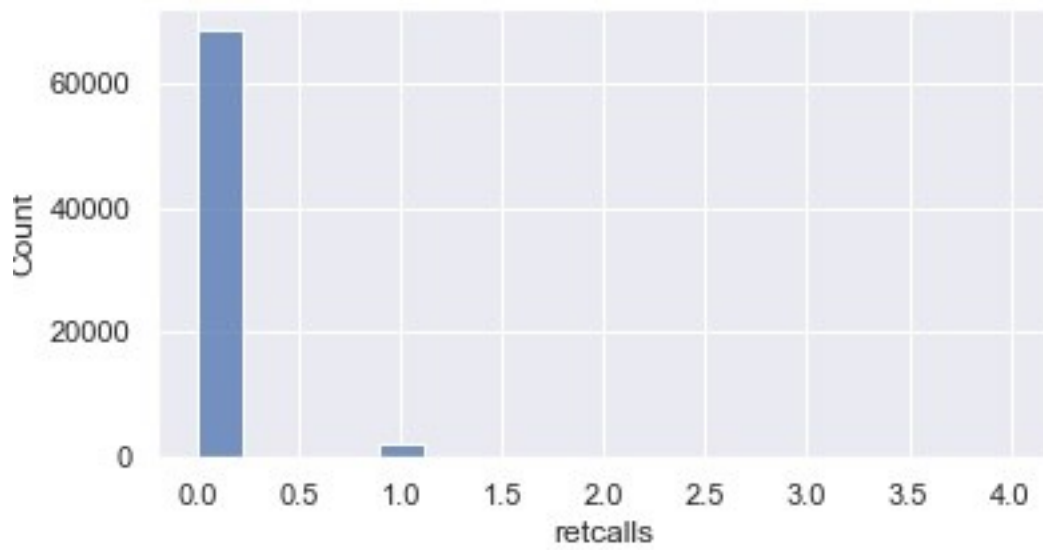
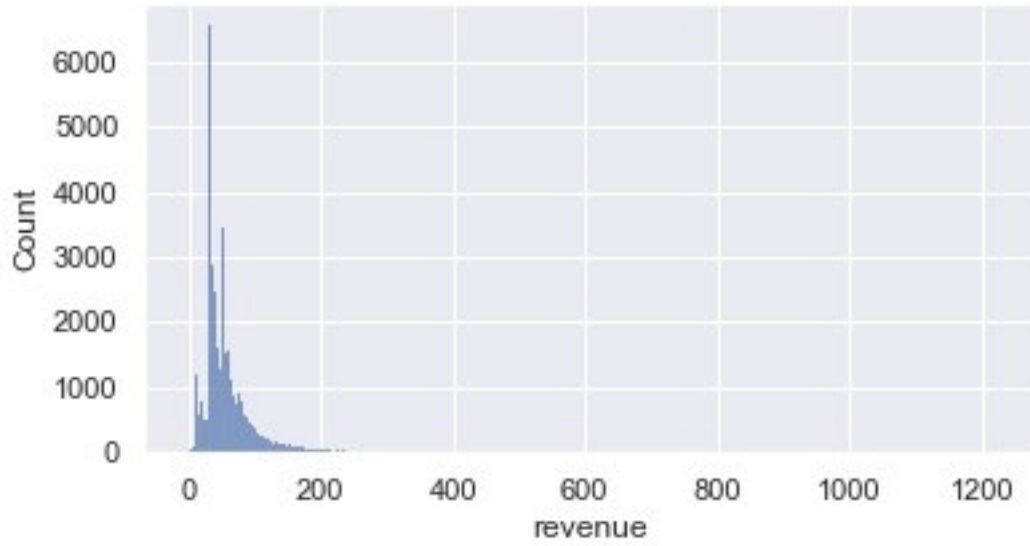
Numerical Variable Visualization

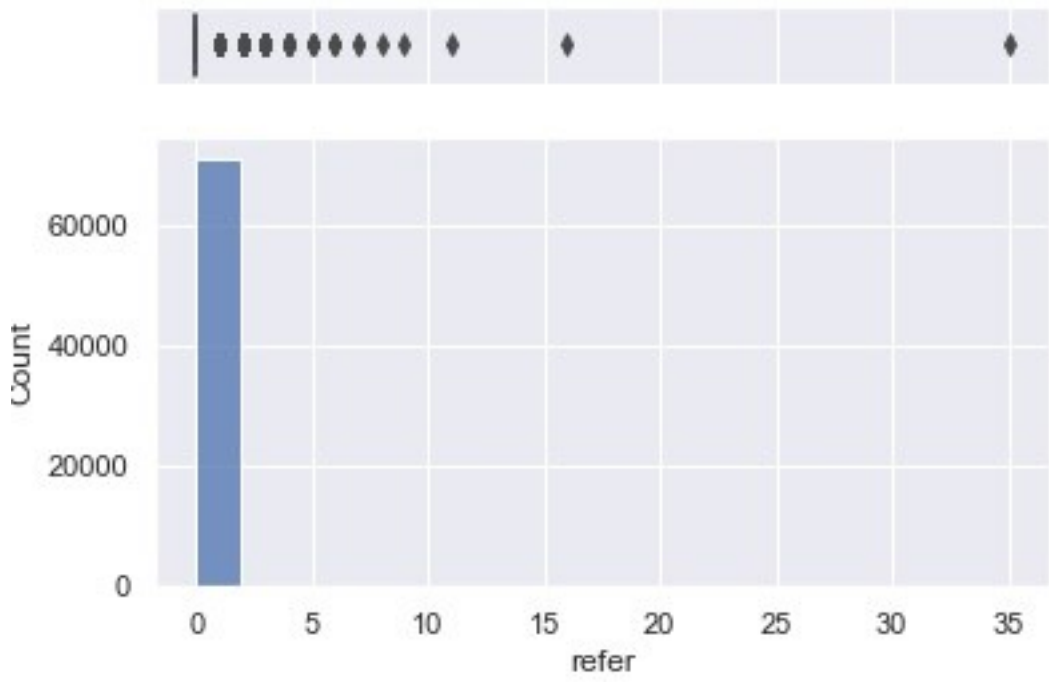
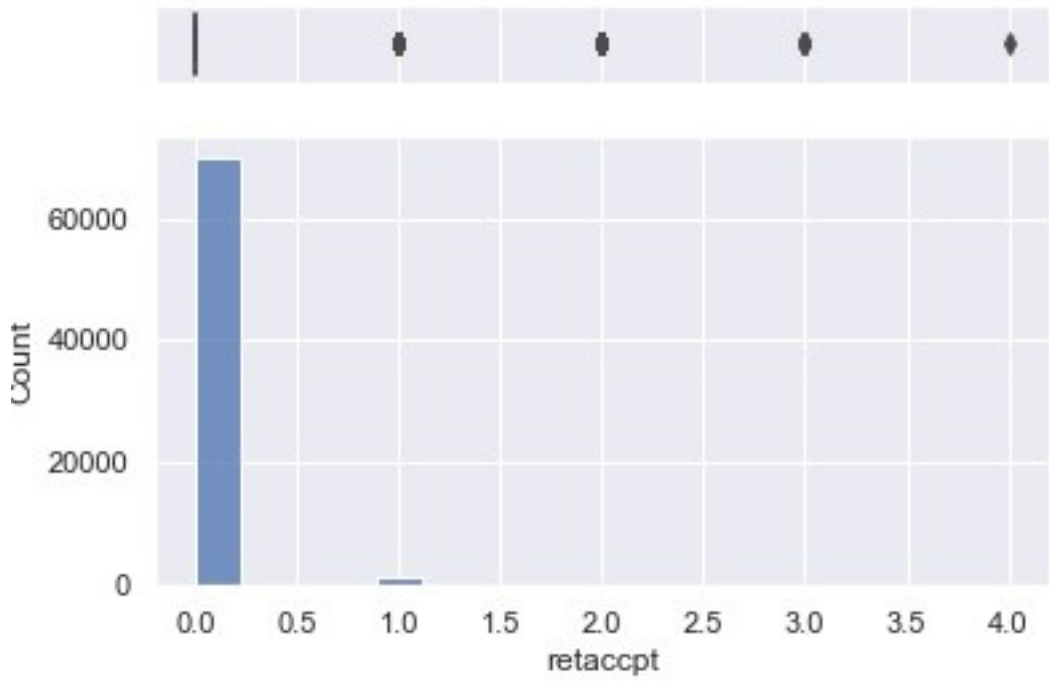


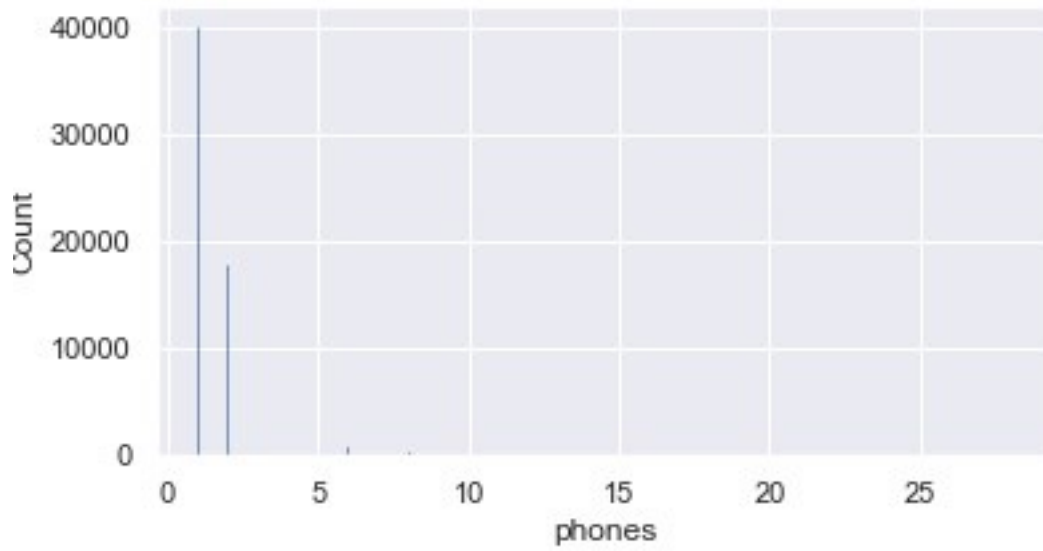
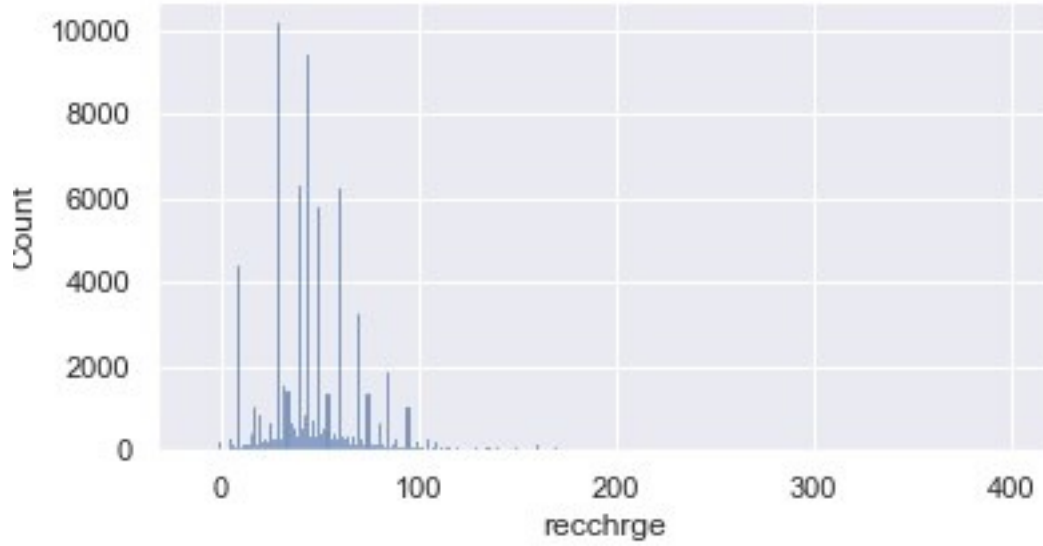


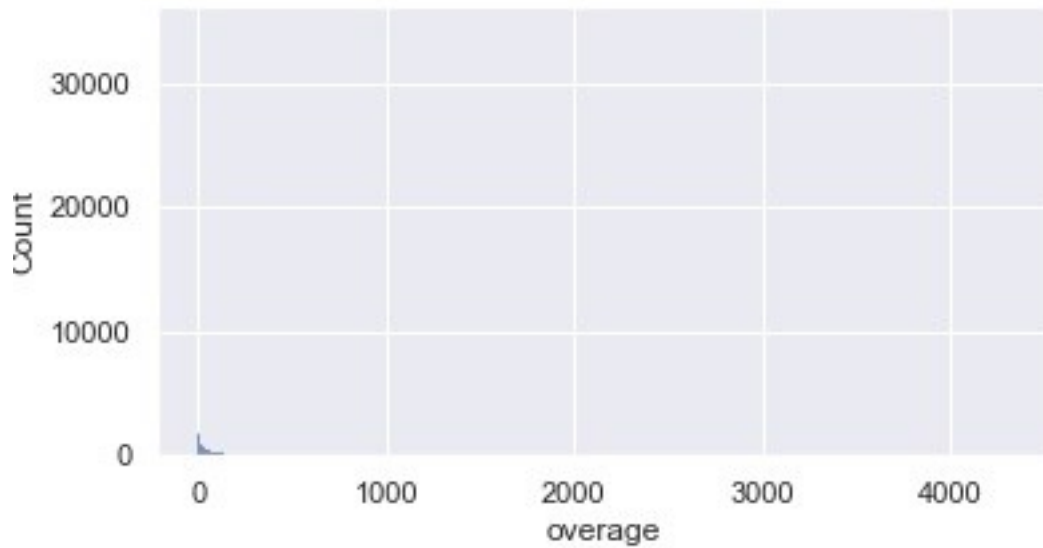
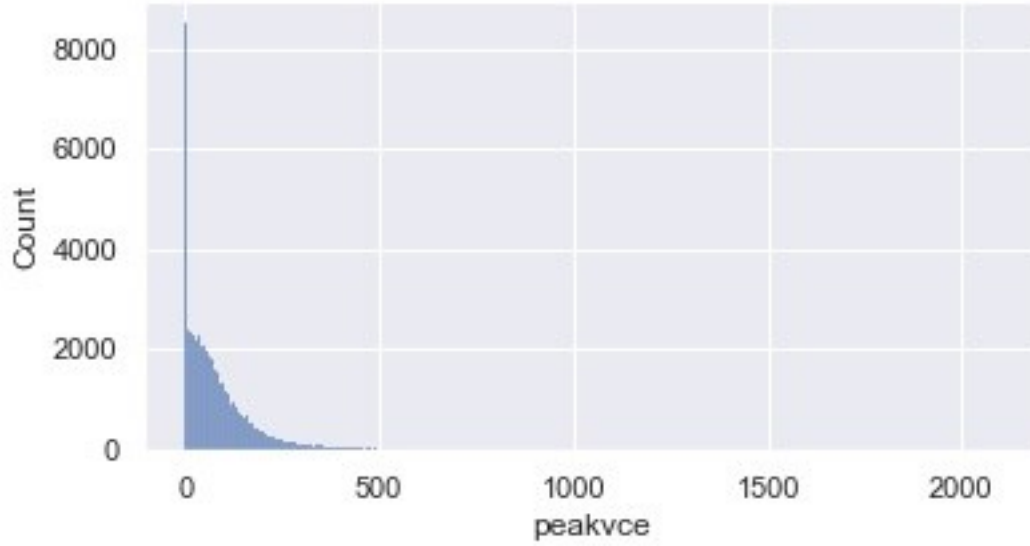


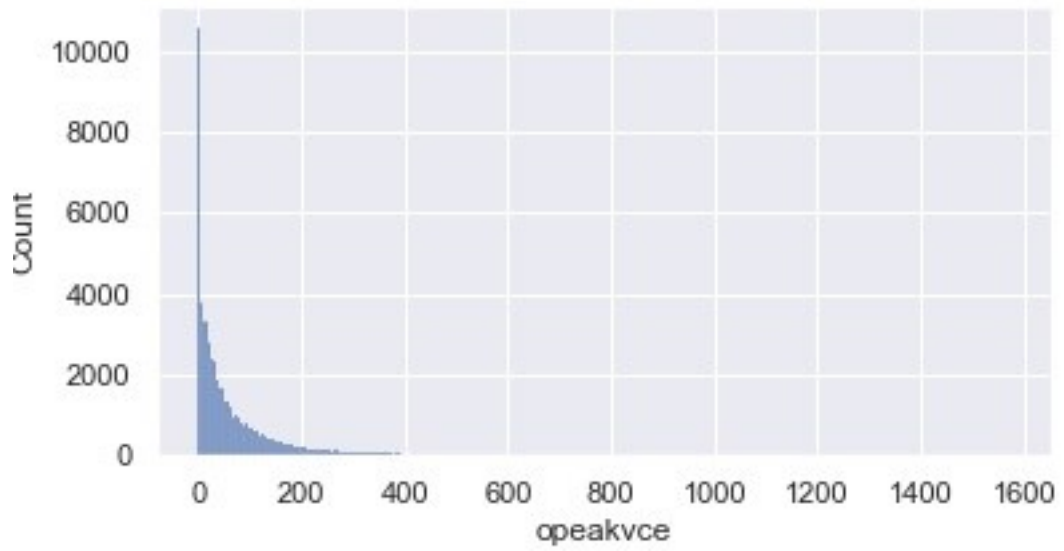
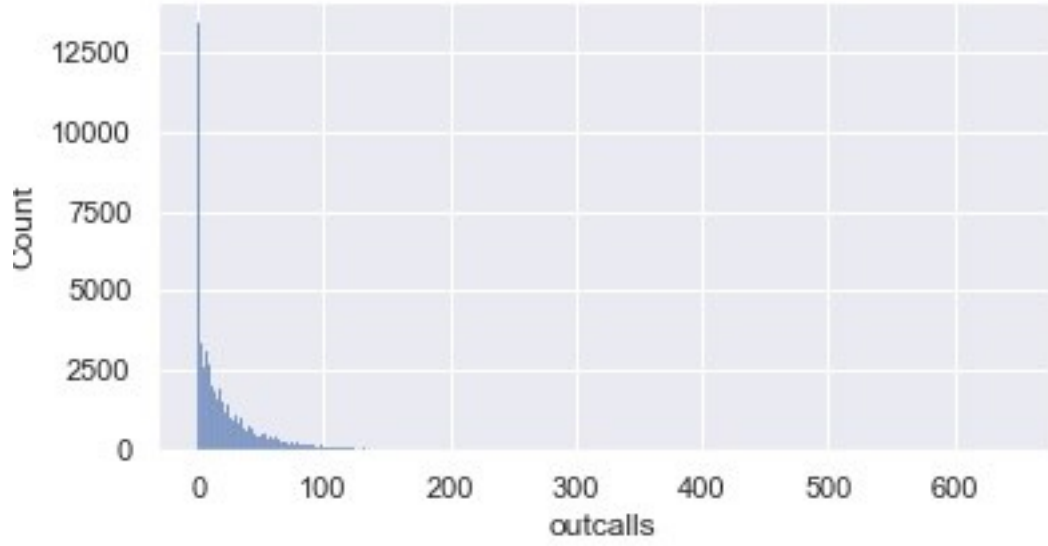


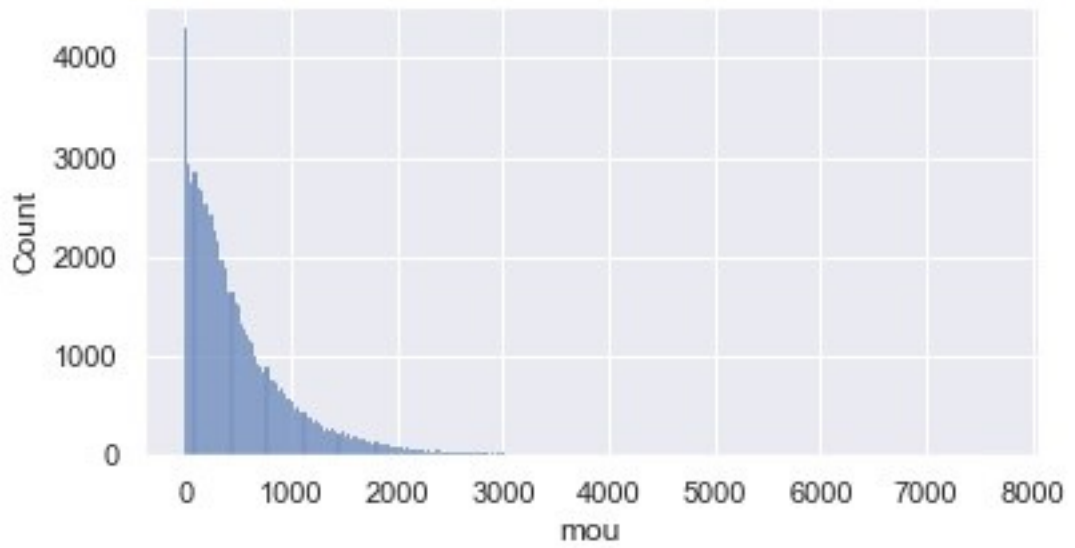
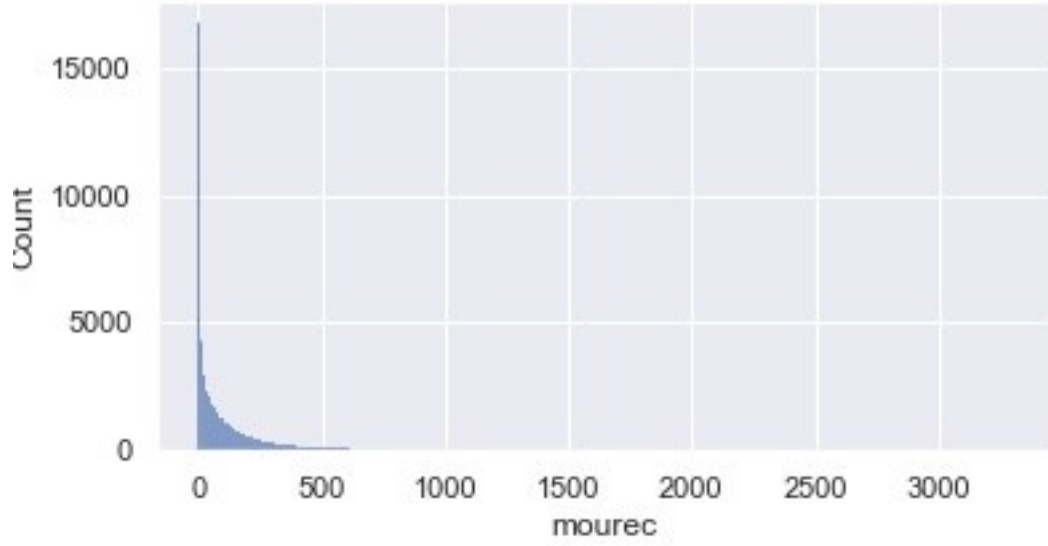


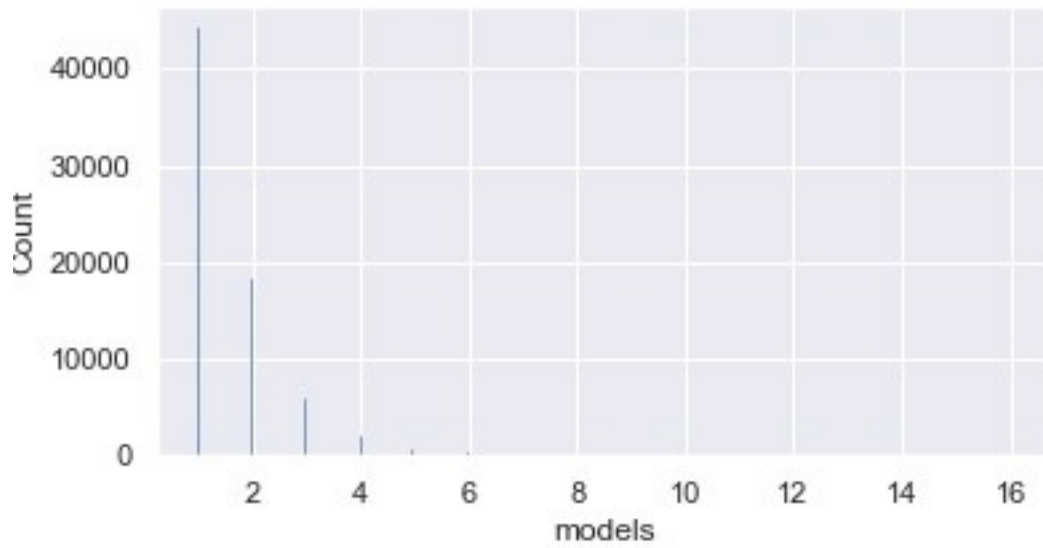
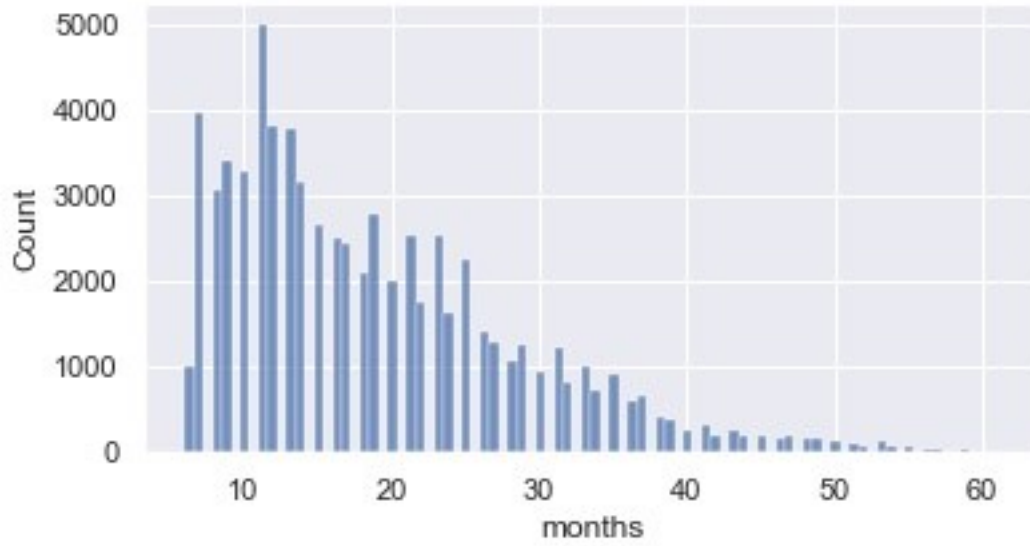


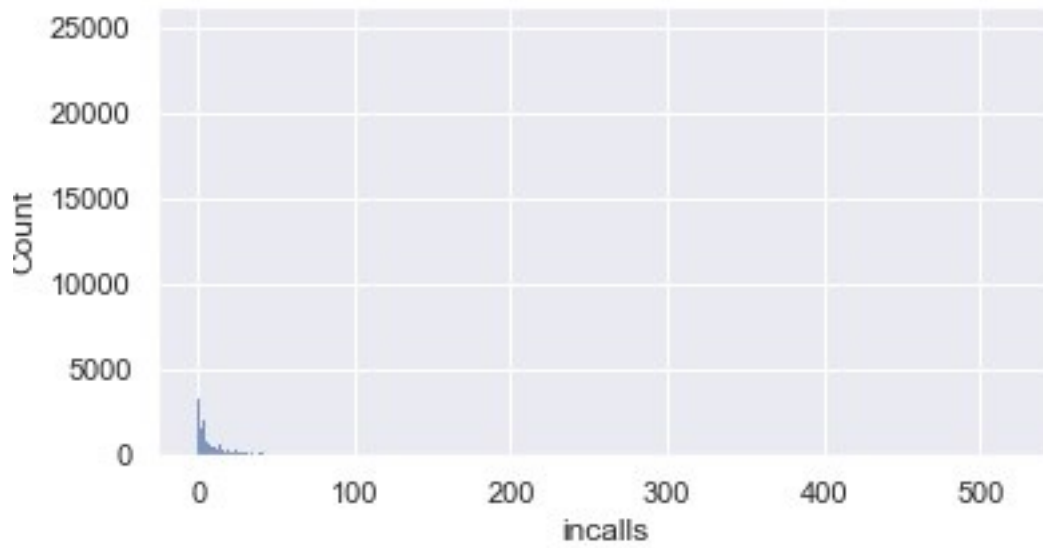
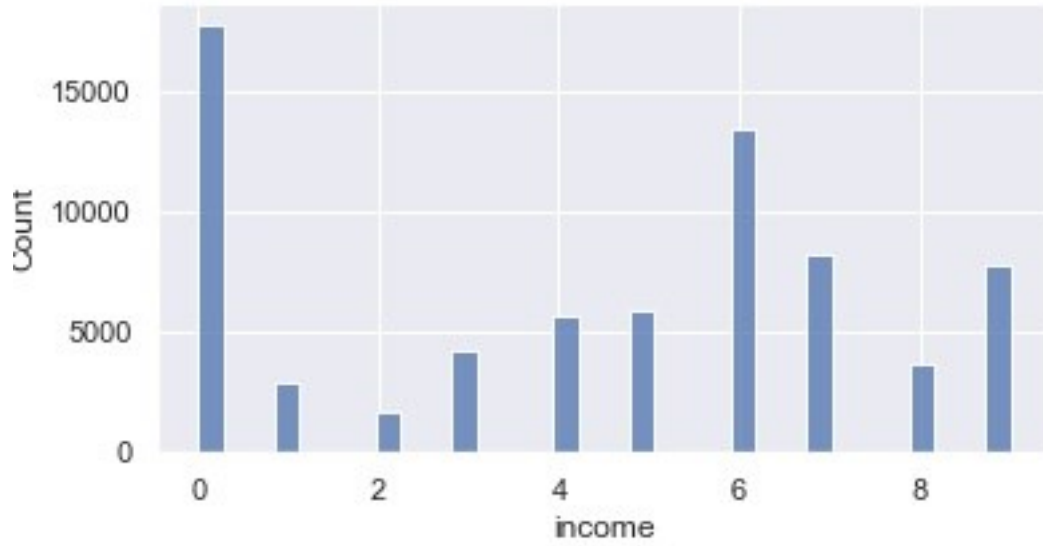


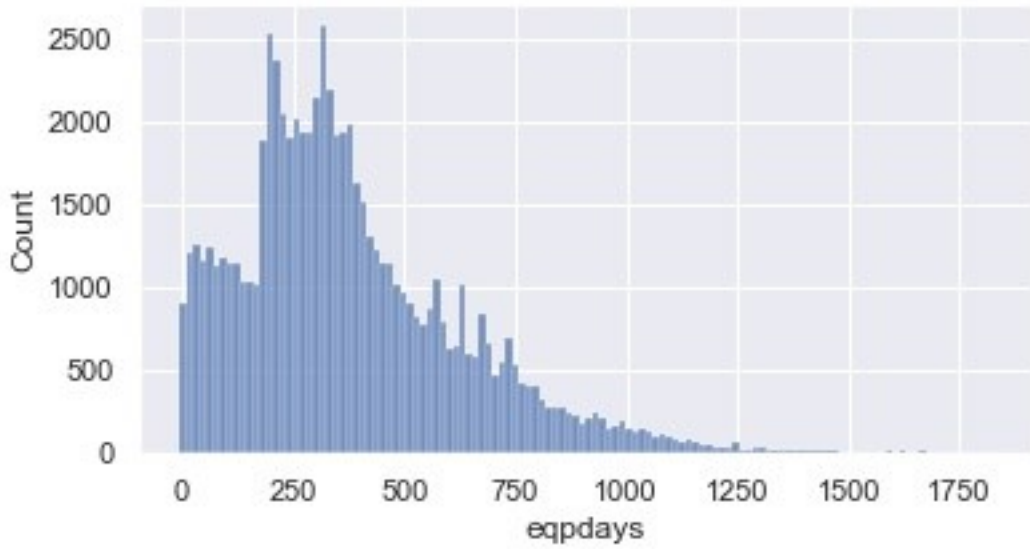
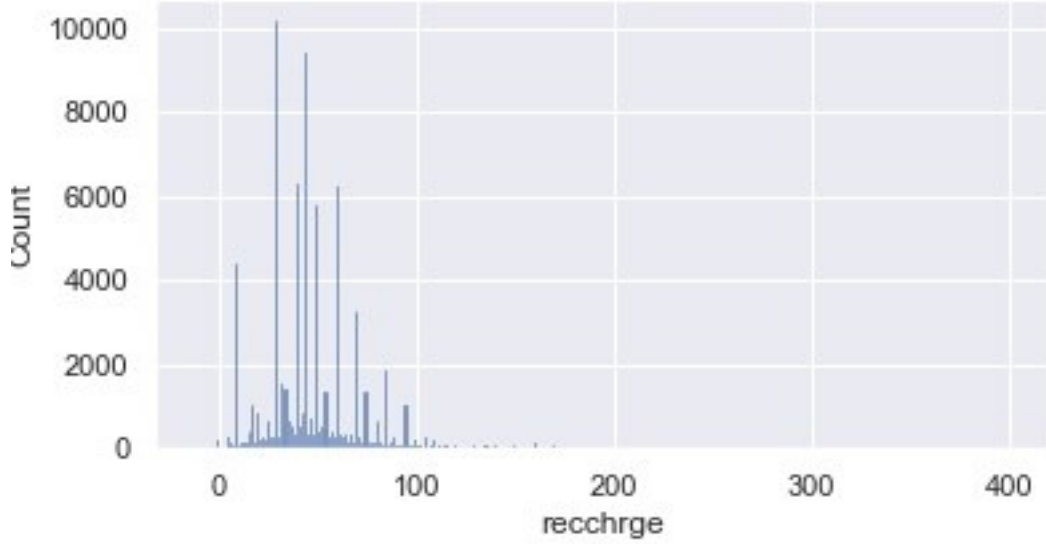


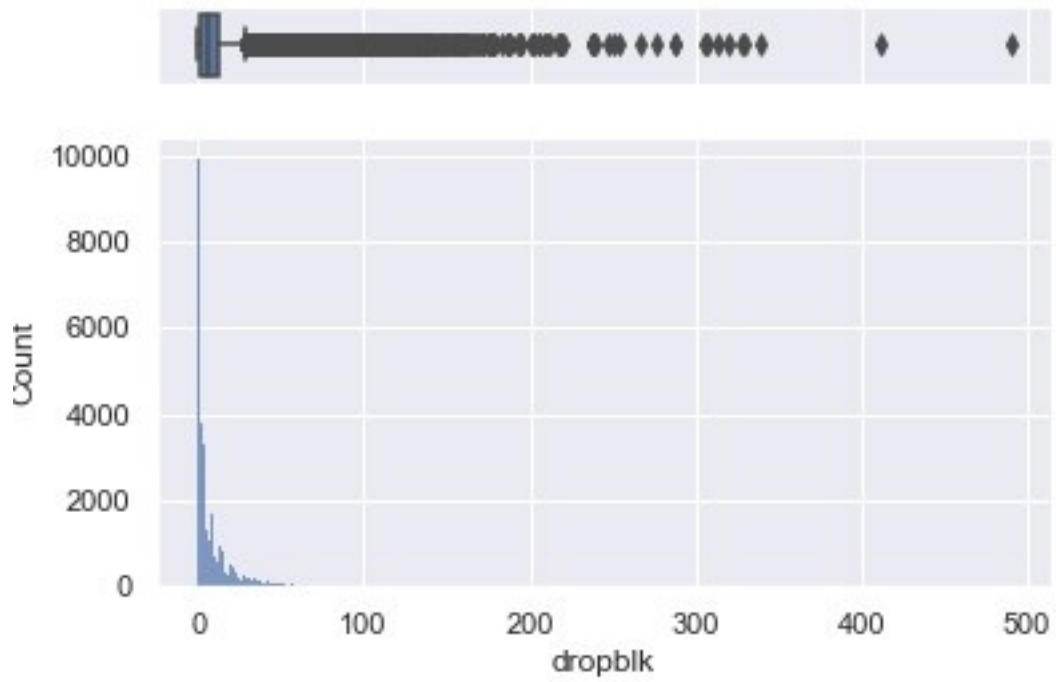
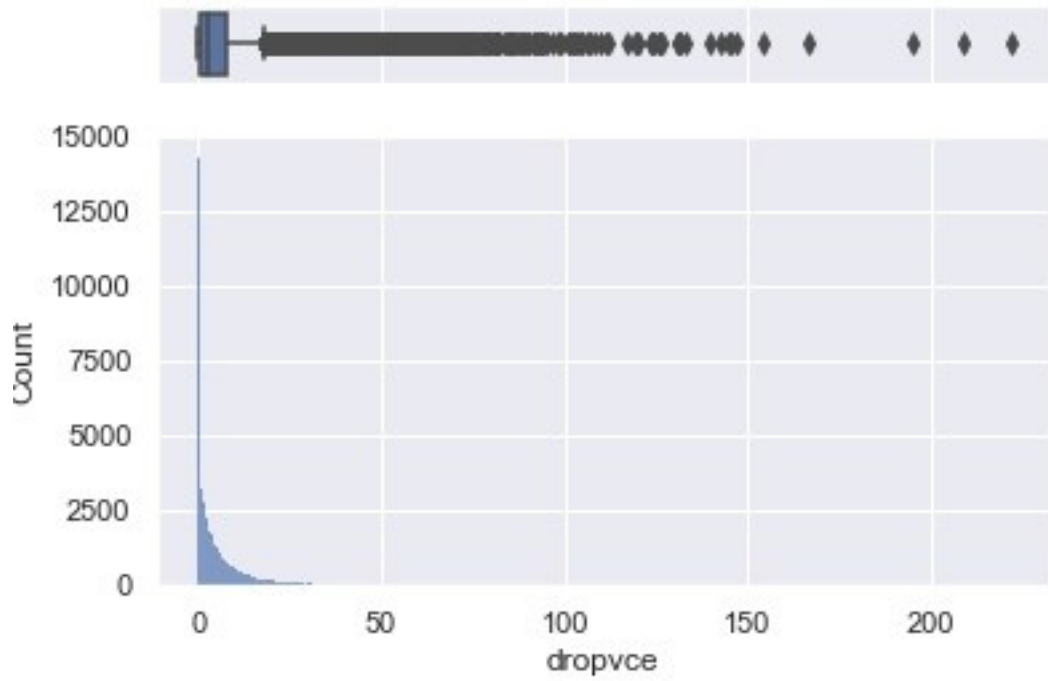


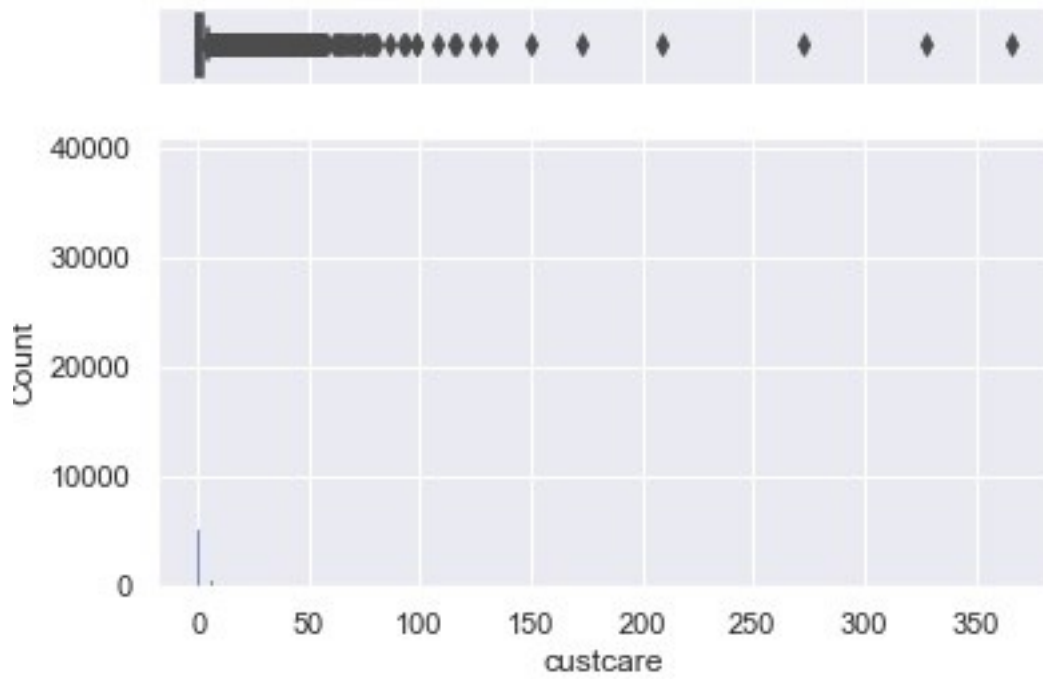
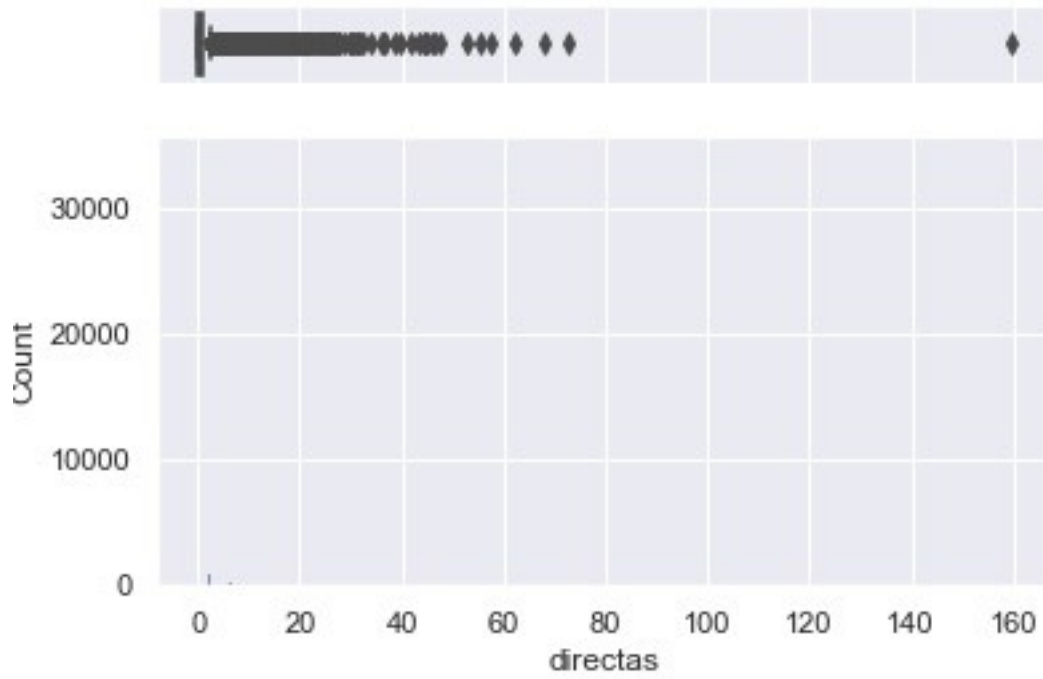


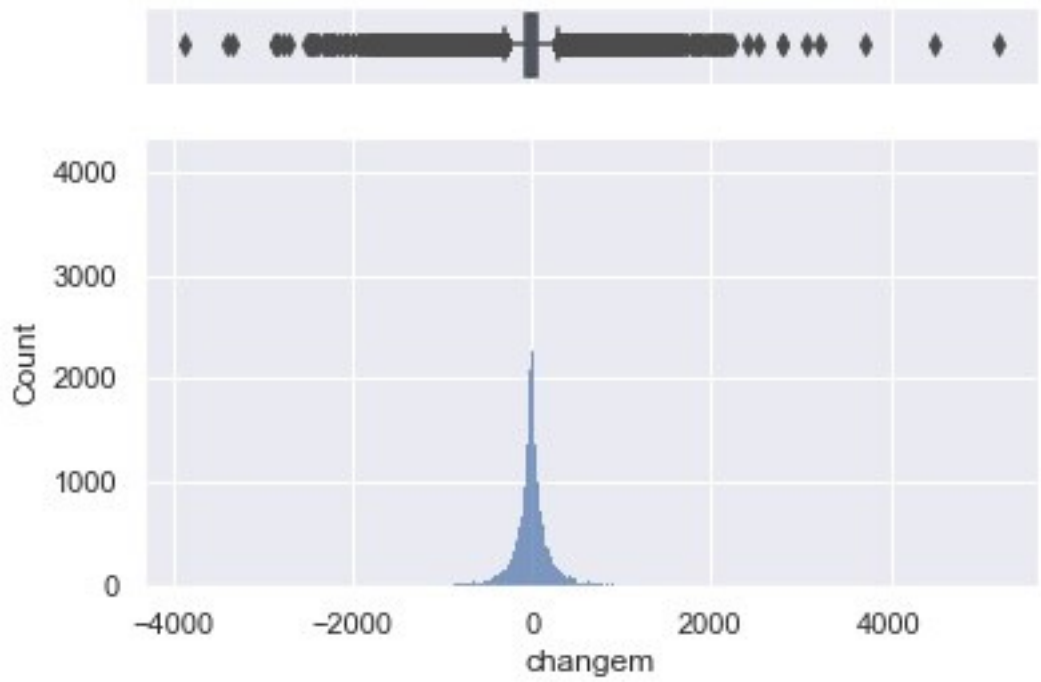
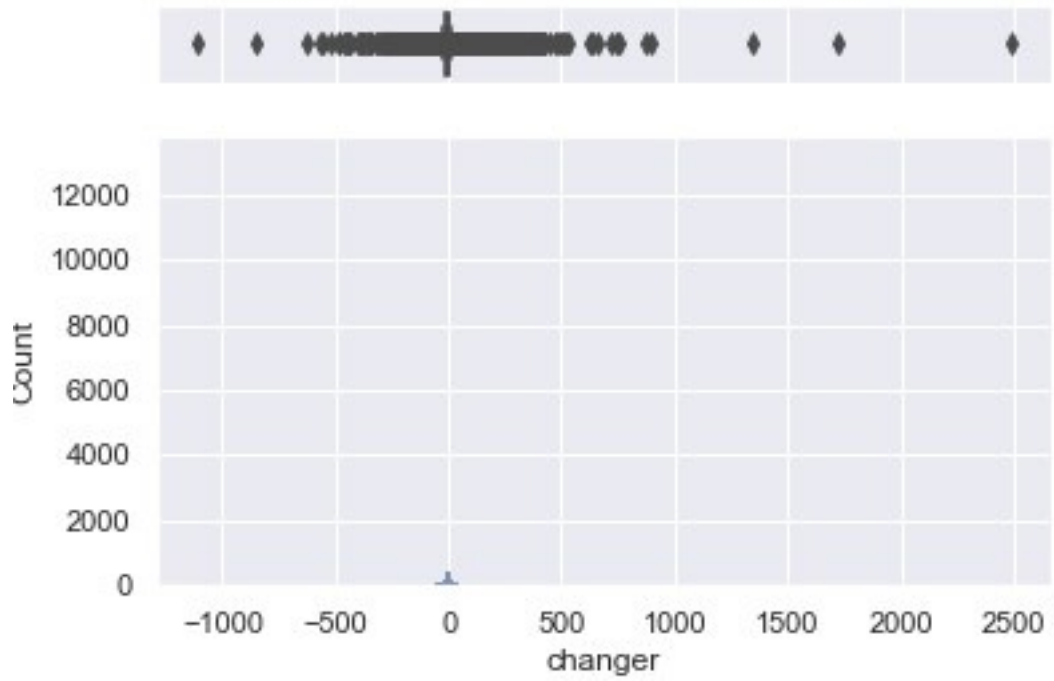


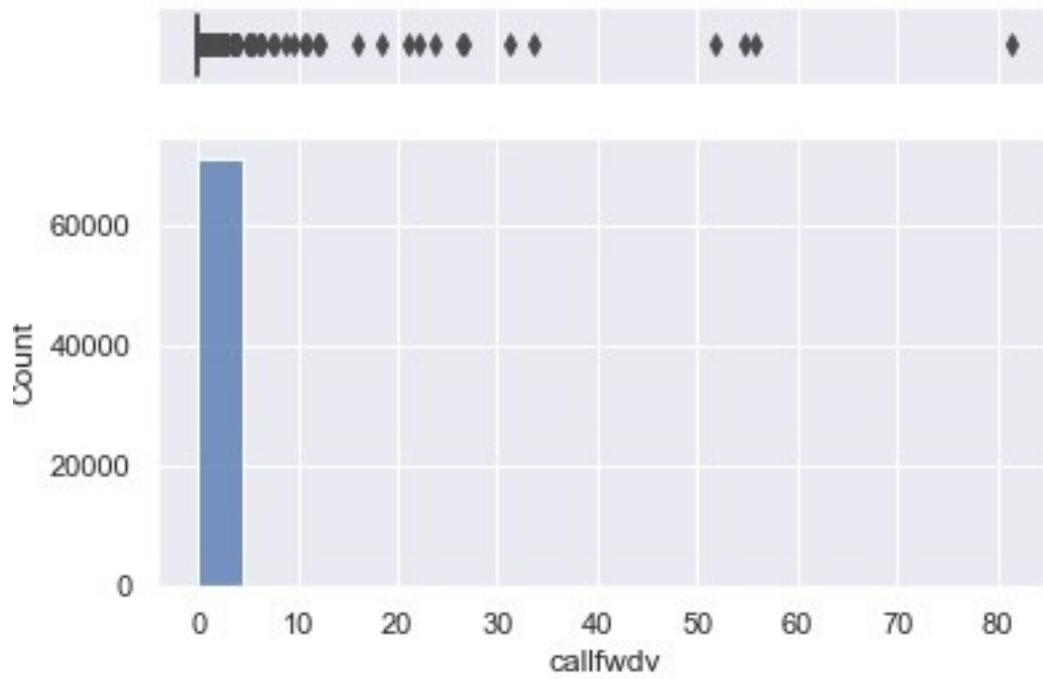
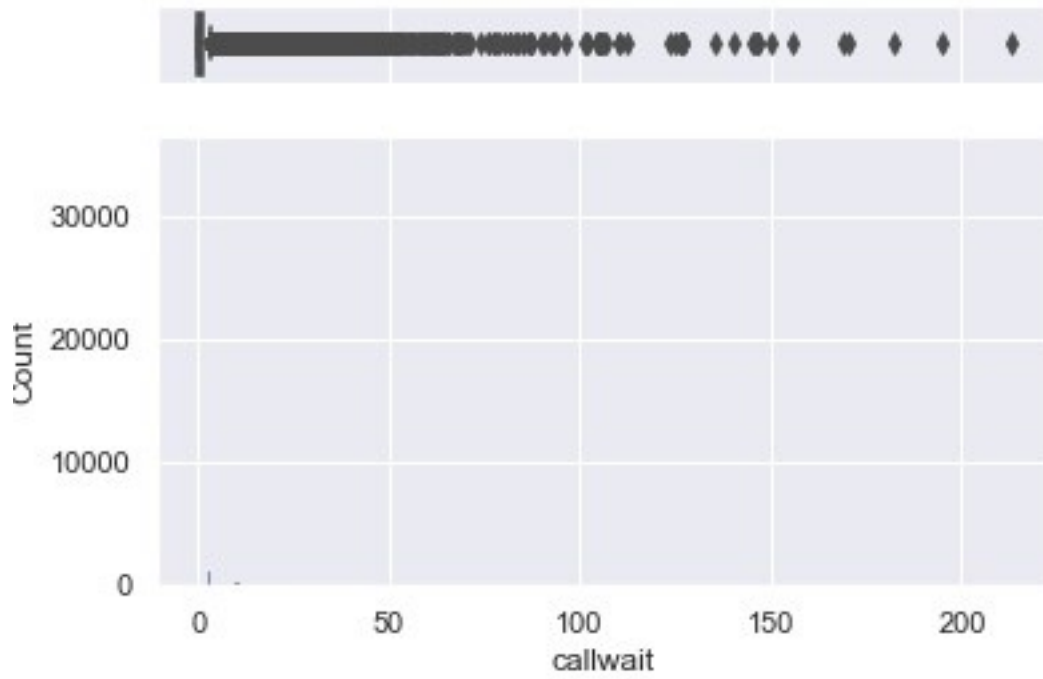


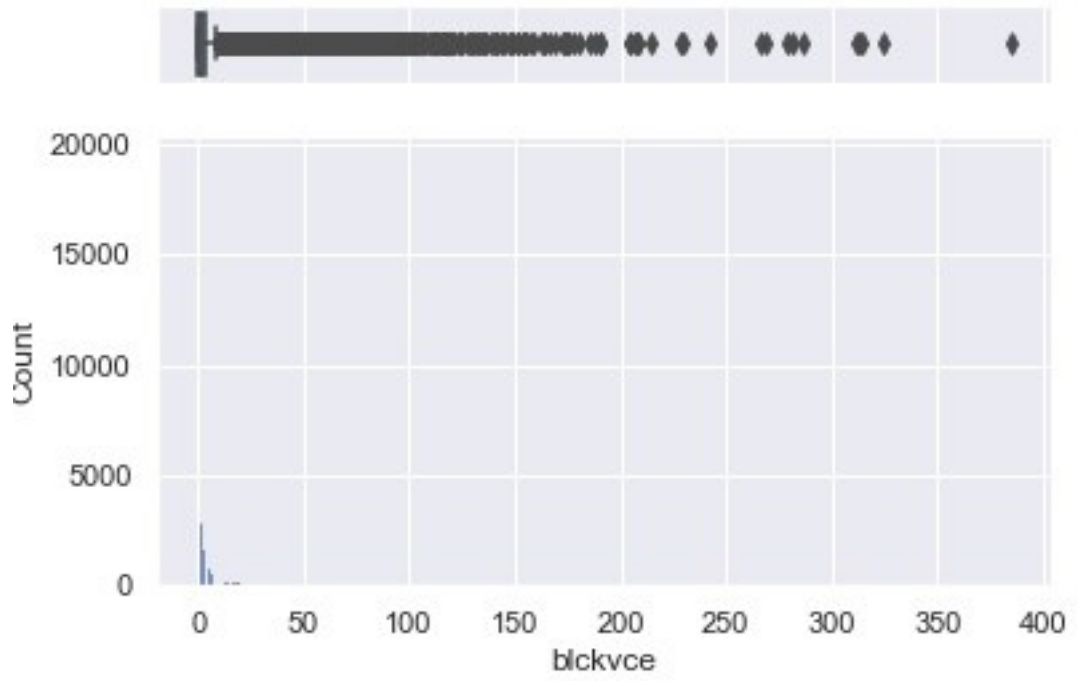




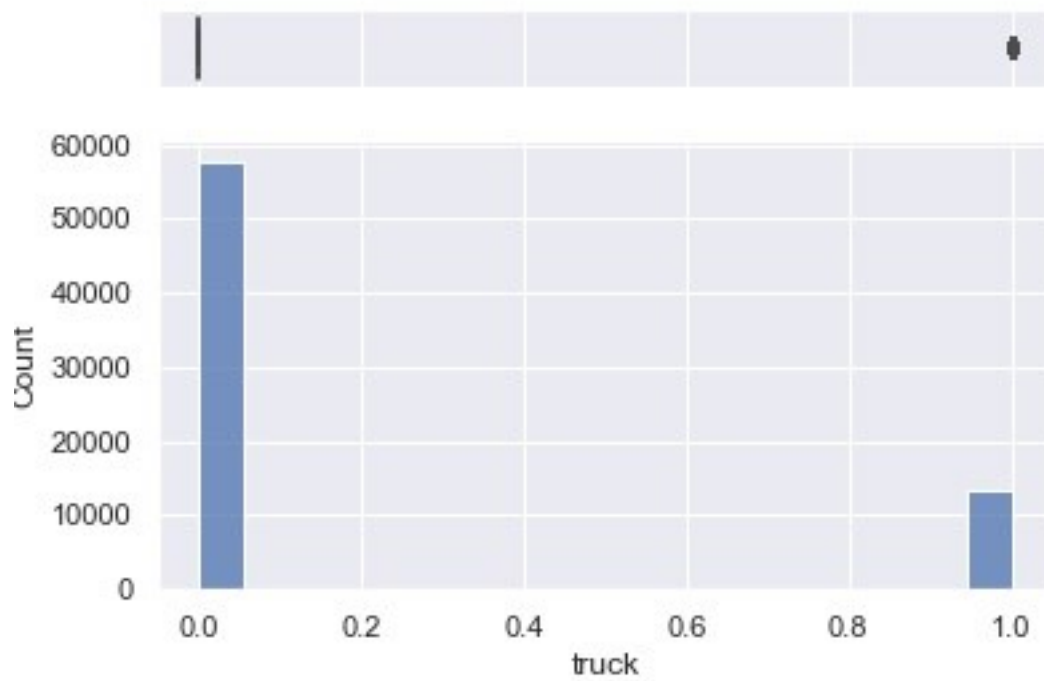
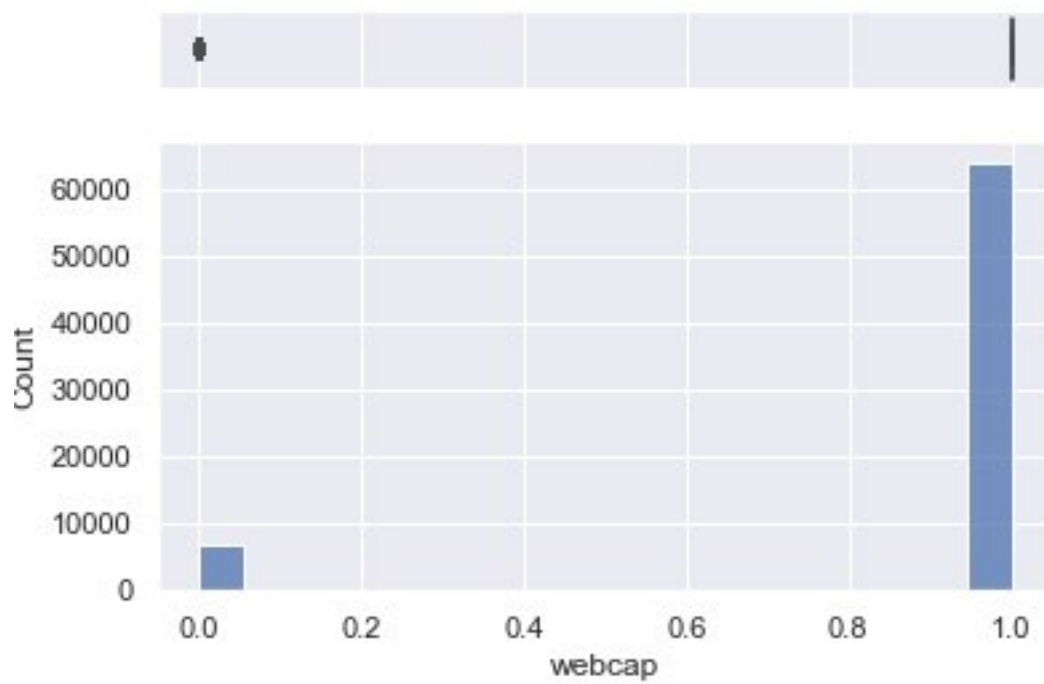


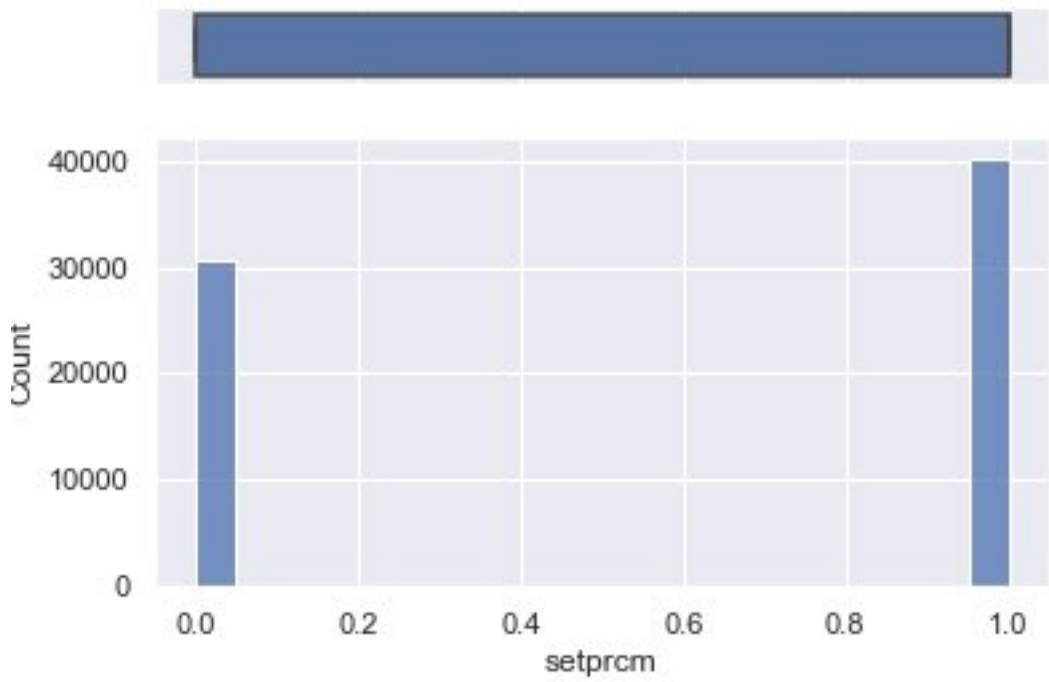
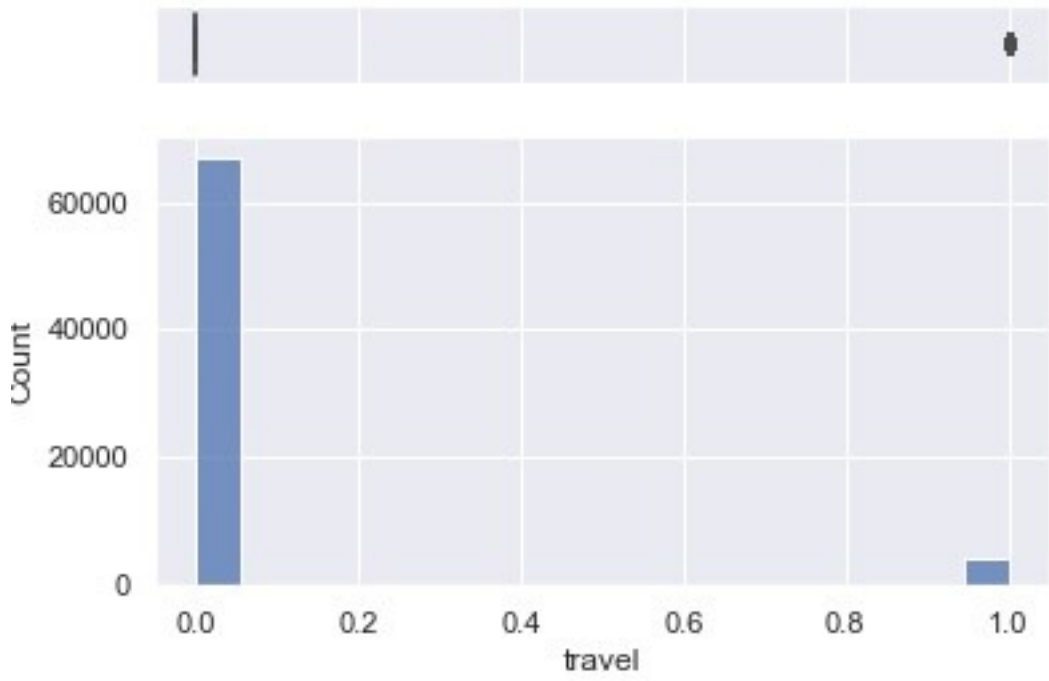


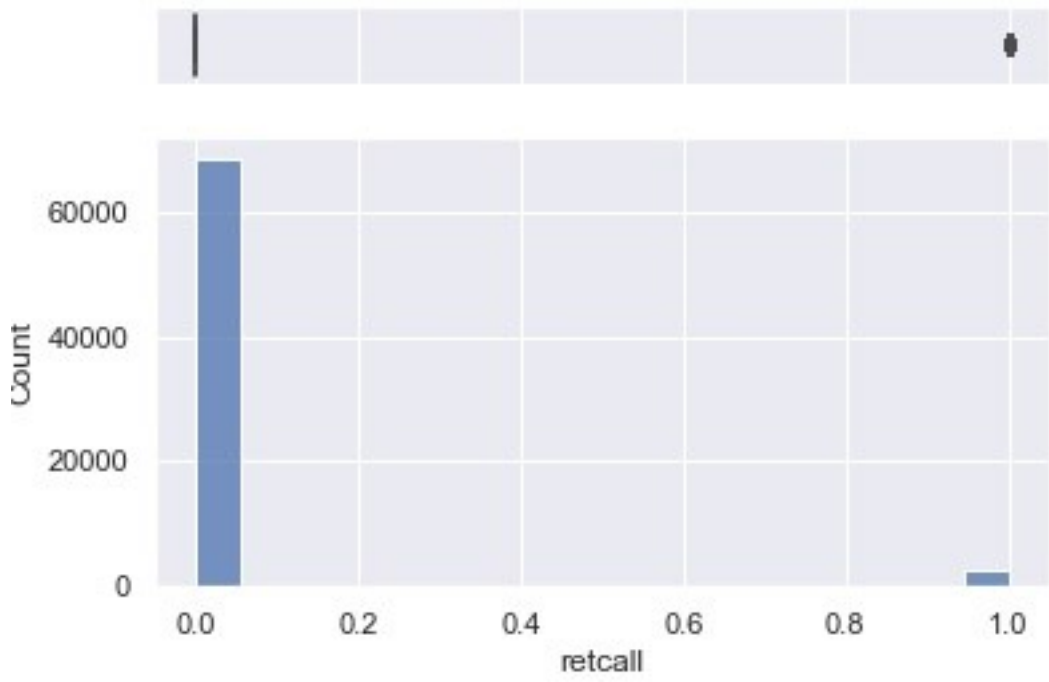
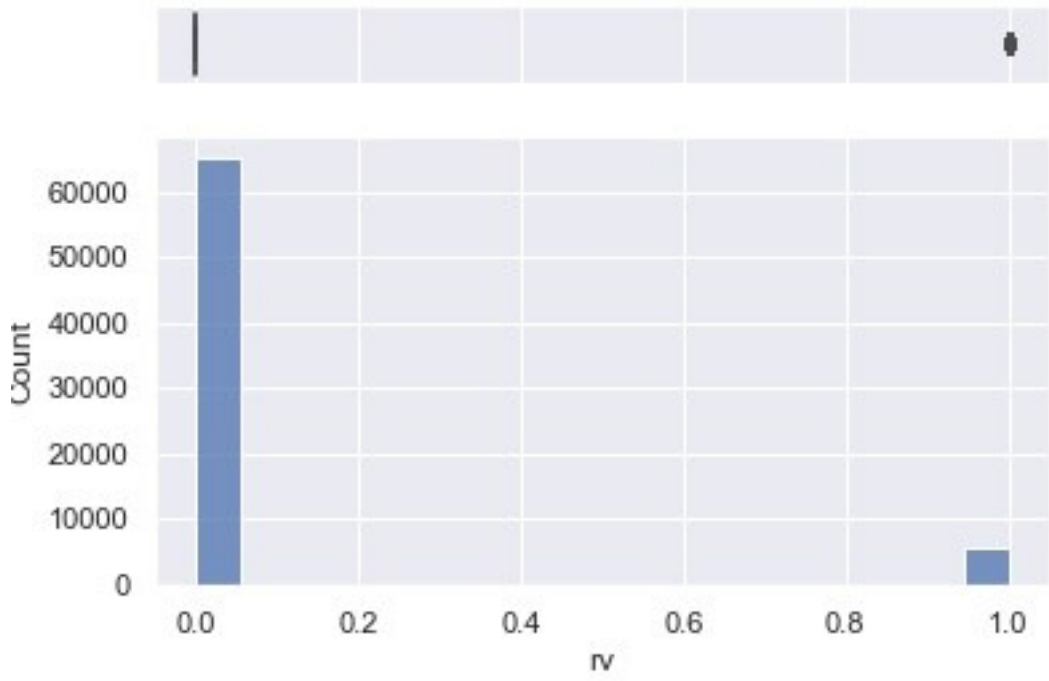


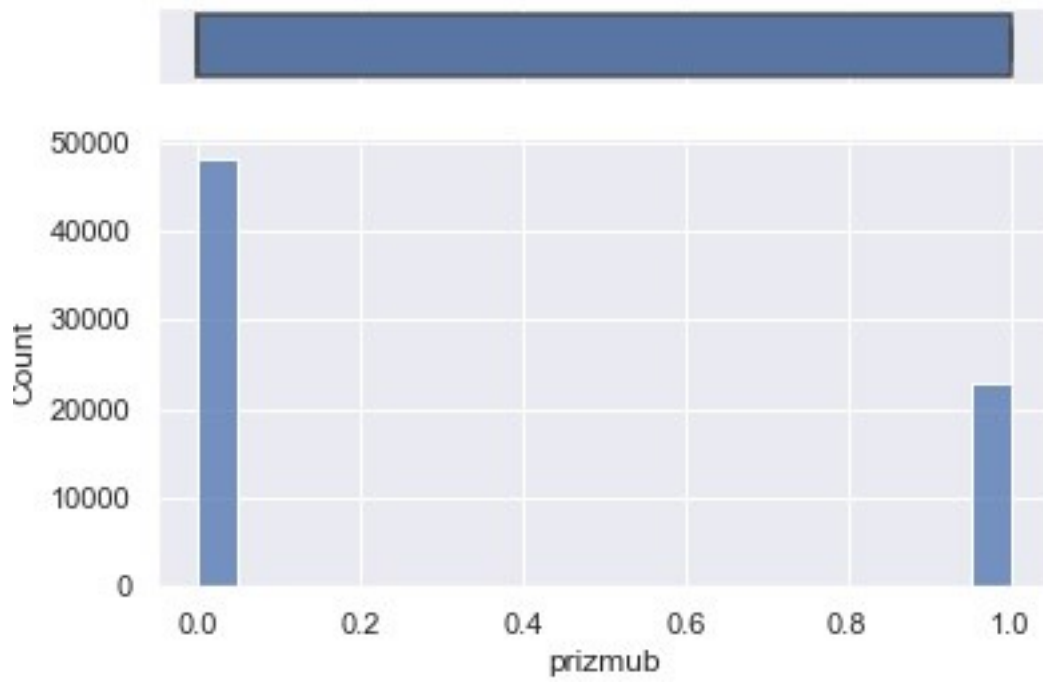
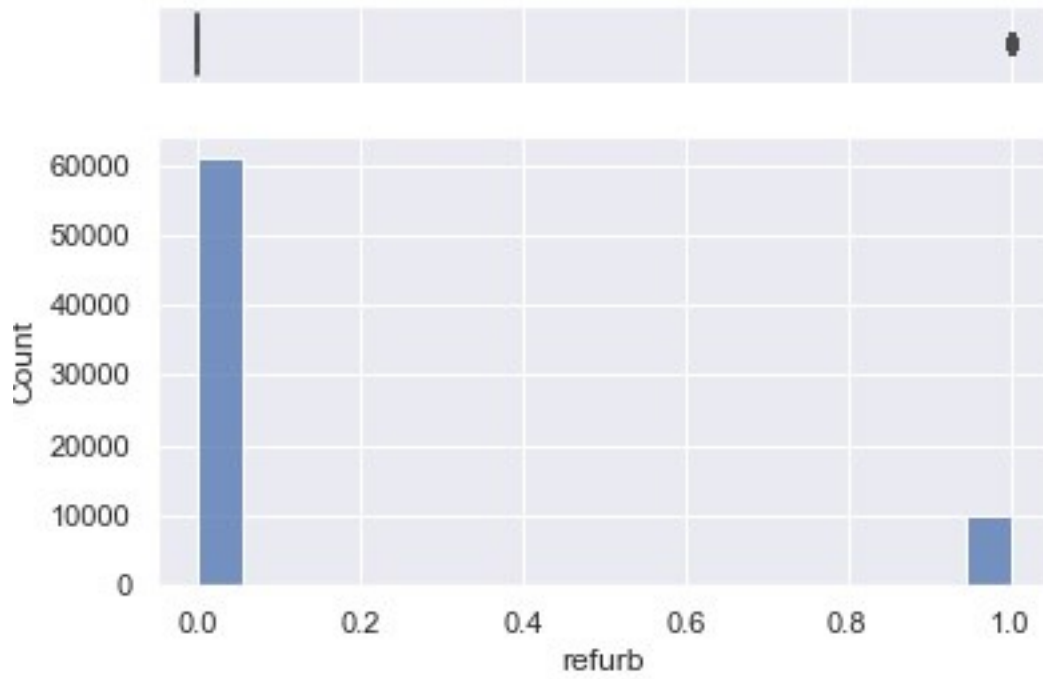


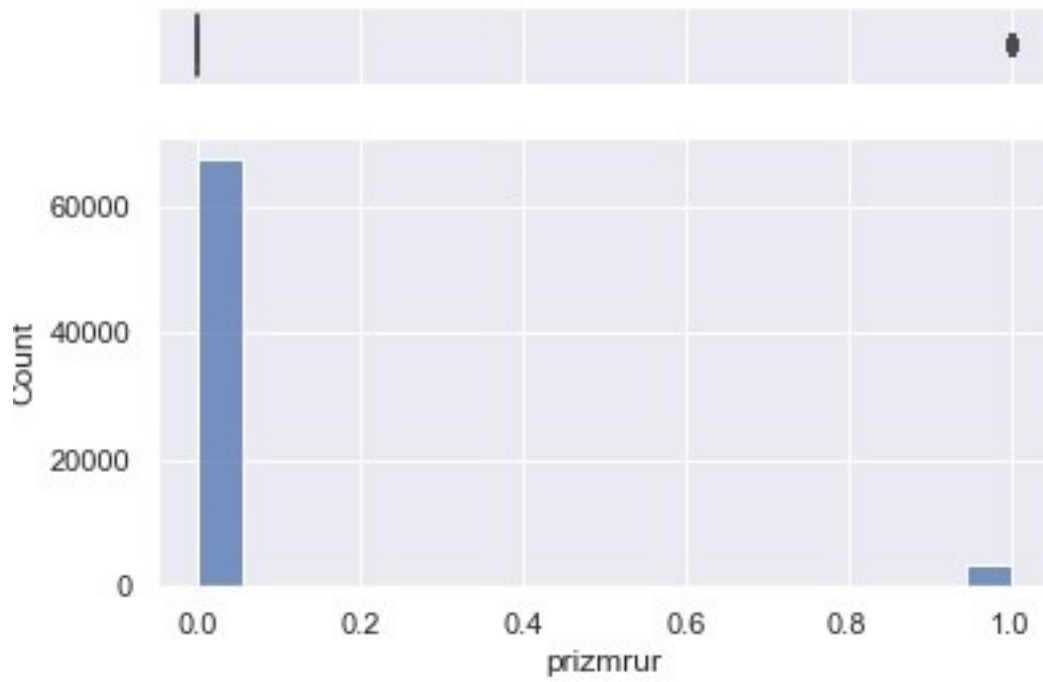
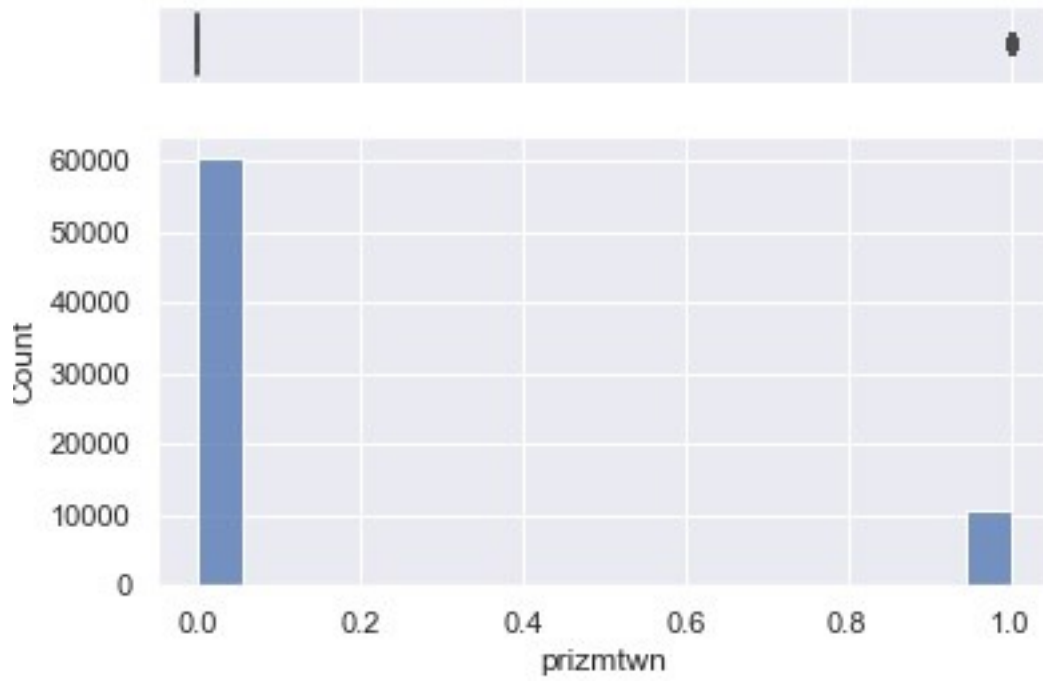
Categorical Variables visualization

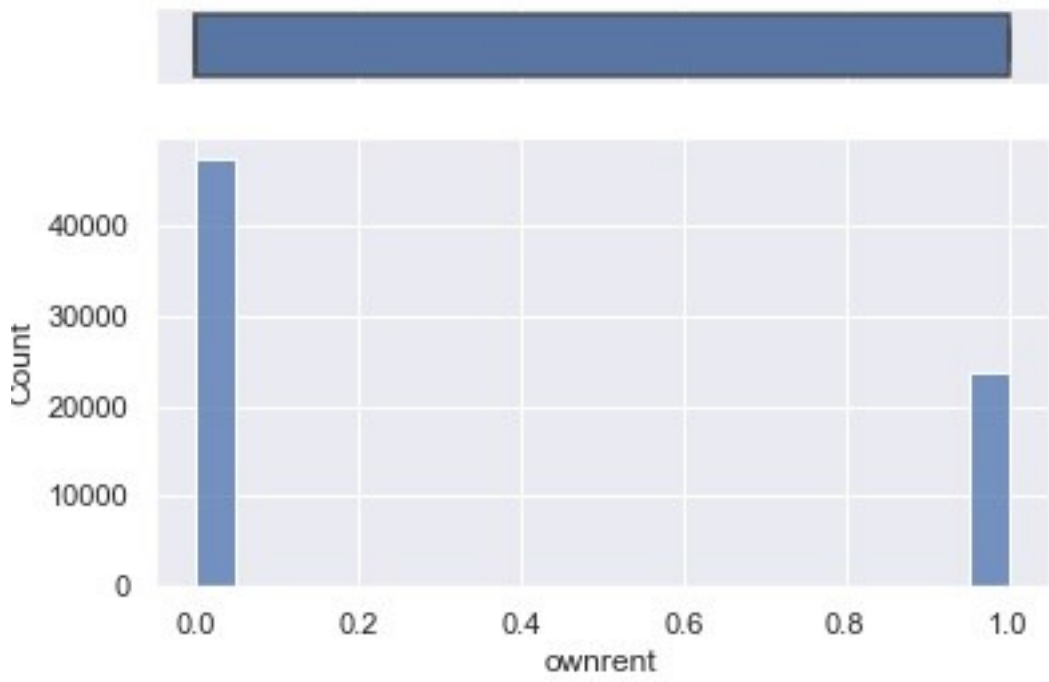
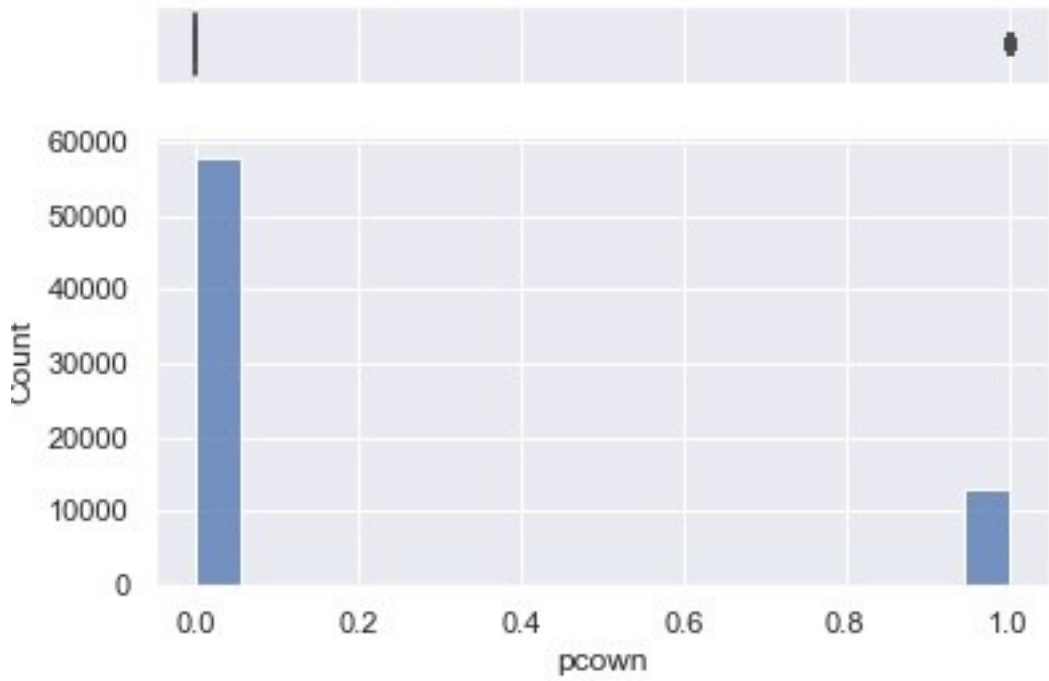


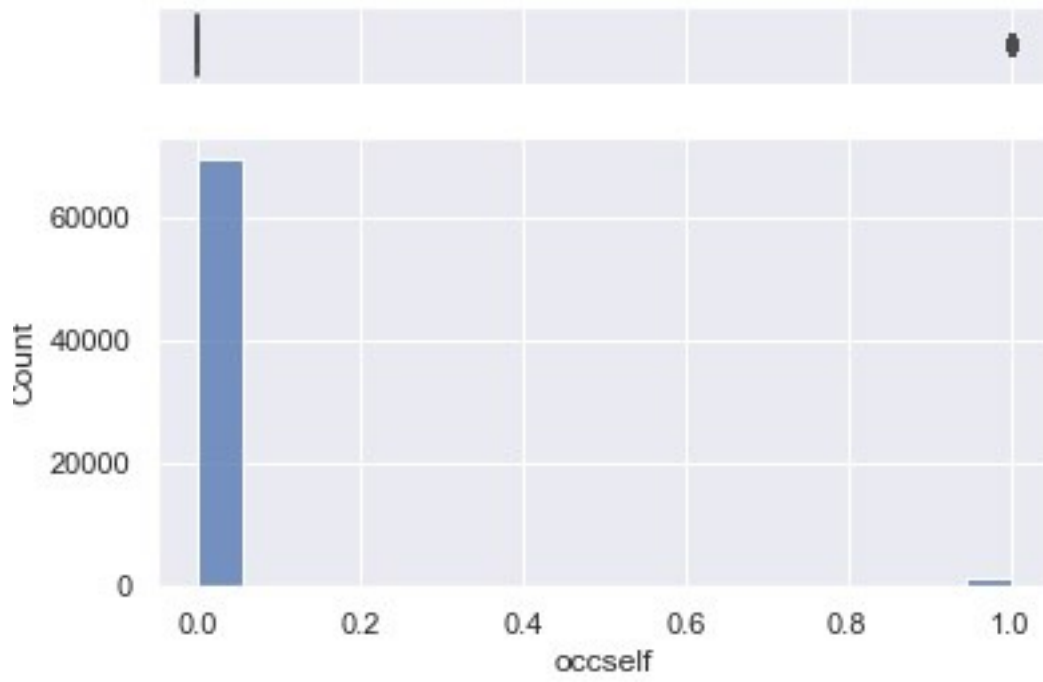
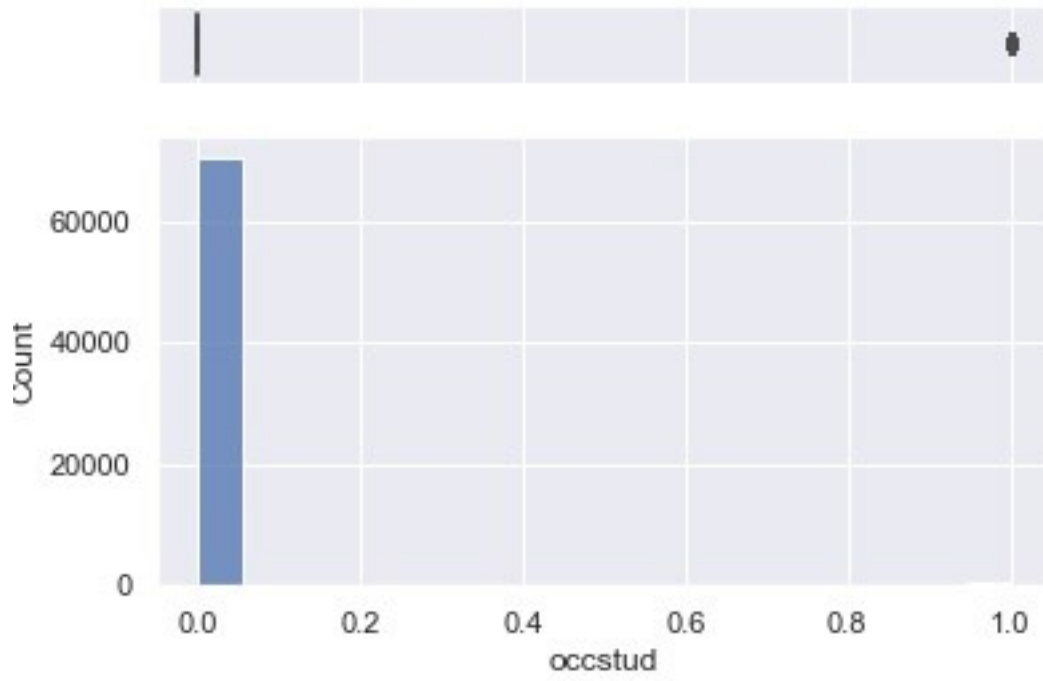


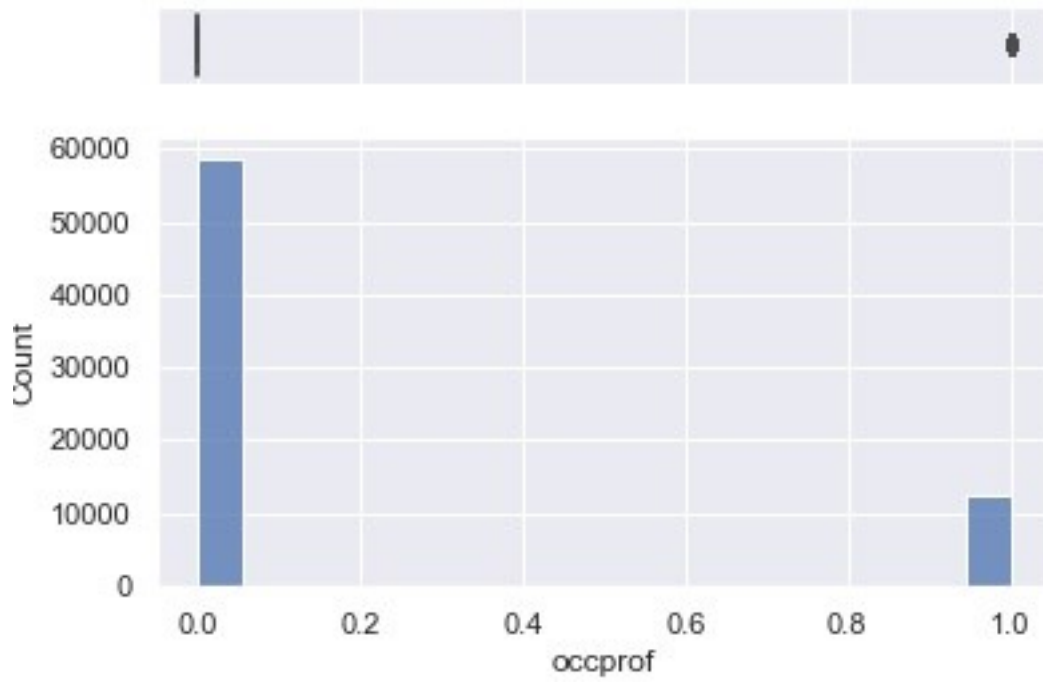
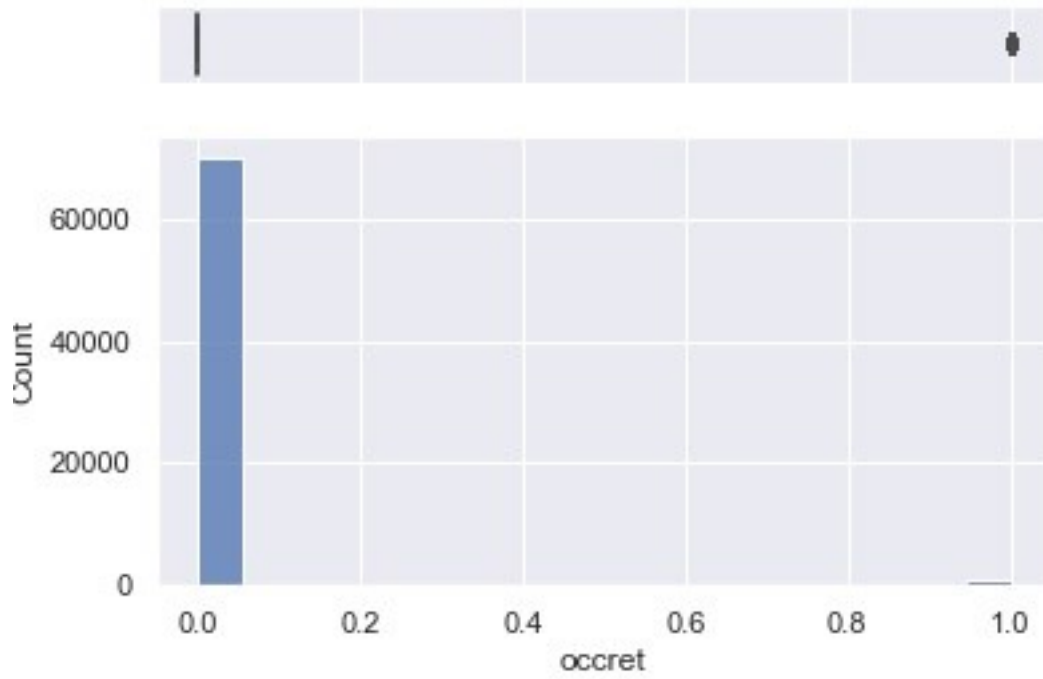


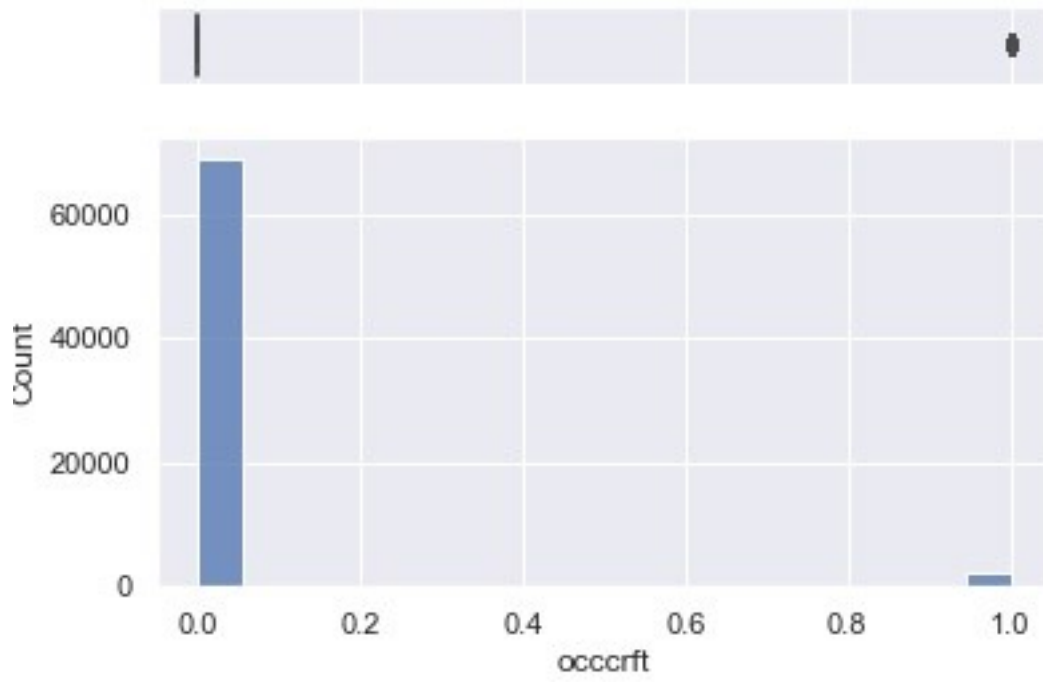
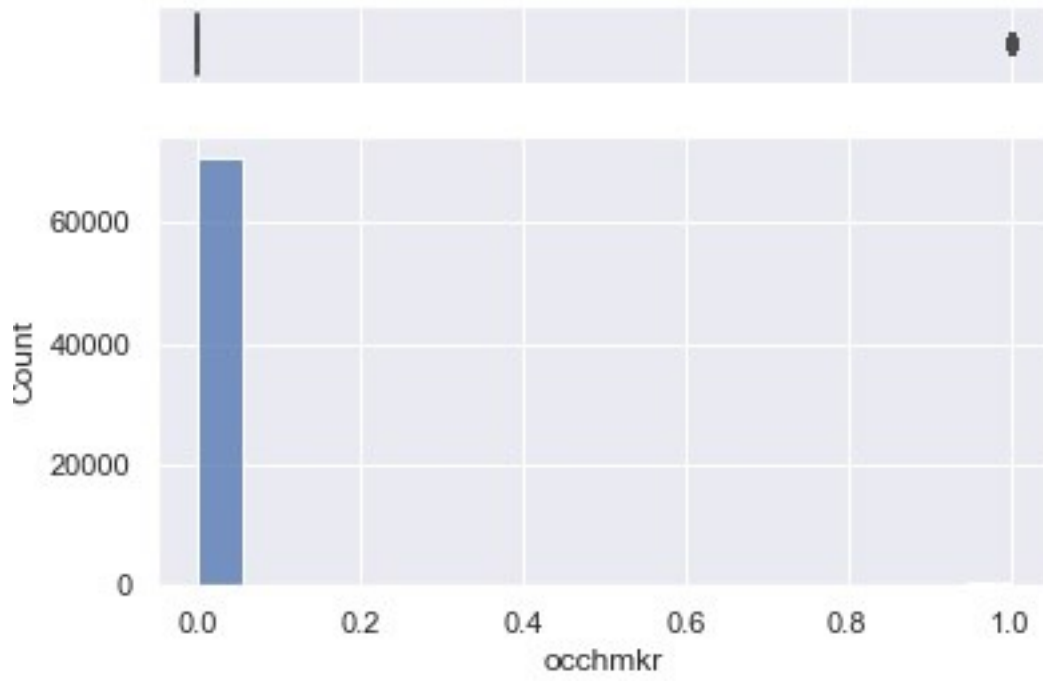


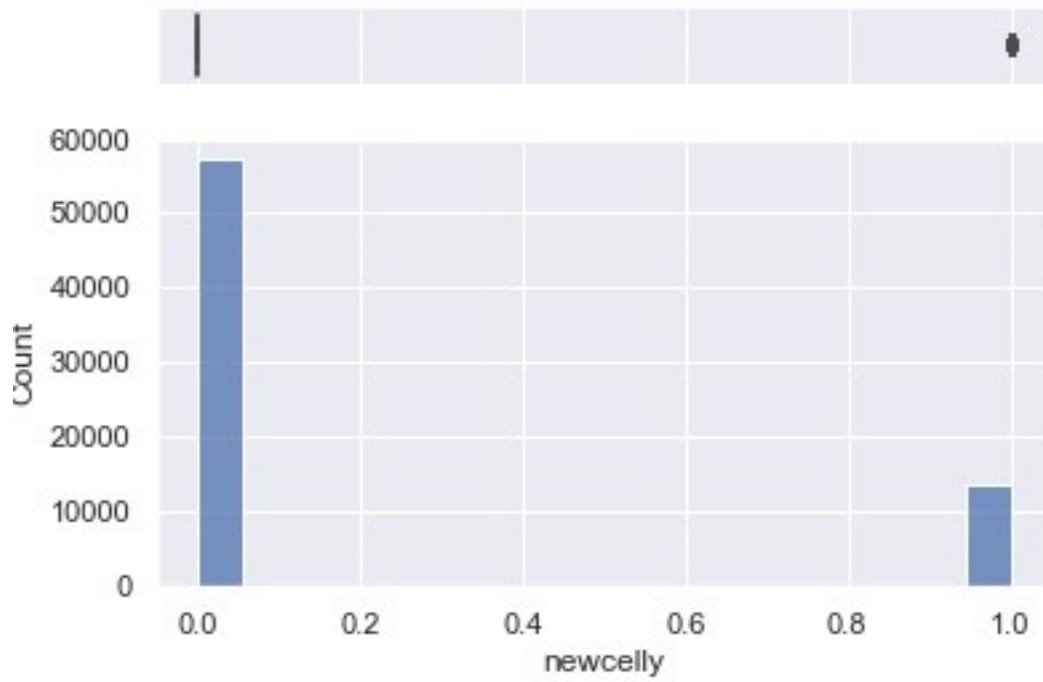
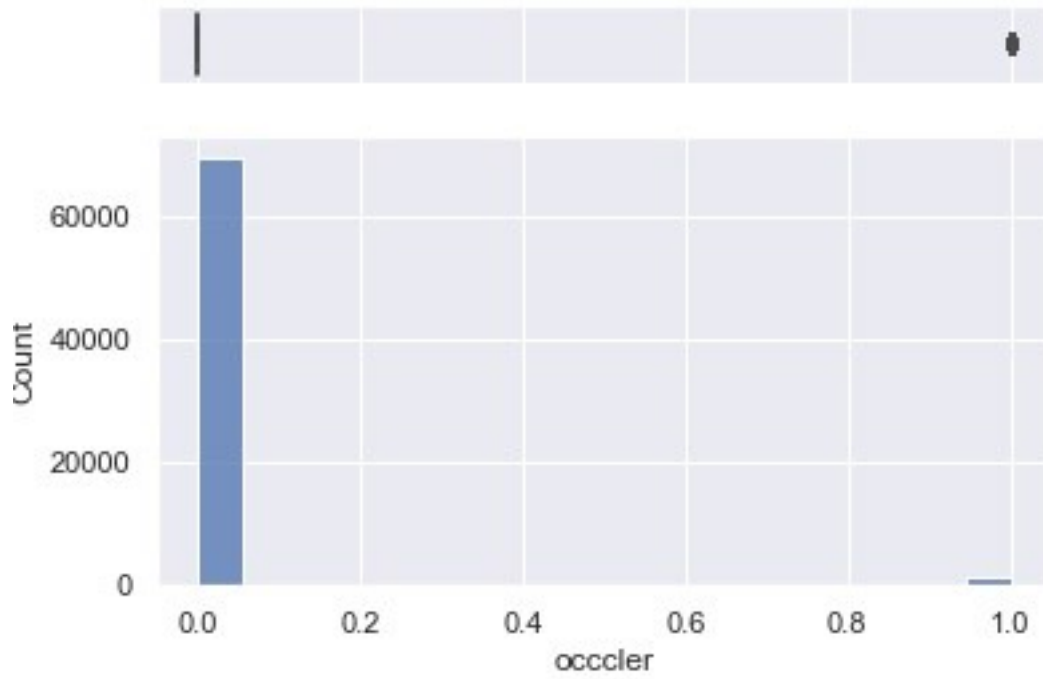


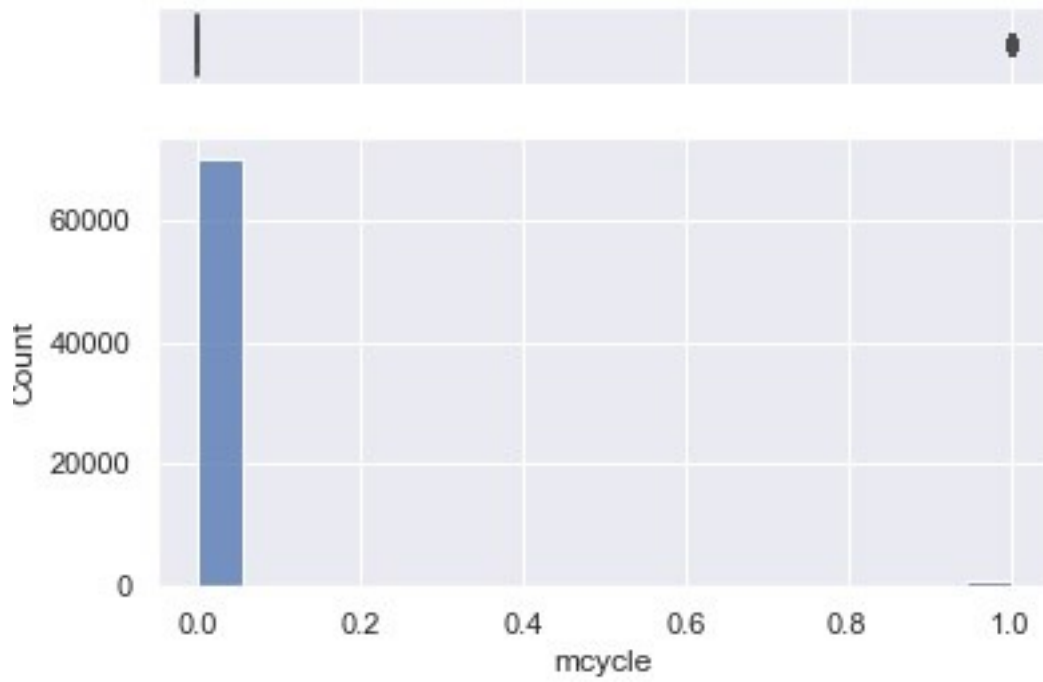
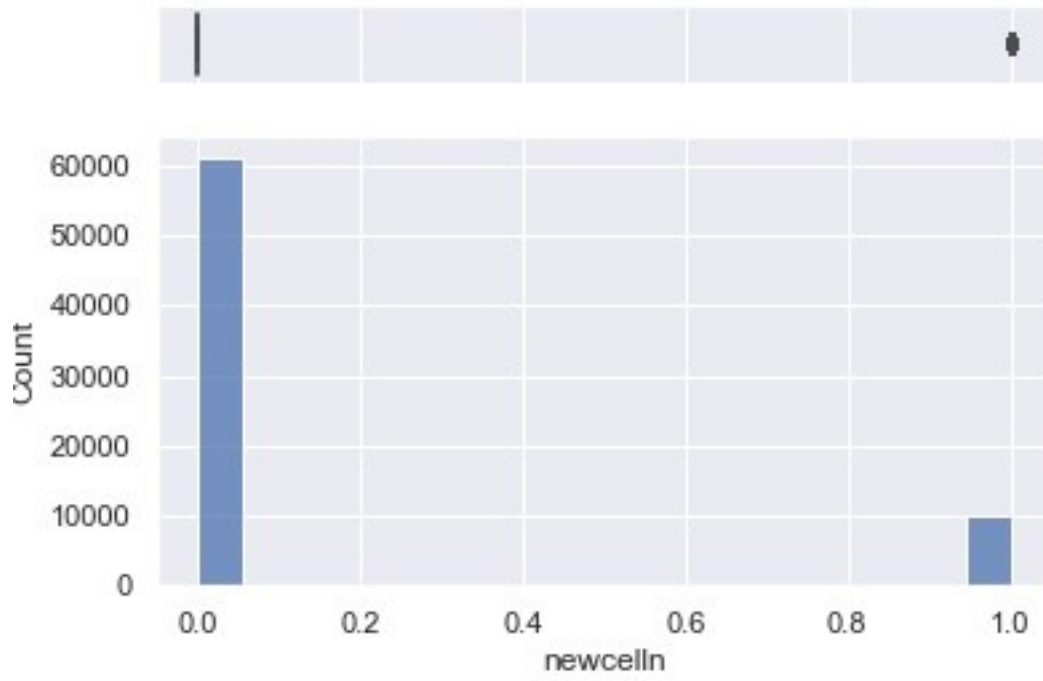


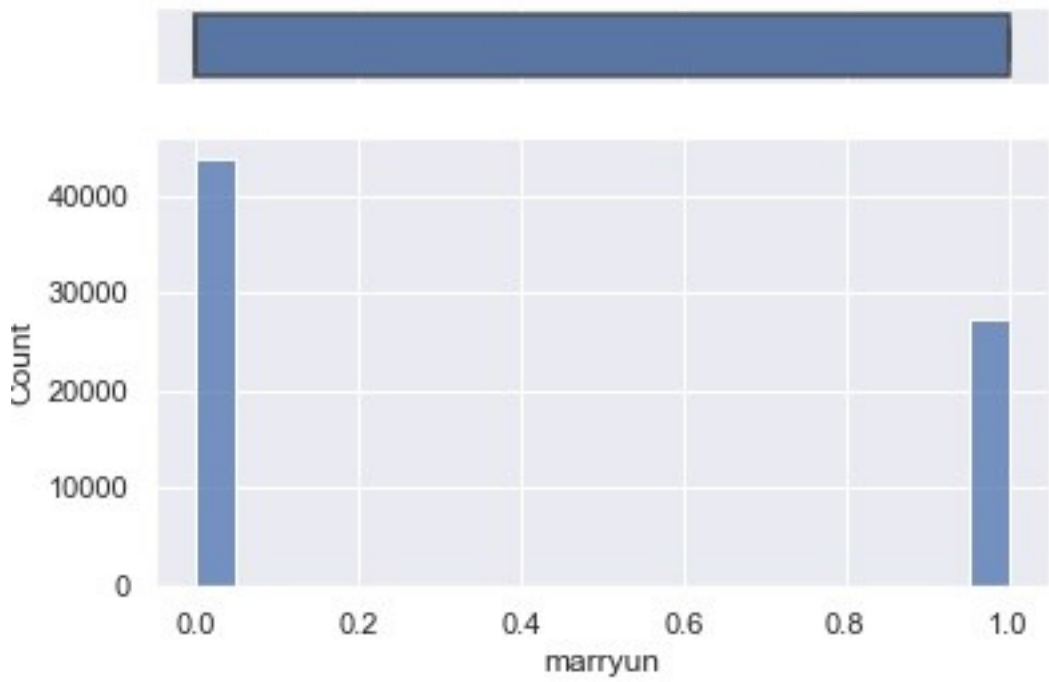
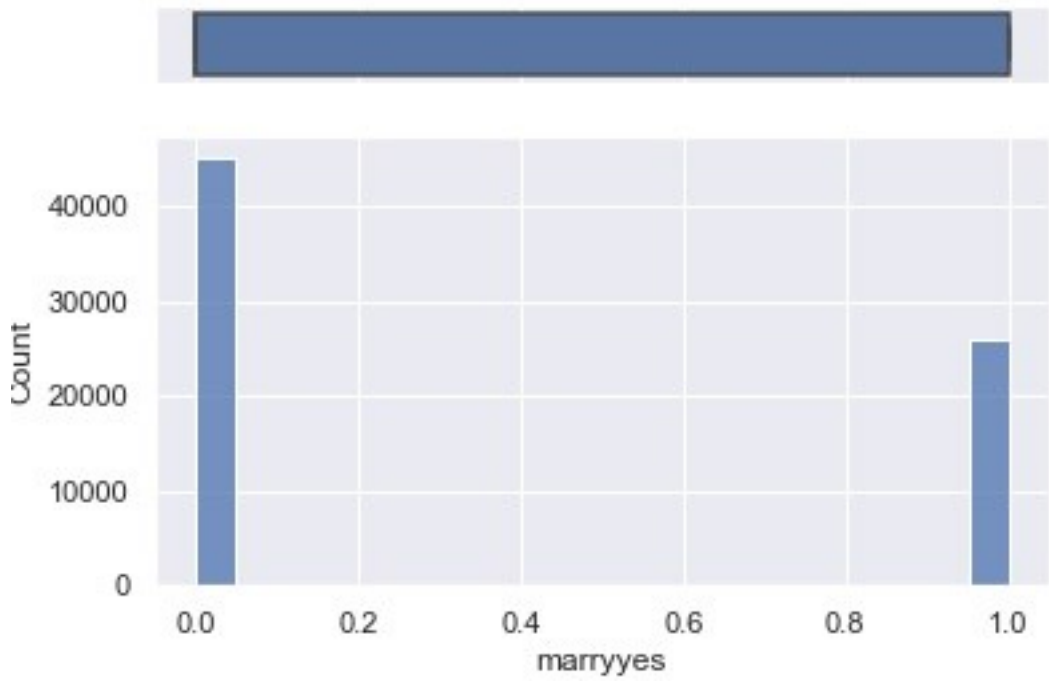


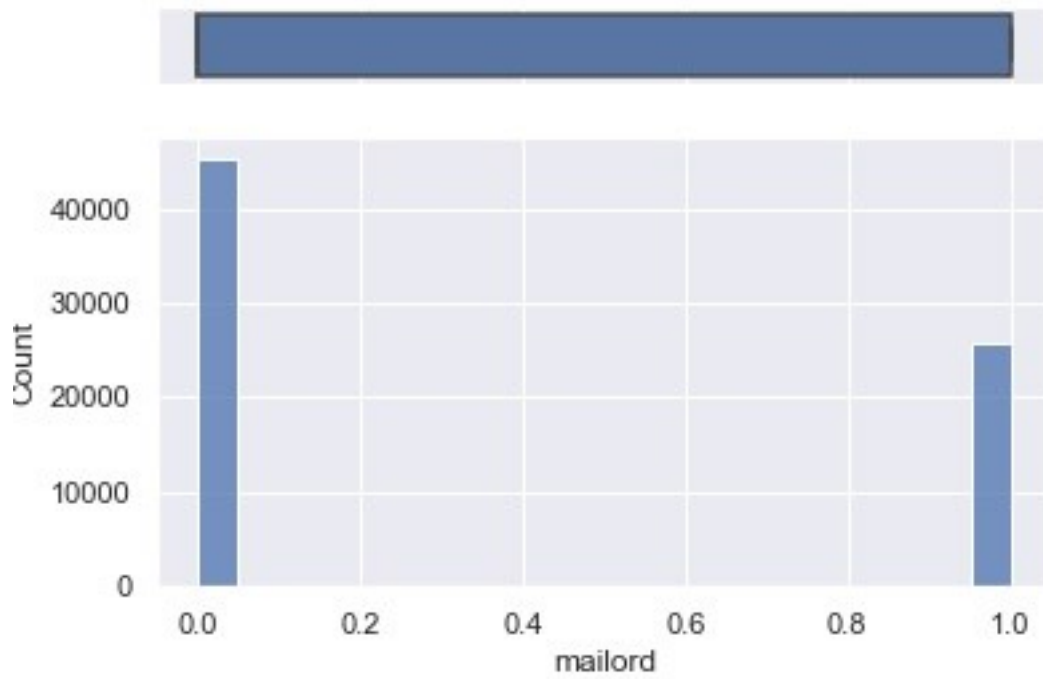
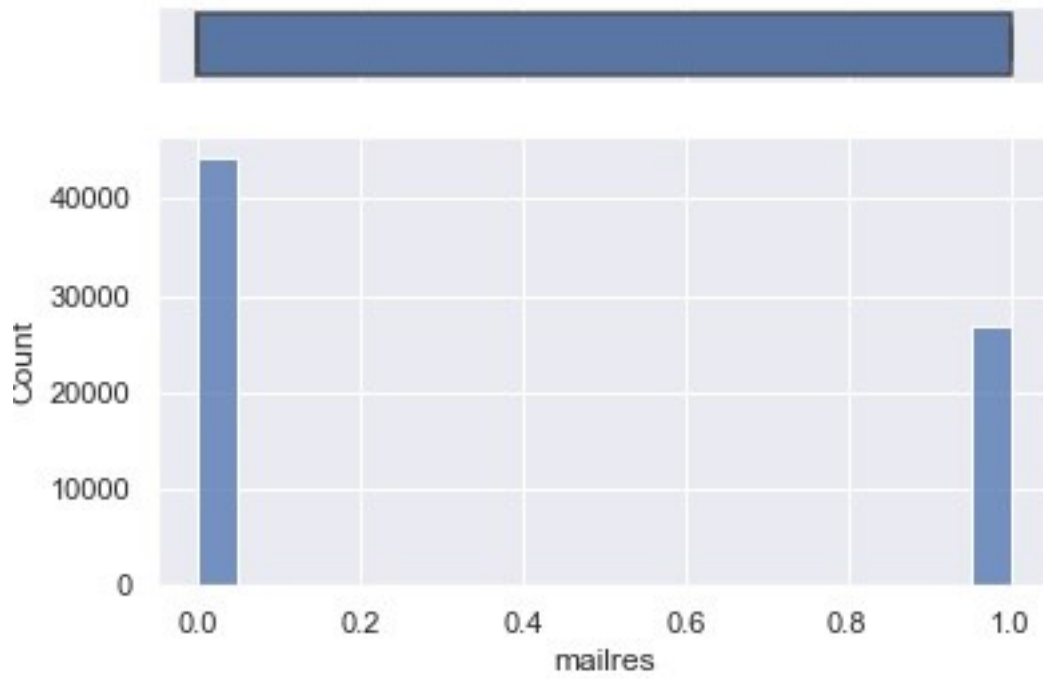


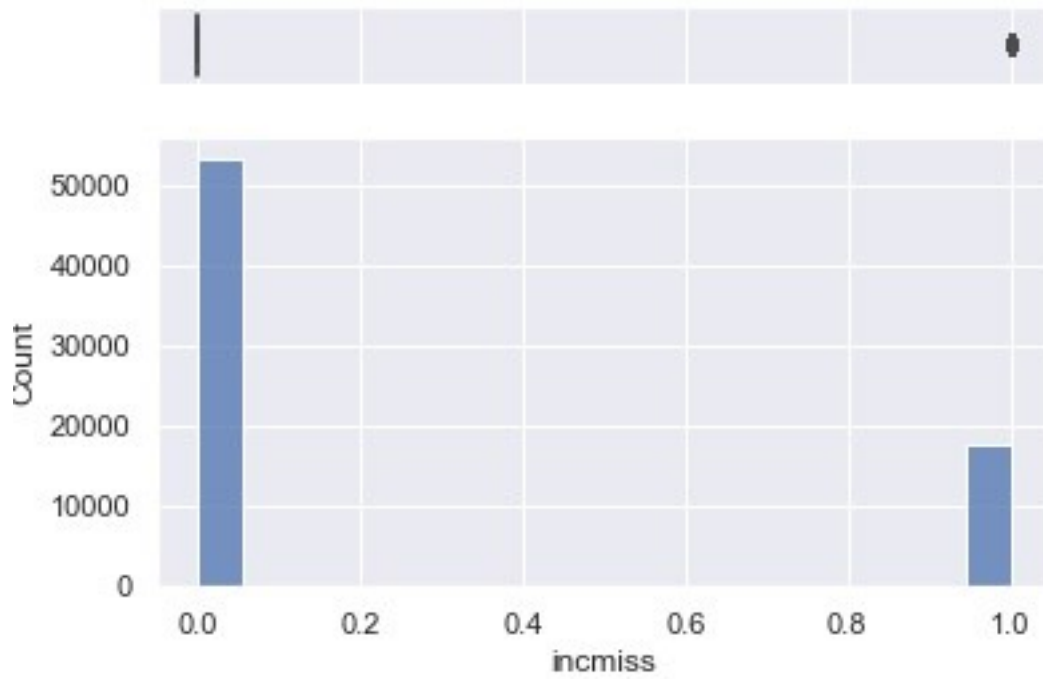
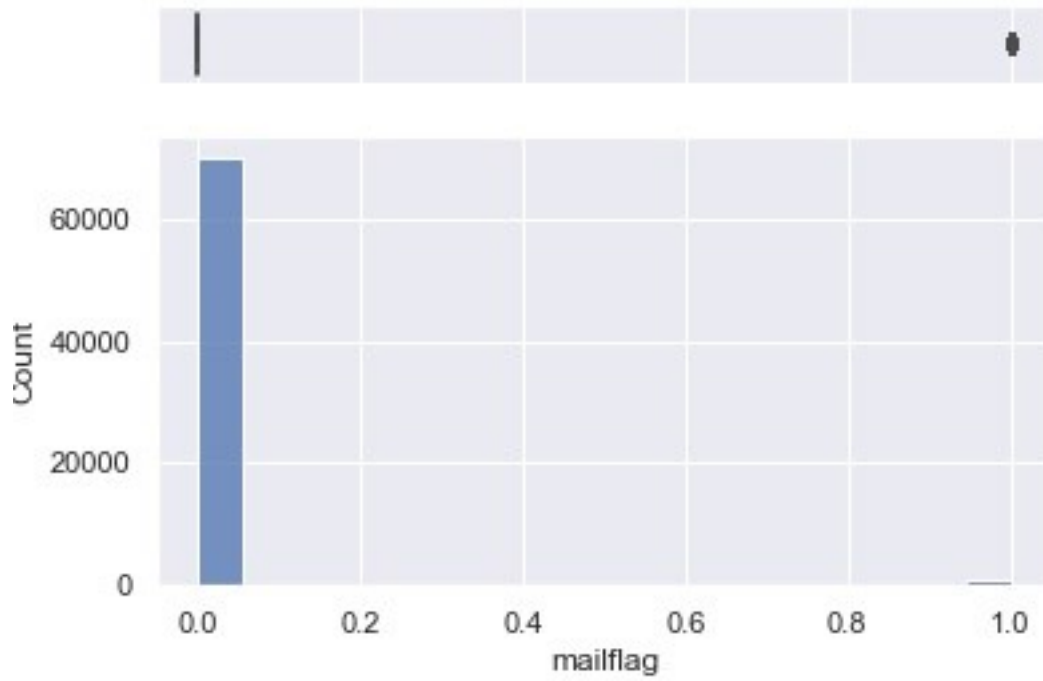


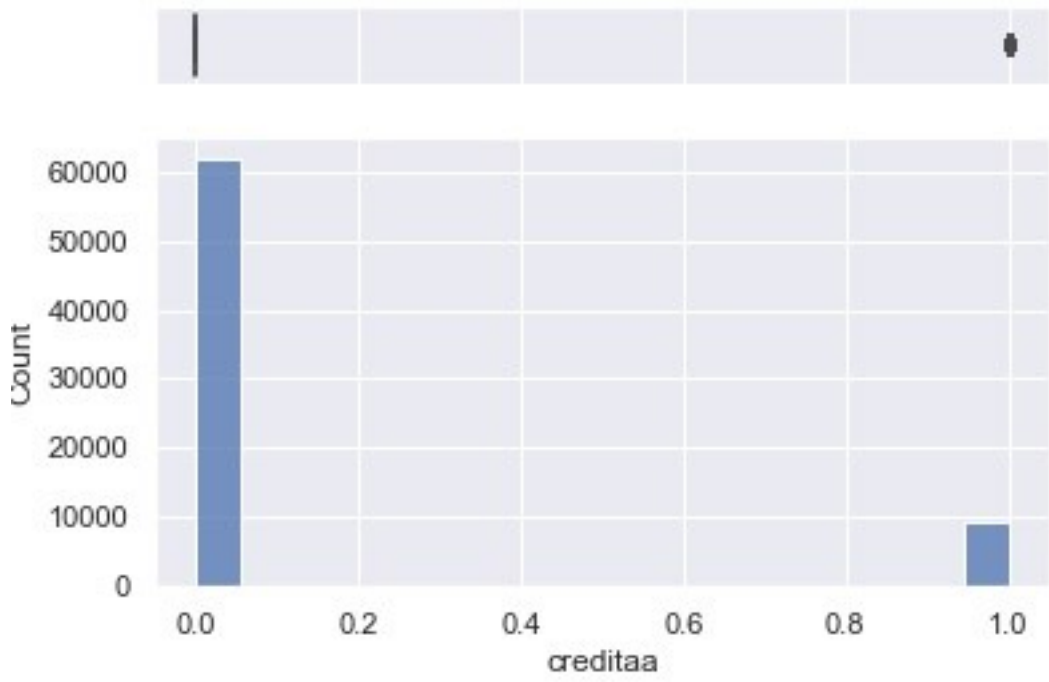
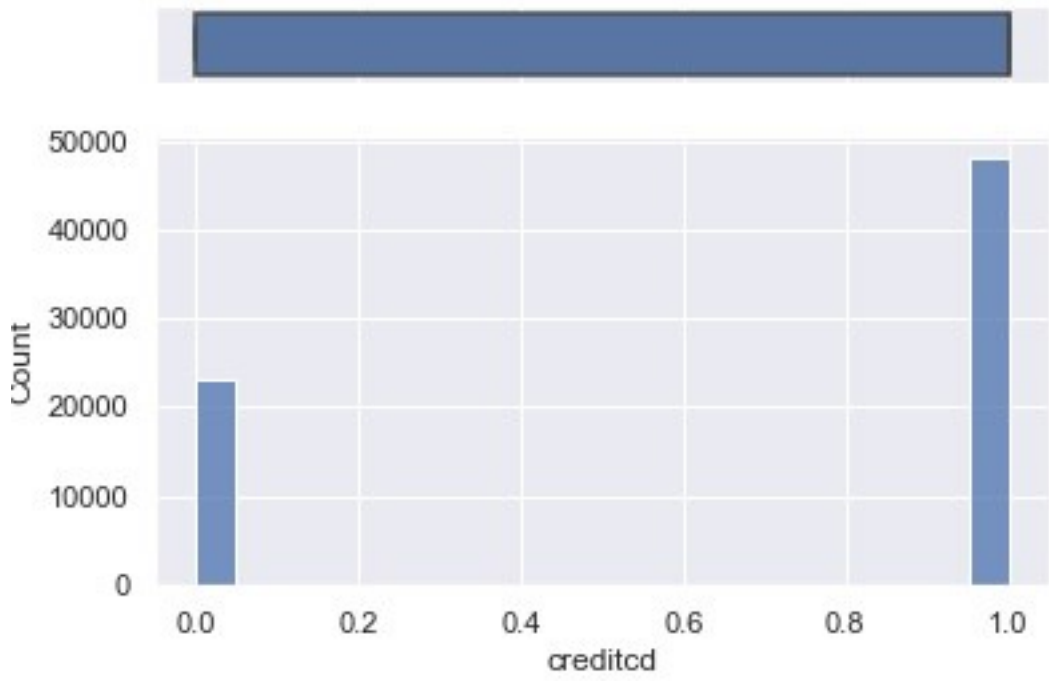


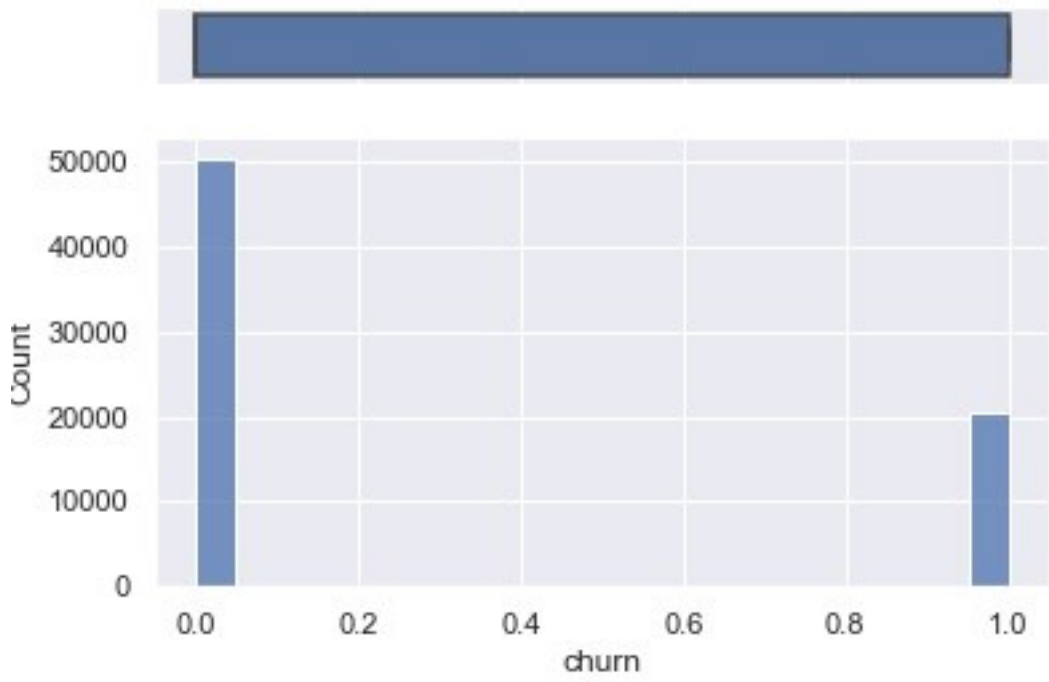
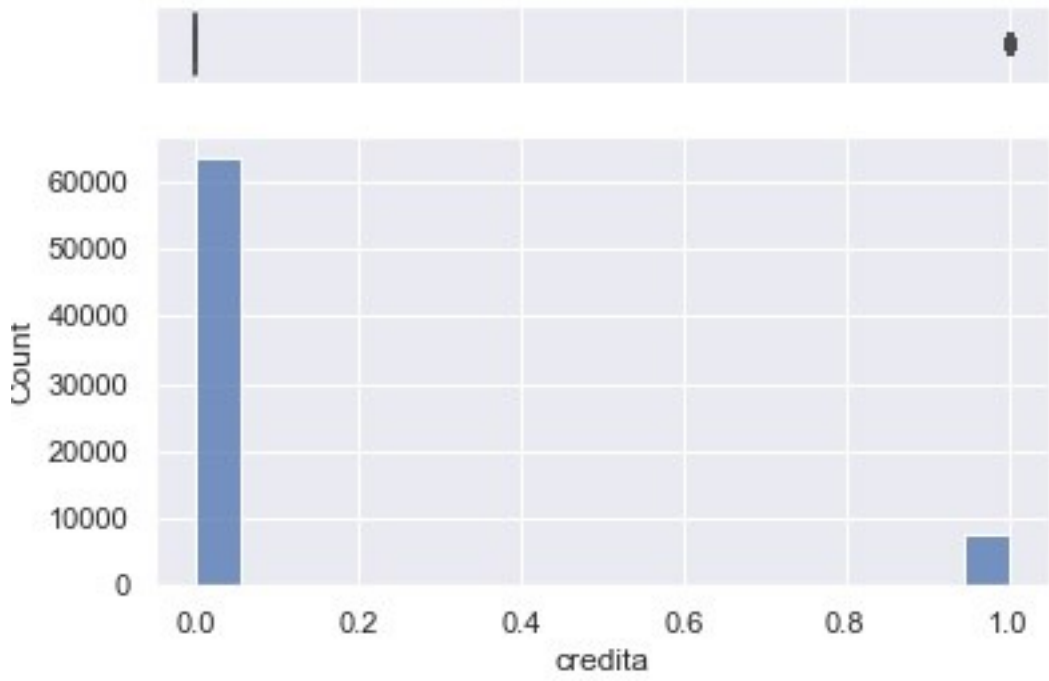


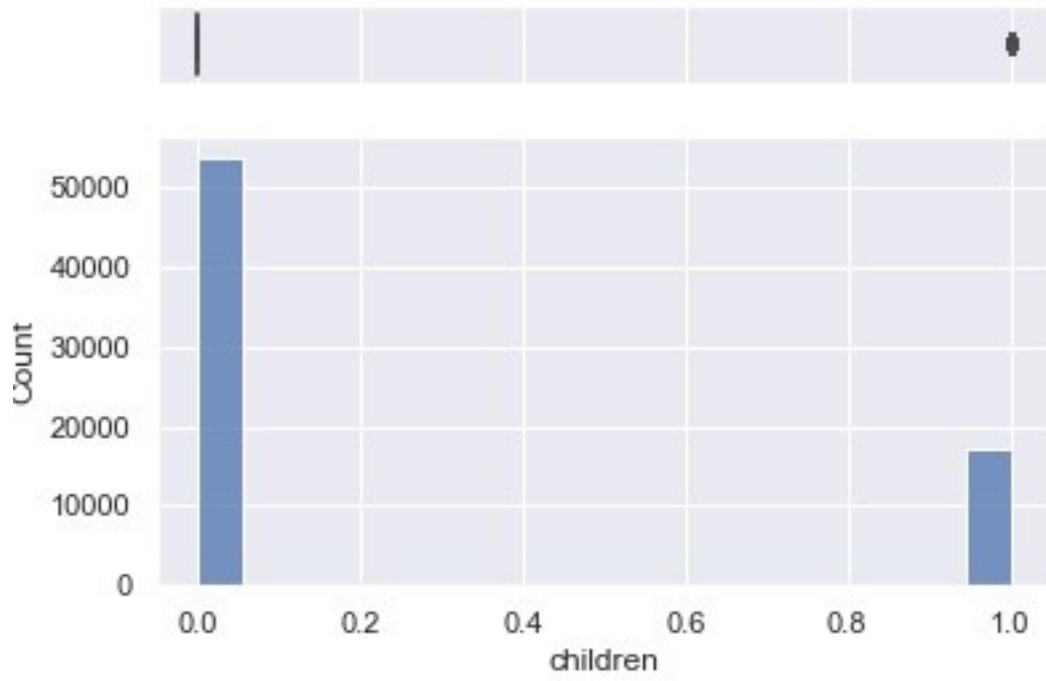












Appendix 02

```
Python code.
TCP Server
# -*- coding: utf-8 -*-
"""
Created on Sun May 23 16:00:54 2021

@author: madhawa
"""

import socket
import sys
import time
import Data_Manage_2 as DM

# Create a TCP/IP socket
sock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)

# Bind the socket to the port

HOST = '127.0.0.1' # Standard loopback interface address (localhost)
PORT = 65432      # Port to listen on (non-privileged ports are >
1023)

server_address = (HOST, PORT)
print (sys.stderr, 'starting up on %s port %s' % server_address)
sock.bind(server_address)

# Listen for incoming connections
sock.listen(1)

while True:
    # Wait for a connection
    # print (sys.stderr, 'waiting for a connection')
    connection, client_address = sock.accept()
    try:
        # print (sys.stderr, 'connection from', client_address)

        # Receive the data in small chunks and retransmit it
        while True:
            data = connection.recv(1024)
            # print (sys.stderr, 'received "%s"' % data)
            if data:
                # print (sys.stderr, 'sending data back to the client')
                txt= data.decode("utf-8")
                D1 = txt.split(", ")
                print("01 txt =====" , txt)
                print("02 Split Data List =====" ,D1)
                if D1[0]=="CLIENT_SEND_CONFIG_DATA":
                    Lower = float(D1[1])
                    Upper = float(D1[2])
                    Ratio = float(D1[3])
                    No_of_Variable =int(D1[4])

                    Customer= ["NULL"]
                    Out_put_data,Out_Fea_List,Load_Parametres =
DM.main(Lower,Upper,Ratio,No_of_Variable,Customer)
```

```

#             Send_data = "SERVER_SEND_CONFIG_OUT_DATA" + ","
+"68" + "," + "71000" + "," + "67" + "," + "25000" + "," + "66" + ","
+ "20000" + "," + "65" + "," + "5000" + "," + "64" + "," + "15000" +
"," + "63" + "," + "11000" + "," + "62" + "," + "4000" + "," + "1" + ","
+"2" + "," + "3" + "," + "4" + "," + "5" + "," + "6"
Send_data1 = "SERVER_SEND_CONFIG_OUT_DATA"
for i in range (26):
    A= Out_put_data[i+1]
    if isinstance(A, str):
        Send_data1=Send_data1+","+A
    else:
        Send_data1=Send_data1+","+str(A)

print("\n\n==== Data Send to Server
===== ", Send_data1 )
connection.sendall(bytes(Send_data1, 'utf-8'))

time.sleep(1)

Numer_FL , Bool_FL = Out_Fea_List
Nume_ls = Numer_FL["Specs"].tolist()
Bool_ls = Bool_FL["Specs"].tolist()

Send_data2 = "SERVER_SEND_FEA_LIST_DATA"
for i in range (len(Nume_ls)):
    A= Nume_ls[i]
    if isinstance(A, str):
        Send_data2=Send_data2+","+A
    else:
        Send_data2=Send_data2+","+str(A)

for i in range (len(Bool_ls)):
    A= Bool_ls[i]
    if isinstance(A, str):
        Send_data2=Send_data2+","+A
    else:
        Send_data2=Send_data2+","+str(A)

connection.sendall(bytes(Send_data2, 'utf-8'))

#'''===== Customer Data LOAD===== '''

if D1[0]=="CUSTMER_DATA_LOAD":
    print("\n\nCUSTMER_DATA_VALI",)
    Lower = float(D1[1])
    Upper = float(D1[2])
    Ratio = float(D1[3])
    No_of_Variable =int(D1[4])

    Customer= ["LOAD"]

    Out_put_data,Out_Fea_List,OUT_Parametres =
DM.main(Lower,Upper,Ratio,No_of_Variable,Customer)

Send_data3 = "CUSTMER_DATA_LOAD_OUT"
for i in range (len(OUT_Parametres)):
    Load= OUT_Parametres[i]
    print('==== ', i , '=====',Load)

```

```

        if isinstance(Load, str):
            Send_data3=Send_data3+" " + Load
        else:

Send_data3=Send_data3+" "+str(round(Load,3))

        print("\n\n==== Data Send to Server
===== ", Send_data3 )
        time.sleep(1)
        connection.sendall(bytes(Send_data3, 'utf-8'))
        time.sleep(1)

#'''===== Customer Data validation=====
'''

        if D1[0]=="CUSTMER_DATA_VALI":
            print("\n\nCUSTMER_DATA_VALI :",txt)
            Lower = float(D1[1])
            Upper = float(D1[2])
            Ratio = float(D1[3])
            No_of_Variable =int(D1[4])

            var_n=[]
            print("lenDi-----",len(D1), '====', ((len(D1)-
5)/2))

            for i in range (int((len(D1)-5)/2)) :
                var_n= var_n+ [str(D1[i+5])]

            var_V=[]
            for i in range (int((len(D1)-5)/2)) :
                var_V=var_V +[ float(D1[i+19])]

            Customer= ["VALID"] + var_n + var_V

            Out_put_data,Out_Fea_List,OUT_Parametres =
DM.main(Lower,Upper,Ratio,No_of_Variable,Customer)

            Send_data4 = "CUSTMER_DATA_VALI_OUT"
            print("\n\n\nOUT_Parametres==== : ",
OUT_Parametres)

            for i in range (len(OUT_Parametres)):
                Load= OUT_Parametres[i]
                print('==== ', i , '=====',Load)
                if isinstance(Load, str):
                    Send_data4=Send_data4+" " + Load
                else:

Send_data4=Send_data4+" "+str(round(Load,3))

#            Send_ACC=""
#            for i in range (5):
#                A= Out_put_data[i+19]
#                if isinstance(A, str):
#                    Send_ACC=Send_ACC+" "+A
#                else:
#                    Send_ACC=Send_ACC+" "+str(A)
#
#            Send_data4=Send_data4+Send_ACC
#

```

```

        print("\n\n ==== Data Send to Server
===== ", Send_data4 )
        time.sleep(1)
        connection.sendall(bytes(Send_data4, 'utf-8'))
        time.sleep(1)

    finally:
        # Clean up the connection
        connection.close()

Data management module
# -*- coding: utf-8 -*-
"""
Created on Thu Jun 17 08:07:30 2021

@author: madhawa
"""

#
=====
=====
""" This module is toload the Data Set and

    """
#
=====
=====

import numpy as np
import pandas as pd
import xlswriter

#import Visulization as vis
import Feature_selction as fea_sel
import Gaussian_Class_1 as Gas_Cla

from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn import metrics

from sklearn.decomposition import PCA

import sys
import warnings

from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn import tree
from sklearn import svm
from sklearn.ensemble import VotingClassifier
from sklearn.model_selection import cross_val_score

if not sys.warnoptions:
    warnings.simplefilter("ignore")

```

```

def main(Lower,Upper,Ratio,No_of_Variable,Customer):

    Lower= Lower
    Upper= Upper

    Ratio=Ratio
    No_of_Variable=No_of_Variable

    missing_values = ["n/a", "na", "--"]
    df = pd.read_csv("data_set_21K.csv" ,delimiter=',',na_values =
missing_values)
    DF_Shape = df.shape
    (DF_No_Row ,DF_No_Col)=df.shape
    df = df.apply(pd.to_numeric,errors='coerce')
    Column_List=df.columns.tolist()
    print("Original Data Set Size",df.shape)
    np_index=Column_List[1]
    np_body=np.append(Column_List[6:],Column_List[4])
    df= df[np.append(np_index,np_body)]

    print("Remove irrelevant coumns",df.shape)

    Bool_Column =
["children","credita","creditaa","prizmrur","prizmub","prizmtwn","refur
b","webcap","truck","rv","occprof","occcler","occcrft","occstud","occhm
kr","occret","occsself","ownrent","marryun","marryyes","mailord","mailre
s","mailflag","travel","pcown","creditcd","newcelly","newcelln","incmis
s","mcycle","setprcm","retcall","churn"]
    Num_Column =
["revenue","mou","recchrge","directas","overage","roam","changem","chan
ger","dropvce","blckvce","unansvce","custcare","threeway","mourec","out
calls","incalls","peakvce","opeakvce","dropblk","callfwdv","callwait","
months","uniqsubs","actvsbs","phones","models","eqpdays","age1","age2"
,"retcalls","retacct","refer","income","setprc"]

    row = 0
    col = 0

    workbook = xlswriter.Workbook("Report_06_18"+'.xlsx')
    worksheet = workbook.add_worksheet("INI_DET")

    cell_format0 = workbook.add_format({'bold': True, 'font_color':
'red','border': True})
    cell_format1 = workbook.add_format({'bold': True, 'font_color':
'blue','border': True})

    worksheet.write(row, col, "Numerical
ini",workbook.add_format({'bold': True, 'font_color': 'black','border':
True}))
    col = 4
    worksheet.write(row, col, "Normalized-Numerical
ini",workbook.add_format({'bold': True, 'font_color': 'black','border':
True}))
    col = 8

```

```

worksheet.write(row, col, "Normalized-Num ini -LWD
Outlier",workbook.add_format({'bold': True, 'font_color':
'black','border': True}))
col = 0

##### x = x[x.between(x.quantile(.15), x.quantile(.85))]

row += 2

df_n =df.copy()
df_n[Num_Column] = df_n[Num_Column].apply(lambda x: (x - x.min()) /
(x.max() - x.min()))

df_n_o= df_n.copy()
df_n_o= df_n_o.loc[ : , ["X"]+Num_Column]

Q1 = df_n_o.quantile(0.25)
Q3 = df_n_o.quantile(0.75)
IQR = Q3 - Q1

df_n_o = df_n_o[~((df_n_o < (Q1 - Lower * IQR)) |(df_n_o > (Q3 +
Upper * IQR))).any(axis=1)]
df_Numeric_Out_rem =df_n_o
df_Boo1 =df_n_o.copy()
df_Boo1 =df_Boo1.loc[ : , ["X"]+Boo1_Column]

df_n_o =pd.merge(df_Numeric_Out_rem, df_Boo1, on='X')

"""df_n_o is row data set -->> Normalized (0,1) --->> remove
Outliers based on Q+ IQR*Factor """

for i in Num_Column:
# vis.Plot_His_and_Box(df,i)
df_temp = df[i]
df_sub = df_temp.describe()
Values = df_sub.values.tolist()+[ df_temp.isnull().sum()]
Data_lable = ["count","mean","std","min","Q1-25%","Q2-50%","Q3-
75%","max","null"]

df_temp_n = df_n[i]
df_sub_n = df_temp_n.describe()
Values_n = df_sub_n.values.tolist()+[ df_temp_n.isnull().sum()]

df_temp_n_o = df_n_o[i]
df_sub_n_o = df_temp_n_o.describe()
Values_n_o = df_sub_n_o.values.tolist()+[
df_temp_n_o.isnull().sum()]

worksheet.write(row, col, i,cell_format0)
col = 4
worksheet.write(row, col, i,cell_format0)
col = 8
worksheet.write(row, col, i,cell_format0)
col = 0

row += 1
for j in range(len(Data_lable)):
worksheet.write(row, col, Data_lable[j],cell_format1)

```

```

        worksheet.write(row, col + 1, Values[j],cell_format1)
        col = 4
        worksheet.write(row, col, Data_lable[j],cell_format1)
        worksheet.write(row, col + 1, Values_n[j],cell_format1)
        col = 8
        worksheet.write(row, col, Data_lable[j],cell_format1)
        worksheet.write(row, col + 1, Values_n_o[j],cell_format1)

        col = 0

#         print (i," === ",Data_lable[j]," ", " ",Values[j])
        row += 1

        row += 2

row = 0
col = 12
worksheet.write(row, col, "Bool before
process",workbook.add_format({'bold': True, 'font_color':
'black','border': True}))

row += 2

for i in Bool_Column:
#     vis.Plot_His_and_Box(df,i)
    df_temp = df[i]
    df_sub = df_temp.describe()
    Values_temp = df_sub.values.tolist()
    Count_1 = df_temp.sum()
    Count_0 = Values_temp[0]-df_temp.sum()
    mode = df_temp.mode()
    Null= df_temp.isnull().sum()
    Data_lable = ["count","mean","std","1-count","0-
Count","mode","",""," ", "Null "]
    Values = Values_temp[0:3]+ [Count_1,Count_0,mode," ", " ",Null]
    worksheet.write(row, col, i,cell_format0)

    row += 1
    for j in range(len(Data_lable)):
        worksheet.write(row, col, Data_lable[j],cell_format1)
        worksheet.write(row, col + 1, Values[j],cell_format1)
#         print (i," === ",Data_lable[j]," ", " ",Values[j])
        row += 1

    row += 2

#for i in Bool_Column:
#     df_temp = df[i]
#     print (df_temp.describe())

#for i in Num_Column:
#     vis.Plot_His_and_Box(df,i)
#
#for i in Bool_Column:
#     vis.Plot_His_and_Box(df,i)
#
#
#

```



```

=====
List Wise Deltion
=====
"""

df_LD = df.copy()
df_LD_n = df.copy()
df_LD.dropna(inplace=True)
df_LD_n.dropna(inplace=True)

df_LD_n[Num_Column] = df_LD_n[Num_Column].apply(lambda x: (x -
x.min()) / (x.max() - x.min()))

df_LD_o= df_LD_n.copy()
df_LD_o= df_LD_o.loc[ : , ["X"]+Num_Column]

Q1 = df_LD_o.quantile(0.25)
Q3 = df_LD_o.quantile(0.75)
IQR = Q3 - Q1

df_LD_o = df_LD_o[~((df_LD_o < (Q1 - Lower * IQR)) |(df_LD_o > (Q3
+ Upper * IQR))).any(axis=1)]
df_Numeric_Out_rem_LD =df_LD_o
df_Bool_LD =df_LD.copy()
df_Bool_LD =df_Bool_LD.loc[ : , ["X"]+Bool_Column]

df_LD_o =pd.merge(df_Numeric_Out_rem_LD, df_Bool_LD, on='X')

"""df_LD_o is row data set---->> Remove missing value List wise
deltion
-->> Normalized (0,1) --->> remove Outliers based on Q+ IQR*Factor
"""

#df_LD_n= df_LD_n.div(df_LD_n.max(),1)

worksheet1 = workbook.add_worksheet("List_DEL")

row = 0
col = 0

worksheet1.write(row, col, "Numerical list wise
deletion",workbook.add_format({'bold': True, 'font_color':
'black','border': True}))
col = 4
worksheet1.write(row, col, "Numerical list wise deletion-
>Normlaized-",workbook.add_format({'bold': True, 'font_color':
'black','border': True}))
col = 8
worksheet1.write(row, col, "Numerical list wise deletion-
>Normlaized-> outlier",workbook.add_format({'bold': True, 'font_color':
'black','border': True}))
col = 0

```

```

row += 2

for i in Num_Column:
    # vis.Plot_His_and_Box(df,i)
    df_temp = df_LD[i]
    df_sub = df_temp.describe()
    Values = df_sub.values.tolist()+[ df_temp.isnull().sum()]
    Data_lable = ["count","mean","std","min","Q1-25%","Q2-50%","Q3-
75%","max","null"]

    df_temp_n = df_LD_n[i]
    df_sub_n = df_temp_n.describe()
    Values_n = df_sub_n.values.tolist()+[ df_temp_n.isnull().sum()]

    df_temp_n_o = df_LD_o[i]
    df_sub_n_o = df_temp_n_o.describe()
    Values_n_o = df_sub_n_o.values.tolist()+[
df_temp_n_o.isnull().sum()]

worksheet1.write(row, col, i,cell_format0)
col = 4
worksheet1.write(row, col, i,cell_format0)
col = 8
worksheet1.write(row, col, i,cell_format0)
col = 0

row += 1
for j in range(len(Data_lable)):
    worksheet1.write(row, col, Data_lable[j],cell_format1)
    worksheet1.write(row, col + 1, Values[j],cell_format1)

    col = 4
    worksheet1.write(row, col, Data_lable[j],cell_format1)
    worksheet1.write(row, col + 1, Values_n[j],cell_format1)
    col = 8
    worksheet1.write(row, col, Data_lable[j],cell_format1)
    worksheet1.write(row, col + 1, Values_n_o[j],cell_format1)

    col = 0

    row += 1

row += 2

row = 0
col = 12
worksheet1.write(row, col, "Bool list wise
deletion",workbook.add_format({'bold': True, 'font_color':
'black','border': True}))
row += 2

for i in Bool_Column:
    # vis.Plot_His_and_Box(df,i)
    df_temp = df_LD[i]
    df_sub = df_temp.describe()
    Values_temp = df_sub.values.tolist()

```

```

Count_1 = df_temp.sum()
Count_0 = Values_temp[0]-df_temp.sum()
mode = df_temp.mode()
Null= df_temp.isnull().sum()
Data_lable = ["count","mean","std","1-count","0-
Count","mode","","","Null "]
Values = Values_temp[0:3]+ [Count_1,Count_0,mode," "," ",Null]
worksheet1.write(row, col, i,cell_format0)

row += 1
for j in range(len(Data_lable)):
    worksheet1.write(row, col, Data_lable[j],cell_format1)
    worksheet1.write(row, col + 1, Values[j],cell_format1)
#     print (i," === ",Data_lable[j]," , ",Values[j])
    row += 1

row += 2

```

```

=====
=====

```

Mean , Mode imputation

```

=====
=====

```

```

df_MMI = df.copy()
df_MMI_n = df.copy()
df_MMI_n[Num_Column] = df_MMI_n[Num_Column].apply(lambda x: (x -
x.min()) / (x.max() - x.min()))

df_MMI_o= df_MMI_n.copy()
df_MMI_o= df_MMI_o.loc[ : , ["X"]+Num_Column]

Q1 = df_MMI_o.quantile(0.25)
Q3 = df_MMI_o.quantile(0.75)
IQR = Q3 - Q1

df_MMI_o = df_MMI_o[~((df_MMI_o < (Q1 - Lower * IQR)) |(df_MMI_o >
(Q3 + Upper * IQR))).any(axis=1)]
df_Numeric_Out_rem_MMI =df_MMI_o
df_Bool_MMI =df_MMI.copy()
df_Bool_MMI =df_Bool_MMI.loc[ : , ["X"]+Bool_Column]

df_MMI_o =pd.merge(df_Numeric_Out_rem_MMI, df_Bool_MMI, on='X')

"""df_MMI_o is row data set---->> Mean mode impute
-->> Normalized (0,1) ---->> remove Outliers based on Q+ IQR*Factor
"""

#df_MMI_n= df_MMI_n.div(df_MMI_n.max(),1)

worksheet1 = workbook.add_worksheet("M_M Impute")

```

```

row = 0
col = 0

worksheet1.write(row, col, "Numerical Mean mode
Imputation",workbook.add_format({'bold': True, 'font_color':
'black','border': True}))
col = 4
worksheet1.write(row, col, "Numerical Mean mode Imputation-
Normalized",workbook.add_format({'bold': True, 'font_color':
'black','border': True}))
col = 8
worksheet1.write(row, col, "Numerical Mean mode Imputation-
Normalized-Outlier",workbook.add_format({'bold': True, 'font_color':
'black','border': True}))

col = 0
row += 2

for i in Num_Column:
# vis.Plot_His_and_Box(df,i)
df_MMI[i].fillna(df_MMI[i].mean(), inplace=True)
df_temp=df_MMI[i]
df_sub = df_temp.describe()
Values = df_sub.values.tolist()+[ df_temp.isnull().sum()]
Data_label = ["count","mean","std","min","Q1-25%","Q2-50%","Q3-
75%","max","null"]

df_MMI_n[i].fillna(df_MMI_n[i].mean(), inplace=True)
df_temp_n=df_MMI_n[i]
df_sub_n = df_temp_n.describe()
Values_n = df_sub_n.values.tolist()+[ df_temp_n.isnull().sum()]

df_MMI_o[i].fillna(df_MMI_o[i].mean(), inplace=True)
df_temp_o=df_MMI_o[i]
df_sub_o = df_temp_o.describe()
Values_o = df_sub_o.values.tolist()+[ df_temp_o.isnull().sum()]

worksheet1.write(row, col, i,cell_format0)
col = 4
worksheet1.write(row, col, i,cell_format0)
col = 8
worksheet1.write(row, col, i,cell_format0)

col = 0

row += 1
for j in range(len(Data_label)):
worksheet1.write(row, col, Data_label[j],cell_format1)
worksheet1.write(row, col + 1, Values[j],cell_format1)

col = 4
worksheet1.write(row, col, Data_label[j],cell_format1)
worksheet1.write(row, col + 1, Values_n[j],cell_format1)

col = 8
worksheet1.write(row, col, Data_label[j],cell_format1)
worksheet1.write(row, col + 1, Values_o[j],cell_format1)

```

```

        col = 0

        row += 1

    row += 2

row = 0
col = 12
worksheet1.write(row, col, "Bool list wise
deletion",workbook.add_format({'bold': True, 'font_color':
'black','border': True}))
row += 2

for i in Bool_Column:
#     vis.Plot_His_and_Box(df,i)
df_MMI[i].fillna(df_MMI[i].mode(), inplace=True)
df_temp=df_MMI[i]

df_sub = df_temp.describe()
Values_temp = df_sub.values.tolist()
Count_1 = df_temp.sum()
Count_0 = Values_temp[0]-df_temp.sum()
mode = df_temp.mode()
Null= df_temp.isnull().sum()
Data_lable = ["count", "mean", "std", "1-count", "0-
Count", "mode", "", " ", "Null "]
Values = Values_temp[0:3]+ [Count_1,Count_0,mode, " ", " ",Null]
worksheet1.write(row, col, i,cell_format0)

row += 1
for j in range(len(Data_lable)):
    worksheet1.write(row, col, Data_lable[j],cell_format1)
    worksheet1.write(row, col + 1, Values[j],cell_format1)
#     print (i," === ",Data_lable[j]," ", " ",Values[j])
    row += 1

row += 2

print("Before imputation")
print(df.shape)
print("After imputation")
print(df_MMI.shape)

Raw_Data = df.copy()

Raw_Norm = df_n.copy()
Raw_Norm_OutLier=df_n_o.copy()

Lst_Del_Norm= df_LD_n.copy()
Lst_Del_Norm_OutLier =df_LD_o.copy()

MeMo_Imp_Norm= df_MMI_n.copy()
MeMo_Imp_Norm_OutLier= df_MMI_o.copy()

print("\n\n")
print ("Raw_Data===== : ",Raw_Data.shape )
print ("Raw_Norm===== : ",Raw_Norm.shape )
print ("Raw_Norm_OutLier===== :
",Raw_Norm_OutLier.shape )

```

```

    print ("Lst_Del_Norm===== : ",Lst_Del_Norm.shape )
    print ("Lst_Del_Norm_OutLier===== :
",Lst_Del_Norm_OutLier.shape )
    print ("MeMo_Imp_Norm===== : ",MeMo_Imp_Norm.shape )
    print ("MeMo_Imp_Norm_OutLier===== :
",MeMo_Imp_Norm_OutLier.shape )

worksheet1 = workbook.add_worksheet("Fea_List")

col = 0
row = 0

worksheet1.write(row, col, "List_Delete_Normlized Numerical
",cell_format0)
col = 4
worksheet1.write(row, col, "List_Delete_Normlized Bool
",cell_format0)
col = 0

row += 1
FS_Lst_Del_Norm = fea_sel.Anova_Feature_selction (
Lst_Del_Norm.loc[ : , Num_Column+["churn"]],Ratio,No_of_Variable)
FS_Lst_Del_Bool = fea_sel.Chi_Feature_selction (
Lst_Del_Norm.loc[ : , Bool_Column+["churn"]],Ratio,No_of_Variable+1)
for i in range(len(FS_Lst_Del_Norm["Specs"])):
    A=FS_Lst_Del_Norm["Specs"]
    B=FS_Lst_Del_Norm["Score"]
    worksheet1.write(row, col , A.iloc[i] ,cell_format1)
    worksheet1.write(row, col + 1, B.iloc[i] ,cell_format1)

col = 4
A=FS_Lst_Del_Bool["Specs"]
B=FS_Lst_Del_Bool["Score"]
if (B.iloc[i]==np.inf):
    B="inf"
else:
    B=B.iloc[i]
    worksheet1.write(row, col , A.iloc[i] ,cell_format1)
    worksheet1.write(row, col + 1, B ,cell_format1)
col = 0
row += 1

row+=2
worksheet1.write(row, col, "List_Delete_Normlized Outlier Numerical
",cell_format0)
col = 4
worksheet1.write(row, col, "List_Delete_Normlized Outlier Bool
",cell_format0)
col = 0
row += 1
FS_Lst_Del_Norm_out = fea_sel.Anova_Feature_selction (
Lst_Del_Norm_OutLier.loc[ : ,
Num_Column+["churn"]],Ratio,No_of_Variable)
FS_Lst_Del_Bool_out = fea_sel.Chi_Feature_selction (
Lst_Del_Norm_OutLier.loc[ : ,
Bool_Column+["churn"]],Ratio,No_of_Variable+1)
for i in range(len(FS_Lst_Del_Norm_out["Specs"])):

```

```

A=FS_Lst_Del_Norm_out["Specs"]
B=FS_Lst_Del_Norm_out["Score"]
worksheet1.write(row, col, A.iloc[i], cell_format1)
worksheet1.write(row, col + 1, B.iloc[i], cell_format1)

col = 4
A=FS_Lst_Del_Bool_out["Specs"]
B=FS_Lst_Del_Bool_out["Score"]
if (B.iloc[i]==np.inf):
    B="inf"
else:
    B=B.iloc[i]
worksheet1.write(row, col, A.iloc[i], cell_format1)
worksheet1.write(row, col + 1, B, cell_format1)
col = 0

row += 1

row+=2
worksheet1.write(row, col, "Mean ,Mode impute Normlized Numerical", cell_format0)
col = 4
worksheet1.write(row, col, "Mean ,Mode impute Normlized Bool", cell_format0)
col = 0

row += 1
FS_MMI_Norm = fea_sel.Anova_Feature_selction (
MeMo_Imp_Norm.loc[:, Num_Column+"churn"],Ratio,No_of_Variable)
FS_MMI_Bool = fea_sel.Chi_Feature_selction (
MeMo_Imp_Norm.loc[:, Bool_Column+"churn"],Ratio,No_of_Variable+1)
for i in range(len(FS_MMI_Norm["Specs"])):
    A=FS_MMI_Norm["Specs"]
    B=FS_MMI_Norm["Score"]
    worksheet1.write(row, col, A.iloc[i], cell_format1)
    worksheet1.write(row, col + 1, B.iloc[i], cell_format1)

col = 4
A=FS_MMI_Bool["Specs"]
B=FS_MMI_Bool["Score"]
if (B.iloc[i]==np.inf):
    B="inf"
else:
    B=B.iloc[i]
worksheet1.write(row, col, A.iloc[i], cell_format1)
worksheet1.write(row, col + 1, B, cell_format1)
col = 0
row += 1

row+=2
worksheet1.write(row, col, "Mean ,Mode impute Normlized Numerical Outl", cell_format0)
col = 4
worksheet1.write(row, col, "Mean ,Mode impute Normlized Bool Outl", cell_format0)
col = 0
row += 1
FS_MMI_Norm_out = fea_sel.Anova_Feature_selction (
MeMo_Imp_Norm_OutLier.loc[:, Num_Column+"churn"],Ratio,No_of_Variable)

```

```

    FS_MMI_Bool_out = fea_sel.Chi_Feature_selction (
MeMo_Imp_Norm_OutLier.loc[ : ,
Bool_Column+["churn"]],Ratio,No_of_Variable+1)

for i in range(len(FS_MMI_Norm_out["Specs"])):
    A=FS_MMI_Norm_out["Specs"]
    B=FS_MMI_Norm_out["Score"]
    worksheet1.write(row, col , A.iloc[i] ,cell_format1)
    worksheet1.write(row, col + 1, B.iloc[i] ,cell_format1)

    col = 4
    A=FS_MMI_Bool_out["Specs"]
    B=FS_MMI_Bool_out["Score"]
    if (B.iloc[i]==np.inf):
        B="inf"
    else:
        B=B.iloc[i]
    worksheet1.write(row, col , A.iloc[i] ,cell_format1)
    worksheet1.write(row, col + 1, B , cell_format1)
    col = 0
    row += 1

def check_availability(element, collection: iter):
    return element

##=====
#####
    #print ("\n\nLst_Del_Norm===== :
",Lst_Del_Norm.shape )
    #
    #List_Del_Nor_FL=
FS_Lst_Del_Norm["Specs"].tolist()+FS_Lst_Del_Bool["Specs"].tolist()
    #if check_availability('churn', List_Del_Nor_FL)=='churn':
    #    List_Del_Nor_FL.remove('churn')
    #
    #Temp_LD_N_DF =Lst_Del_Norm[List_Del_Nor_FL+["churn"]].copy()
    #print("No of Featues ===== : ",len(List_Del_Nor_FL))
    #

    #X = Temp_LD_N_DF.iloc[:, :-1] #independent columns
    #y = Temp_LD_N_DF.iloc[:, -1] #target column i.e price range
    #X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=Ratio, random_state=0)
    #A=Gas_Cla.NB_Classifires(X_train, X_test, y_train, y_test)

##=====
#####

    print ("\n\nLst_Del_Norm_OutLier===== :
",Lst_Del_Norm_OutLier.shape)
    (DF_LD_O_No_Row,DF_LD_O_No_Col )=Lst_Del_Norm_OutLier.shape

    List_Del_Nor_OUT_FL=
FS_Lst_Del_Norm_out["Specs"].tolist()+FS_Lst_Del_Bool_out["Specs"].toli
st()
    # if check_availability('churn', List_Del_Nor_OUT_FL)=='churn':
    #     List_Del_Nor_OUT_FL.remove('churn')

```



```

Temp_LD_N_O_DF=Lst_Del_Norm_OutLier[List_Del_Nor_OUT_FL+["churn"]].copy
()
    print("No of Featues  ==== : ",len(List_Del_Nor_OUT_FL))

    X = Temp_LD_N_O_DF.iloc[:, :-1]    #independent columns
    y = Temp_LD_N_O_DF.iloc[:, -1]    #target column i.e price range
    X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=Ratio, random_state=0)

    (DF_LD_O_No_Train_Row ,DF_LD_O_No_Train_Col )=X_train.shape
    (DF_LD_O_No_Test_Row ,DF_LD_O_No_Test_Col )=X_test.shape

#    LD_N_O_Acc=Gas_Cla.NB_Classifires(X_train, X_test, y_train,
y_test)

'''=====LD_N_O_Acc
====='''

    clf_GNB = GaussianNB()
    clf_GNB.fit(X_train, y_train)
    y_pred = clf_GNB.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy GNB ==1=====
: ",metrics.accuracy_score(y_test, y_pred))
    NB_Acc =metrics.accuracy_score(y_test, y_pred)

    clf_DT = tree.DecisionTreeClassifier(criterion="entropy",
max_depth=5)
    clf_DT.fit(X_train, y_train)
    y_pred= clf_DT.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy DT ===== : ",
metrics.accuracy_score(y_test, y_pred))
    DT_Acc =metrics.accuracy_score(y_test, y_pred)

    clf_SVM = svm.SVC(kernel='linear') # Linear Kernel
    clf_SVM.fit(X_train, y_train)
    y_pred = clf_SVM.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy SVM ===== : ",
metrics.accuracy_score(y_test, y_pred))
    SVM_Acc =metrics.accuracy_score(y_test, y_pred)

    clf_NB_DT = VotingClassifier(estimators=[('GNB', clf_GNB), ('DT',
clf_DT)],voting='soft', weights=[2, 1])
    clf_NB_DT.fit(X_train, y_train)
    y_pred= clf_NB_DT.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy Voting (GNB + DT)===== : ",
metrics.accuracy_score(y_test, y_pred))
    NBDT_Acc =metrics.accuracy_score(y_test, y_pred)

```

```

    clf_SVM_DT = VotingClassifier(estimators=[ ('SVM', clf_SVM), ('DT',
clf_DT)], voting='hard')
    clf_SVM_DT.fit(X_train, y_train)
    y_pred= clf_SVM_DT.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy Voting (SVM + DT )===== :",
metrics.accuracy_score(y_test, y_pred))
    DTSVM_Acc      =metrics.accuracy_score(y_test, y_pred)

    clf_NB_SVM = VotingClassifier(estimators=[ ('SVM', clf_SVM),
('GNB', clf_GNB)], voting='hard')
    clf_NB_SVM.fit(X_train, y_train)
    y_pred= clf_NB_SVM.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy Voting (SVM + GNB)===== :",
metrics.accuracy_score(y_test, y_pred))
    SVMNB_Acc      =metrics.accuracy_score(y_test, y_pred)

#     SVM_Acc      = 1.000
#     DTSVM_Acc   = 2.000
#     SVMNB_Acc   = 3.000

LD_N_O_Acc =
[round(NB_Acc,4) ,round(DT_Acc,4) ,round(SVM_Acc,4) ,round(NBDT_Acc,4) ,rou
nd(DTSVM_Acc,4) ,round(SVMNB_Acc,4) ]

'''=====LD_N_O_Acc
====='''

print("LD_N_O_Acc",LD_N_O_Acc)

[NB_LD_Acc,DT_LD_Acc,SVM_LD_Acc,NBDT_LD_Acc,DTSVM_LD_Acc
,SVMNB_LD_Acc]=LD_N_O_Acc

##=====
=====

#print ("\n\nMean Mode impute Norm_===== :
",MeMo_Imp_Norm.shape)
#MMI_Nor_FL=
FS_MMI_Norm["Specs"].tolist()+FS_MMI_Bool["Specs"].tolist()
#if check_availability('churn', MMI_Nor_FL)=='churn':
#     MMI_Nor_FL.remove('churn')
#
#Temp_MMI_N_DF=MeMo_Imp_Norm[MMI_Nor_FL+["churn"]].copy()
#print("No of Featues  ==== : ",len(MMI_Nor_FL))

#X = Temp_MMI_N_DF.iloc[:, :-1] #independent columns
#y = Temp_MMI_N_DF.iloc[:, -1] #target column i.e price range
#X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=Ratio, random_state=0)
#A=Gas_Cla.NB_Classifires(X_train, X_test, y_train, y_test)

```

```

##=====
=====

    print ("\n\nMeMo_Imp_Norm_OutLier===== :
",MeMo_Imp_Norm_OutLier.shape )
    MMI_Nor_OUT_FL=
FS_MMI_Norm_out["Specs"].tolist()+FS_MMI_Bool_out["Specs"].tolist()
#     if check_availability('churn', MMI_Nor_OUT_FL)=='churn':
#         MMI_Nor_OUT_FL.remove('churn')

    (DF_MMI_O_No_Row,DF_MMI_O_No_Col)=MeMo_Imp_Norm_OutLier.shape

Temp_MMI_N_O_DF=MeMo_Imp_Norm_OutLier[MMI_Nor_OUT_FL+["churn"]].copy()
    print("No of Featues  ==== : ",len(MMI_Nor_OUT_FL))

    X = Temp_MMI_N_O_DF.iloc[:, :-1]    #independent columns
    y = Temp_MMI_N_O_DF.iloc[:, -1]    #target column i.e price range

    print("No of Shape  ==== : ",(X.shape))
    print("No of Head  ==== : ",(X.head()))

    X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=Ratio, random_state=0)

    (DF_MMI_O_No_Train_Row, DF_MMI_O_No_Train_Col)=X_train.shape
    (DF_MMI_O_No_Test_Row, DF_MMI_O_No_Test_Col)=X_test.shape

#     MMI_N_O_Acc =Gas_Cla.NB_Classifires(X_train, X_test, y_train,
y_test)

'''=====MMI N O Acc
====='''

    clf_GNB = GaussianNB()
    clf_GNB.fit(X_train, y_train)
    y_pred = clf_GNB.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy GNB ==1=====
:",metrics.accuracy_score(y_test, y_pred))
    NB_Acc = metrics.accuracy_score(y_test, y_pred)
    NB_Acc = (tn+ fp+ tp)/(tn+ fp+ fn+ tp)

    clf_DT = tree.DecisionTreeClassifier(criterion="entropy",
max_depth=5)
    clf_DT.fit(X_train, y_train)
    y_pred= clf_DT.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy DT  ===== :",
metrics.accuracy_score(y_test, y_pred))
    DT_Acc =metrics.accuracy_score(y_test, y_pred)

    clf_SVM = svm.SVC(kernel='linear') # Linear Kernel
    clf_SVM.fit(X_train, y_train)
    y_pred = clf_SVM.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()

```

```

    print ("Accuracy SVM ===== :",
metrics.accuracy_score(y_test, y_pred))
    SVM_Acc =metrics.accuracy_score(y_test, y_pred)

    clf_NB_DT = VotingClassifier(estimators=[('GNB', clf_GNB), ('DT',
clf_DT)],voting='soft', weights=[2, 1])
    clf_NB_DT.fit(X_train, y_train)
    y_pred= clf_NB_DT.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy Voting (GNB + DT)===== :",
metrics.accuracy_score(y_test, y_pred))

    clf_SVM_DT = VotingClassifier(estimators=[ ('SVM', clf_SVM), ('DT',
clf_DT)], voting='hard')
    clf_SVM_DT.fit(X_train, y_train)
    y_pred= clf_SVM_DT.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy Voting (SVM + DT )===== :",
metrics.accuracy_score(y_test, y_pred))
    DTSVM_Acc    =metrics.accuracy_score(y_test, y_pred)

    clf_NB_SVM = VotingClassifier(estimators=[ ('SVM', clf_SVM),
('GNB', clf_GNB)], voting='hard')
    clf_NB_SVM.fit(X_train, y_train)
    y_pred= clf_NB_SVM.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy Voting (SVM + GNB)===== :",
metrics.accuracy_score(y_test, y_pred))
    SVMNB_Acc    =metrics.accuracy_score(y_test, y_pred)

#     SVM_Acc    = 1.000
#     DTSVM_Acc = 2.000
#     SVMNB_Acc = 3.000

MMI_N_O_Acc =
[round(NB_Acc,4),round(DT_Acc,4),round(SVM_Acc,4),round(NBDT_Acc,4),rou
nd(DTSVM_Acc,4),round(SVMNB_Acc,4) ]

'''=====MMI_N_O_Acc
====='''

[NB_MMI_Acc,DT_MMI_Acc,SVM_MMI_Acc,NBDT_MMI_Acc,DTSVM_MMI_Acc
,SVMNB_MMI_Acc] = MMI_N_O_Acc
workbook.close()

if Customer[0]=="VALID" :
#     X1= X_test.iloc[50:51,: ]

    X1_Name = Customer[1:15]
    X1_Val=[]
    for i in range(14) :
        X1_Val =X1_Val +[ Customer[i+15]]
    X1_Val=[X1_Val]

```

```

X1=pd.DataFrame(X1_Val, columns =X1_Name)

# Temp_MMI_N_O_DF.iloc[:, :-1]
print ("\n\nX1 shape      : = " , X1)

[y_pred_GNB] = clf_GNB.predict(X1)
[y_pred_DT] = clf_DT.predict(X1)
[y_pred_SVM] = clf_SVM.predict(X1)
[y_pred_NB_DT] = clf_NB_DT.predict(X1)
[y_pred_DT_SVM] = clf_SVM_DT.predict(X1)
[y_pred_SVM_NB] = clf_NB_SVM.predict(X1)

print ("y_pred_GNB      : = " , y_pred_GNB)
print ("y_pred_DT       : = " , y_pred_DT)
print ("y_pred_SVM       : = " , y_pred_SVM)
print ("y_pred_NB_DT     : = " , y_pred_NB_DT)
print ("y_pred_DT_SVM    : = " , y_pred_DT_SVM)
print ("y_pred_SVM_NB    : = " , y_pred_SVM_NB)

ACC_LD=[NB_MMI_Acc,DT_MMI_Acc,SVM_MMI_Acc,NBDT_MMI_Acc,DTSVM_MMI_Acc
,SVMNB_MMI_Acc]

Classifire_Pred =
[y_pred_GNB,y_pred_DT,y_pred_SVM,y_pred_NB_DT,y_pred_DT_SVM,y_pred_SVM_
NB]+ACC_LD

df_Lst_Del_min= df.copy()

FS_Selceted =FS_Lst_Del_Norm_out["Specs" ].tolist()+
FS_Lst_Del_Bool_out["Specs" ].tolist()
# print ("FS_Selceted",FS_Selceted)
df_Lst_Del_min=df_Lst_Del_min[FS_Selceted]
Lst_Del_min =df_Lst_Del_min.min()
Lst_Del_max =df_Lst_Del_min.max()
print ("\n\nX1 Lst_      : =\n" ,FS_Selceted,
Lst_Del_min.tolist(),Lst_Del_max.tolist())
Load_Parametres =FS_Selceted+
Lst_Del_min.tolist()+Lst_Del_max.tolist()+Classifire_Pred
else:
df_Lst_Del_min= df.copy()

FS_Selceted =FS_Lst_Del_Norm_out["Specs" ].tolist()+
FS_Lst_Del_Bool_out["Specs" ].tolist()
# print ("FS_Selceted",FS_Selceted)
df_Lst_Del_min=df_Lst_Del_min[FS_Selceted]
Lst_Del_min =df_Lst_Del_min.min()
Lst_Del_max =df_Lst_Del_min.max()

print ("\n\nX1 Lst_      : =\n" ,FS_Selceted,
Lst_Del_min.tolist(),Lst_Del_max.tolist())
Load_Parametres =FS_Selceted+
Lst_Del_min.tolist()+Lst_Del_max.tolist()

# return ([DF_Shape,DF_No_Col ,DF_No_Row
,DF_LD_O_No_Row,DF_LD_O_No_Col, DF_LD_O_No_Train_Row

```

```

,DF_LD_O_No_Train_Col,DF_LD_O_No_Test_Row
,DF_LD_O_No_Test_Col,DF_MMI_O_No_Row,DF_MMI_O_No_Col,DF_MMI_O_No_Train_
Row, DF_MMI_O_No_Train_Col,DF_MMI_O_No_Test_Row, DF_MMI_O_No_Test_Col,
NB_LD_Acc,DT_LD_Acc,SVM_LD_Acc,NBDT_LD_Acc ,DTSVM_LD_Acc
,SVMNB_LD_Acc])
    return ([DF_Shape,DF_No_Col ,DF_No_Row
,DF_LD_O_No_Col,DF_LD_O_No_Row, DF_LD_O_No_Train_Col
,DF_LD_O_No_Train_Row,DF_LD_O_No_Test_Col
,DF_LD_O_No_Test_Row,DF_MMI_O_No_Col,DF_MMI_O_No_Row,DF_MMI_O_No_Train_
Col, DF_MMI_O_No_Train_Row,DF_MMI_O_No_Test_Col, DF_MMI_O_No_Test_Row,
NB_LD_Acc,DT_LD_Acc,SVM_LD_Acc,NBDT_LD_Acc ,DTSVM_LD_Acc
,SVMNB_LD_Acc,NB_MMI_Acc,DT_MMI_Acc,SVM_MMI_Acc,NBDT_MMI_Acc,DTSVM_MMI_
Acc
,SVMNB_MMI_Acc],[FS_Lst_Del_Norm_out,FS_Lst_Del_Boolean_out],Load_Parametr
es)

#
#C_DATA = ["VALID"] + ['eqpdays', 'changem', 'mou', 'age1', 'recchrg',
'months', 'age2', 'ownrent', 'incmiss', 'occprof', 'mailres',
'mailord', 'marryun', 'travel']+[812.0, 2192.25, 2336.25, 89.0,
159.92999269999996, 20.0, 49.0, 1.0, 0.0, 0.0, 0.0, 1.0, 1.0, 1.0]
##C_DATA = ["NULL"]
#
#A,B,C,y_pred_GNB =main(1.5,1.5,0.2,7,C_DATA)

Classifiers module
# -*- coding: utf-8 -*-
"""
Created on Sat Jun 19 16:01:25 2021

@author: madhawa
"""

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.api as sm
import os

import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix, plot_confusion_matrix

from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LinearRegression
from sklearn import linear_model

from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
#from sklearn.naive_bayes import MultinomialNB
#from sklearn.naive_bayes import ComplementNB
#from sklearn.naive_bayes import BernoulliNB
from sklearn import tree
from sklearn import svm
from sklearn.ensemble import VotingClassifier
from sklearn.model_selection import cross_val_score

```

```

from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn import metrics

from sklearn.decomposition import PCA

def NB_Classifiers(X_train, X_test, y_train, y_test):

    clf_GNB = GaussianNB()
    clf_GNB.fit(X_train, y_train)
    y_pred = clf_GNB.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy GNB ==1====="
:"metrics.accuracy_score(y_test, y_pred))
    NB_Acc =metrics.accuracy_score(y_test, y_pred)

    clf_DT = tree.DecisionTreeClassifier(criterion="entropy",
max_depth=5)
    clf_DT.fit(X_train, y_train)
    y_pred= clf_DT.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy DT ===== :",
metrics.accuracy_score(y_test, y_pred))
    DT_Acc =metrics.accuracy_score(y_test, y_pred)

    clf_SVM = svm.SVC(kernel='linear') # Linear Kernel
    clf_SVM.fit(X_train, y_train)
    y_pred = clf_SVM.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy SVM ===== :",
metrics.accuracy_score(y_test, y_pred))
    SVM_Acc =metrics.accuracy_score(y_test, y_pred)

    clf_NB_DT = VotingClassifier(estimators=[('GNB', clf_GNB), ('DT',
clf_DT)], voting='soft', weights=[2, 1])
    clf_NB_DT.fit(X_train, y_train)
    y_pred= clf_NB_DT.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy Voting (GNB + DT)===== :",
metrics.accuracy_score(y_test, y_pred))
    NBDT_Acc =metrics.accuracy_score(y_test, y_pred)

    clf_SVM_DT = VotingClassifier(estimators=[ ('SVM', clf_SVM), ('DT',
clf_DT)], voting='hard')
    clf_SVM_DT.fit(X_train, y_train)
    y_pred= clf_SVM_DT.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy Voting (SVM + DT )===== :",
metrics.accuracy_score(y_test, y_pred))
    DTSVM_Acc =metrics.accuracy_score(y_test, y_pred)

```

```

    clf_NB_SVM = VotingClassifier(estimators=[ ('SVM', clf_SVM),
('GNB', clf_GNB)], voting='hard')
    clf_NB_SVM.fit(X_train, y_train)
    y_pred= clf_NB_SVM.predict(X_test)
    M = confusion_matrix(y_test, y_pred)
    tn, fp, fn, tp = M.ravel()
    print ("Accuracy Voting (SVM + GNB)===== :",
metrics.accuracy_score(y_test, y_pred))
    SVMNB_Acc      =metrics.accuracy_score(y_test, y_pred)

#     SVM_Acc      = 1.000
#     DTSVM_Acc   = 2.000
#     SVMNB_Acc   = 3.000

    ACCURACY_01 =
[round(NB_Acc,4),round(DT_Acc,4),round(SVM_Acc,4),round(NBDT_Acc,4),rou
nd(DTSVM_Acc,4),round(SVMNB_Acc,4) ]

    return
([round(NB_Acc,4),round(DT_Acc,4),round(SVM_Acc,4),round(NBDT_Acc,4),ro
und(DTSVM_Acc,4),round(SVMNB_Acc,4) ])
    # plotting the confusion matrix

Feature selection module
# -*- coding: utf-8 -*-
"""
Created on Thu Jun 17 08:07:30 2021

@author: madhawa
"""

#
=====
=====
""" This module is toload the Data Set and

    """
#
=====
=====

import numpy as np
import pandas as pd

from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LinearRegression
from sklearn import linear_model

from sklearn.model_selection import train_test_split
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.feature_selection import f_classif

def Anova_Feature_selction (df,Ratio,No_of_Variable):

```



```

X = df.iloc[:,1:-1] #independent columns
y = df.iloc[:, -1] #target column i.e price range
# print("x----\n\n",X.head())
# print("y----\n\n",y.head())
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=Ratio, random_state=0)

bestfeatures = SelectKBest(score_func=f_classif, k=No_of_Variable)
fit = bestfeatures.fit(X_train,y_train)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X_train.columns)
#concat two dataframes for better visualization
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Specs','Score'] #naming the dataframe
columns
return (featureScores.nlargest(No_of_Variable,'Score')) #print 10
best features

def Chi_Feature_selction (df,Ratio,No_of_Variable):
X = df.iloc[:,1:-1] #independent columns
y = df.iloc[:, -1] #target column i.e price range
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=Ratio, random_state=0)

bestfeatures = SelectKBest(score_func=chi2, k=No_of_Variable)
fit = bestfeatures.fit(X_train,y_train)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X_train.columns)
#concat two dataframes for better visualization
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Specs','Score'] #naming the dataframe
columns
df=featureScores.nlargest(No_of_Variable,'Score')
df=df[df.Specs != "churn"]
print ("*****\n",df)
return (df) #print 10 best features

```