



Retention Prediction of Credit Card Users Using Data Mining and Machine Learning Techniques

**A Dissertation Submitted for the Degree of
Master of Computer Science**

G.A.H. Sandamali

University of Colombo School of Computing

2021



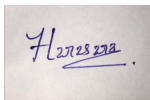
DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: G.A.H Sandamali

Registration Number: 2018/MCS/079

Index Number: 18440792



30/11/2021

Signature of the Student & Date

This is to certify that this thesis is based on the work of Mr. /Ms. G.A.H Sandamali , under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by

Supervisor Name: Dr. M.G.N.A.S Fernando

30-11-2021



Signature of the Supervisor & Date

ACKNOWLEDGEMENT

I would like to acknowledge the limitless support & guidance given by Dr. M.G Noel A.S Fernando who supervised my efforts without any bounds to make this thesis a success.

ABSTRACT

Almost all the business entities thrive their businesses by attracting the customers largely to their businesses day by day. For that, they introduce attractive service campaigns, marketing techniques, and advertising as well. But after all, the business giants found that the churning of the customers from their products or the services can affect largely the profits of their businesses. Hence their existence in the business world depends on the number of customers who retain their services not merely the number of consumers at the beginning.

So, out of all the business domains, this research focuses on the credit card domain which largely affects by the churning of their customers. With the literature, the research identified the number of single machine learning models that were mostly used related to retention prediction. Afterward, the single machine learning models such as Logistic Regression, Random Forest, MLP, k-Nearest Neighbor, Naïve Bayes, Decision Tree, Ada Boost, XGBoost, and LightGBM are used and the performance is evaluated using the metrics such as Accuracy, Area under Curve and Mathews Correlation. Then with the comparison of their performance identification of the weak learners and the meta-model is achieved. Then an ensemble model is created by stacking the weak learners with the meta-model to achieve the increased performance rather than that of the single model.

The built ensemble model guarantees the accuracy of 0.9645, the AUC value of 0.9455 and along with the other performance metrics proving that the ensemble machine learning model is the best solution from the rest of the single models that were being used.

Through this research, it is identified that Total transaction count, the ratio of Total Transaction count over quarter 4 vs quarter 1, Total Revolving Balance, Average Utilization Ratio, and Total Transaction Amount are the features that positively correlated with retention of the credit card users

Keywords: data mining, machine learning, credit card consumer attrition, ensemble model, stacking

TABLE OF CONTENTS

DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT.....	iii
LIST OF FIGURES.....	vii
LIST OF TABLES	ix
LIST OF EQUATIONS	x
ABBREVIATIONS.....	xi
CHAPTER 1: INTRODUCTION.....	1
1.1 Motivation.....	1
1.2 Statement of the problem.....	2
1.3 Research Aims and Objectives	3
1.3.1 Aim.....	3
1.3.2 Objectives.....	3
1.4 Scope.....	4
1.5 Structure of the Thesis	4
CHAPTER 2: LITERATURE REVIEW	5
2.1 A Literature Review.....	5
2.2 Usage of Hybrid or ensemble methods.....	12
2.3 Presentation of Scientific Material	15
2.4 Tools and Technologies	17
CHAPTER 3: METHODOLOGY	19
3.1 Business understanding.....	20
3.2 Data understanding phase	20
3.3 Data Preparation	31
3.3.1 Handling Missing Values	31

3.3.1 Association rule Mining	32
3.3.2 Feature Selection	32
3.3.3 Encoding.....	33
3.3.4 Sampling.....	34
3.3.5 Data Splitting.....	36
3.4 Model Building	37
CHAPTER 4: EVALUATION AND RESULTS	41
4.1 Evaluation of the Results	41
4.1.2 Association rule mining results	41
4.1.3 Feature Importance.....	43
4.1.3 Correlation Matrix with Heat map	44
4.1.4 PCA Analysis	47
4.1.5 Model Evaluation	50
4.2 Strengths of the Research	57
CHAPTER 5: CONCLUSION	58
5.1 Research Overview	58
5.2 Limitations of the Research	60
5.3 Future Work and Recommendations	60
APPENDICES	61
Bibiliography	62

LIST OF FIGURES

Figure 1: Python logo	17
Figure 2: Weka logo	17
Figure 3: Streamlit logo	18
Figure 4: CRISP_DM Process	19
Figure 5: Distribution of churners and non-churners of the Credit card churn dataset	21
Figure 6: Customer Age Distribution	22
Figure 7: Gender Distribution Chart	22
Figure 8: Dependent count Distribution	23
Figure 9: Education Level Distribution	23
Figure 10: Marital Status Distribution	24
Figure 11: Income Category Distribution	24
Figure 12: Card Category Distribution	25
Figure 13: Months_on_book Distribution	25
Figure 14: Total_Relationship_Count Distribution	25
Figure 15: Months_Inactive_12_mon Distribution	26
Figure 16: Contacts_Count_12_mon Distribution	26
Figure 17: Credit Limit Distribution	27
Figure 18: Total_Revolving_Bal Distribution	27
Figure 19: Avg_Open_To_Buy Distribution	28
Figure 20: Total_Amt_Chng_Q4_Q1	28
Figure 21: Total_Ct_Chng_Q4_Q1	29
Figure 22: Total_Trans_Amt Distribution	29
Figure 23: Total_Trans_Ct Distribution	30
Figure 24: Avg_Utilization_Ratio Distribution	30
Figure 25: Attrition Distribution	31
Figure 26: Information of the data columns of credit card churn dataset	32
Figure 27: One hot encoded examples	34
Figure 28: Class distribution	35
Figure 29: Class distribution with SMOTE	35
Figure 30: Imbalanced dataset vs. SMOTE and Undersampled dataset	36
Figure 31: Architectural Diagram for single Machine Learning Model	38

Figure 32: Architecture diagram for Stacking model.....	39
Figure 33: Credit Card Churner Prediction Web Application using Ensembler model.	40
Figure 34: Association rules by Apriori Algorithm with Class Variable as Existing Customers .	42
Figure 35: Association rules by Apriori Algorithm with Class Variable as Attrite Customers	42
Figure 36: Feature Importance Graphical Representation.....	43
Figure 37: Feature Importance with the values of importance	44
Figure 38: Correlation Matrix with Heat map.....	45
Figure 39 Visualization for important features: Churned vs Non Churned Customers	46
Figure 40 Scree Plot	48
Figure 41 Eigen Analysis on Credit card churns data	48
Figure 42: Cumulative Values for each Principal Component.....	48
Figure 43 Correlation matrix plot for loadings.....	49
Figure 44 PC1 against PC2.....	50

LIST OF TABLES

Table 1 Principle component and the relevant attributes for each component.....	49
Table 2 Categorical Attributes and the Categorical value of the data set.....	33
Table 3: Machine learning techniques used in previous researches for churn prediction related to credit card domain	9
Table 4: Decision Tree Classifier	51
Table 5: K-Nearest Neighbor Classifier	51
Table 6: Naive Bayes Classifier	52
Table 7: Multi-layer Perceptron classifier	52
Table 8: Logistic Regression	53
Table 9: Random Forest Classifier	53
Table 10 Ada Boost Classifier.....	54
Table 11: XGB Classifier	54
Table 12: LightGBM Classifier.....	55
Table 13 Summary of the Classifiers for Credit Card churn Dataset.....	55
Table 14: Stack Ensembler	56

LIST OF EQUATIONS

Equation 1 : Monthly Churn.....**Error! Bookmark not defined.**

Equation 2: Accuracy Calculation 15

Equation 3: Precision Calculation 16

Equation 4: Recall Calculation 16

Equation 5: F1 Score Calculation..... 16

Equation 6: MCC calculation formula..... 16

ABBREVIATIONS

Ada-boost - Adaptive Boosting

AI - Artificial Intelligence

ANN – Artificial Neural Network

CRISP-DM - Cross Industry Standard Process for Data Mining

CRM - Customer Relationship Management

KNN - K nearest Neighbor

LGBM - Light Gradient Boosted Machine

LR – Logistic Regression

MCC - Matthews's correlation coefficient

ML - Machine Learning

MLP - Multi-layer Perceptron

SMOTE - Synthetic Minority Oversampling Technique

SOM - Self-organizing map

CHAPTER 1: INTRODUCTION

Customer churn or Attrition is the term that refers to when a user or a customer terminates the relationship with the company. The full cost of churn comprises both lost revenue and the cost involved with the marketing because the company has to incur another cost to replace the lost ones with the new ones. Hence the mitigation of the churning of such customers is a key business goal irrespective of its domain.

The capability of predicting whether, a specific customer is at a high possibility of churning, while there is still a pace to do take some actions for it, signifies a huge additional potential revenue source for every business. Apart from the direct loss of revenue that results from a customer leaving the business, the costs of initially acquiring that customer may not have already been covered by the customer's spending to date. Furthermore, it is always more challenging and expensive to acquire a new customer than it is to retain a current user.

Hence it is essential to acquire the knowledge related to attrition especially by the marketers and retention experts. Thus they can introduce relevant marketing campaigns focusing on the potential non churners to be drawn to their business.

1.1 Motivation

If there is no systematic study is acquired about the churning patterns by the marketers then no prior actions will be imposed for the potential churners in advance. If the marketing strategies followed by the company focuses more on churners due to lack of the knowledge then it can vastly effect on the revenues for no good reason

With the technological advancement massive amount of data can be collected, transferred and stored. With the usage of those large data stores, they can be analyzed and can aid the managers for decision making. With that knowledge, a churn prediction can be done which aids in pinpointing the right customers to market. Further the data scientists can analyze the data stores to reveal the reasons behind the abandoning their products or services

Therefore by considering above facts there is a burning need of investigating the credit card churning faced by credit card providing companies. This research work is a result of the motivation driven by above causes

1.2 Statement of the problem

Any business platform wishes to stay in the pinnacle of financial success and to enhance its competitive advantage over their competitors via retaining their customer base to the fullest level. So, the companies take massive efforts to draw new consumers along with the strategies to retain the already acquired customer base. But with the customer dissatisfaction, they discontinue doing the business with so called companies. When more consumers leave or terminate the usage of the product or the services, the less the business thrives. If any business wants to retain their customers, then they need to address customer churn. Customer attrition has a substantial influence on any business as it drops revenues and profits.

Similarly the rate of churn, also known as the rate of attrition, which is the rate at consumers, ceases its relationship with the business.

Customer attrition can be measured in below aspects:

- Total number or percentage of customers lost during a given period of time
- Recurring business value lost.
- Percentage of recurring value lost.

The customer attrition incurs losses as there's always a cost associated with the acquisition of the customers leading to the by finding new ones hence this need to be addressed immediately to retain the success of any business.

Credit card company domain is one such area which adversely affected with this customer retention. There are numerous categories that can be related with the churn. For an example non-contractual churn where the consumers avoid the purchase without completing the transaction and also the involuntary churn, where a customer cannot pay their credit card bill and no longer stays with the credit card company. Rather than maintaining and upgrading existing customer relationships undoubtedly drawing new customers is considerably very expensive.

Since it is a crucial factor to pay more attention towards of customer churn, more time and effort have to spend by the stakeholders to investigate the reasons behind it. The ideal way to prevent

the attrition is to study the customers and the patterns of the churning through historical data and also via the new customer data.

1.3 Research Aims and Objectives

1.3.1 Aim

Due to the vicious competition in banking industry, to attract customers, bankers tend to offer registration for credit cards with much less consideration given to background checks. Even though the issuers must consider a list of regulations, proper verification of the policy holder's data is not done. Achievement of targets has become more important in the industry more than identifying the right customer eligible for a credit card. With non-banking finance companies also joining in recently, credit risk to the industry has increase significantly.

Most often, customers are tempted to obtain these facilities since it has become much easier over the years. In past, it was considered as a luxury and privileged where only premium customers would enjoy these facilities but nowadays almost each person has access to such facilities due to the poor customer-oriented culture.

Hence, banks lose a lot of money and customers both due to insufficient knowledge on churning of the consumers. Identifying the risk attached with a specific customer profile in a proper systematic way such as data mining can be used to replace standard and obsolete risk rating techniques used in the industry.

The goal of this analysis is to address the problem of customer attrition in the credit card domain by attaining the following objectives:

1.3.2 Objectives

- Identify factors that contribute most to the customer's decision of a credit card service termination.
- Using these factors, perform data mining techniques to understand customer retention patterns by classifying policy holders who are likely to continue or terminate their credit card.
- Building a model to predict the credit card holders who are about to churn

- Predict whether the specific customer will churn or will continue the service with the model prediction and then the customers who are at high risk of attrition to be targeted for promotions to reduce the rate of attrition

1.4 Scope

Although various types of statistical and marketing related research are done related to banking domain, the main purpose of this research is to combine the data mining techniques and machine learning for pattern recognition and cluster identification related to the churn prediction of credit card holders.

Proper research related to the banking domain is necessary to identify the potential factors which contribute towards the research

Hence the first task is to do data collection; data related with the credit cards are from Kaggle website. Now, this dataset consists of around 10,000 customers mentioning their age, salary, marital status, education received, number of dependents, credit card limit, credit card category, credit card usage details for last 12 months etc. Credit card usage details will be focused mostly in the research. All together around 20 features will be considered as mentioned above.

1.5 Structure of the Thesis

The thesis covers the research work along with five chapters which outline the research work. The First chapter explains the introduction and the background of Credit Card Consumer Churning along with the aims, research goals and the scope of the research. Chapter 2 discusses the Literature Survey which comprises the previous research works related to the credit card domain and the churn prediction along with the technical aspects. The Design and Methodology section is detailed in chapter 3. Chapter 4 expresses the evaluation of the results obtained by performing the research methodology. Finally, the chapter 5 presents the conclusion and the impact of the research along with its future works and also the limitations of the research as well.

CHAPTER 2

LITERATURE REVIEW

2.1 A Literature Review

Aforementioned in chapter 1, this research is correlated with many fields. Some of these areas are computer science, data mining, machine learning, the banking and credit card domain and the concepts like churning. For better approach of the research, it is a must, to do a literature survey about recent researches which have been carried out under these correlated fields of study which would be a great advantage when it comes to the implementation phase.

Customer churn

A lot of companies take a key focus on Customer Relationship Management (CRM) with regard to reducing operational costs and maintaining a competitive edge. Anticipating the future behavior of customers is a key factor in CRM. It is highly vital to anticipate future decisions customers are likely to take so that companies can take proactive decisions early. (Glady et al., 2009)

The term “customer churn” according to the Information and Communication Technology (ICT) industry is referred to the customers who are about to leave the existing service provider for a new competitor or terminate their subscription with the current company or the service provider.

Therefore it is very important from the perspective of the real life market to sustain the massive competition in the business world according to Hudaib et al in 2015. And also it is essential to manage the churning as well. Hence churn prediction is very important for the real-life market and to thrive the business competition as well, and it is a necessity to manage (Hudaib et al., 2015)

Jamalian and Foukerd had defined an equation to calculate the loss of consumers or the customers in a given period

$$\text{Monthly churn} = (C_0 + A_1 - C_1) / C_0 \dots\dots\dots (1)$$

Equation 1 : Monthly Churn

In this equation, C_0 means the number of consumers at the beginning of the given period, where C_1 is the number of customers at the end of the period. The number of new consumers in the given period is referred to by A_1 . (Jamalian and Foukerdi, 2018).

With the technological development and the immense competition amongst financial bodies, retention of existing customers is extremely challenging. Retaining customers is highly important in numerous businesses whereas finding new customers is always expensive than retaining the existing customer base. (Kaya, Dong, Suhara, Balsicoy & Bozkaya, 2018).(Clemente et al., 2012)

Detection of possible churners early is a key CRM strategy. Predictive models enable numerous measures that are relevant to each customer with their propensity to churn. The higher the propensity value of a customer, the probability of termination becomes greater. With this information, companies can conduct various marketing activities targeted at retaining customers. (Clemente et al,2014).

Companies have to spend enormously to acquire new customers resulting a key change in the commercialization policies of banks. Hence the focus of the banks has turned towards customer retention since it is less expensive to retain existing customers than to find new ones. Hence, a long term customer will consume more and will be a lot less sensitive to the ongoing competition. Therefore, in order to improve customer retention, it is vital to take proper measures regarding customers who are likely to terminate services. (Cohen et al., 2007)

Churn prediction concepts consist of data mining and predictive analytical models which predict the customers who are most likely to churn. Such models analyze personal, behavioral, and customer transactional data for specific and customer-centered retention marketing activities. (Lejeune, 2001)

Predicting customer churn assists in managing customer relationships (CRM) to minimize the potential churners by introducing retaining policies and offering attractive incentives and other benefit packages. Hence these strategies encourage retaining the possible churners. Hence, any probable revenue loss to the company can be mitigated before it occurs. (Umayaparvathi & Iyakutti, 2012)

Companies, therefore, offer the latest technological advancements and services to improve customer service to ensure customers are retained. Prior to that, it is vital in predicting the customers who are likely to terminate contracts in the future in advance since the loss of these

customers would result in loss of revenue and profits to the company. This process is called Churn Prediction. (Umayaparvathi and Iyakutti, 2016.)

Churning or attrition is a major issue in all domains including banks. Hence the banks always try to keep the track of customer interactions. Therefore they can identify customers who are about to churn. Modeling of customer churn mainly targets the potential churners so that precautionary actions can be taken to avoid the churn. (Oyeniyi & Adeyemo, 2015). (“Customer Churn Analysis In Banking Sector Using Data Mining Techniques | Semantic Scholar,” n.d.)

In a highly competitive business environment, most companies have recognized that their most precious asset is their customer and customer data. Churn predictions are mainly analyzed as a major part of customer relationship management (CRM). Managing churn is an important part of retaining valuable customers. Many business entities including banks and many other service providers are shifting employee focus to better customer service management to ensure policies are formed to address retention of customers. (Hassan and Bin-Nashwan, 2017)

Data mining

Data mining refers to the extraction of meaningful knowledge from an enormous amount of data (Nie et al. 2011). Tsai and Lu (2009) explained that data mining is a technique of discovering interesting patterns within the data and predicting or classifying the behavior founded by the model. But the basic challenge of data mining is how to convert an apparently meaningless flock of data into valuable information and competitive intelligence. (Seng and Chen, 2010)

Many researchers confirm that machine learning technology is effective and efficient in predicting these instances. Data mining is applied via learning from previous data (Umayaparvathi and Iyakutti,2016).

For predicting the churning of the customer accurately, it is crucial to construct effective model which satisfy specific evaluation criteria. Numerous modeling techniques are available for that. These data mining algorithms support in selecting variables and in building models (Hung, Yen, & Wang, 2006).

Regression, Neural Networks, Decision Tree, Markov Model, Cluster Analysis, and optimization are some of these techniques that can be used to perform the modeling (Better, Glover, Kochenberger,& Wang, 2008; Mclain & Aldag, 2009). (Hadden, Tiwari,Roy, & Ruta, 2005)

Hadden et al. (2005), has found out of many algorithms the regression and decision tree are the two most popular algorithms and also they yield better results as well. Even though there are many algorithms, neural networks, support vector machines and logistic regression models are frequently used. Data mining research literature recommends that machine learning techniques such as neural networks should be used for non-parametric datasets as they often beat traditional statistical techniques such as linear and quadratic discriminant analysis approaches (Zoric, 2016).

But there are some inherent challenges related with the churn prediction. One such example in these situations is data are mostly imbalanced. This is due to the fact that the churned portion is much lesser than the non-churners. Moreover the noisy data is another challenge. In any case, churn prediction needs to classify customers according to the probability of churn (Pendharkar, 2009)

As classification techniques some researchers choose individual classifiers. In contrast to the individual classifiers, with the technological advancement especially in the computer hardware, combination of individual classifiers came into action. Random forest (based on bagging) and AdaBoost (based on boosting) are few examples for such combinations. (Clemente et al., 2012)

Logistic Regression is a machine learning algorithm which is used for classification problems and it can cope up with different combination of variables and could be an ideal in predicting the churn with higher accuracy and it is a type of probability statistical classification model (Nie, Rowe, Zhang, Tian & Shi, 2011).

When we discuss about ensemble learning Random forest is foremost for addressing regression and classification problems. Also it uses the technique called bagging to produce the results. It is a suggestion to avoid over fitting which is another advantage of Random Forest (Pretorius, Bierman & Steel, 2016).

According to the (Ahmad et al., 2019), they have found that the hybrid models which combine more than one model outperform the single neural network model from the criteria such as accuracy and the type I and type II errors. As per Ahmad et al hybrid model with ANN + ANN is the best in the performance. However with 2 fuzzy testing sets the Baseline ANN model outperformed the hybrid of SOM + ANN models. Hence the conclusion was the hybrid with two ANN performed better than the baseline ANN model and the hybrid of SOM and ANN.

Therefore they concluded that the stability of ANN+ANN hybrid model is higher than the rest of the models considered

Below table gives a brief of the machine learning techniques that have been identified in the literature as discussed above.

Neural Network

Table 1: Machine learning techniques used in previous researches for churn prediction related to credit card domain

Research Models Used	Conclusions	Further Works
Customer Churn Prediction (Wadikar, 2020.)		
A comparative study on the most popular supervised machine learning methods <ul style="list-style-type: none"> • Random Forest • Support Vector Machine (SVM) • Neural Network • Logistic Regression 	Random Forest has outperformed the other machine learning methods	<ul style="list-style-type: none"> • Explore the other algorithms and analyze with different domains related with churning • Build the time-series model to predict customer churn. • Use unsupervised machine learning technique to examine the data.
Employee attrition prediction using neural network cross validation method (Dutta and Bandyopadhyay, 2020)		
Identify the feasibility of utilizing related parameters and determine the probability of being affected by attrition process. <ul style="list-style-type: none"> • feed-forward neural 	A feed-forward neural network and 10-fold cross validation procedure is provided under a single platform that can determine the attrition process	

<ul style="list-style-type: none"> • Support Vector Machine, • k-Nearest Neighbor, • naïve bayes, • Decision Tree, • Adaboost, • RandomForest classifiers 	beforehand.	
Forecasting Credit Card Attrition using Machine Learning (Rico-Poveda and Galpin, 2020)		
<ul style="list-style-type: none"> • LightGBM • XGBoost • Random Forest • Logistic Regression 	LightGBM has performed better than rest of the selected algorithms	To evaluate the attrition models related to other products related to Banking domain
Employee Turnover Prediction with Machine Learning: A Reliable Approach (Zhao et al., 2019)		
<p>Describing, demonstrating and assessing supervised machine learning methods in terms of employee turnover prediction with below algorithms</p> <ul style="list-style-type: none"> • Naïve Bayes method; • decision tree method • random forest method • gradient boosting trees method; (extreme 	Extreme gradient boosting is favored due to its superior predictive power and speed.	Focus more on feature engineering using techniques such as various data encoding and scaling methods

<p>gradient boosting method; a logistic regression method</p> <ul style="list-style-type: none"> • SVM • linear discriminant analysis; • K-nearest neighbor method • neural networks 		
Customer Churn Prediction:A Survey (Ohny and Mathai, 2017)		
<p>Investigate the customer churn using the algorithms given below</p> <ul style="list-style-type: none"> • statistical based techniques • neural networks, • decision trees, • covering algorithms, regression analysis, • k means 	<p>Each and every churn prediction model has their own pros and cons</p>	<p>To process large inputs with higher dimensions and complex attributes for churn prediction</p>
Assessing classification methods for churn prediction by composite Indicators (Clemente et al., 2012)		
<p>Individual classifiers</p> <ul style="list-style-type: none"> • decision trees • neural networks • logistic regression <p>Combined classifiers</p> <ul style="list-style-type: none"> • AdaBoost • Random forest 	<p>logistic regression and classification trees are simplest to implement with best-cost effectiveness to predict the churn</p>	

Credit card churn forecasting by logistic regression and decision tree (Nie et al., 2011)		
<ul style="list-style-type: none"> • logistic regression • decision tree 	Regression performs a little better than decision tree	
Domain knowledge integration in data mining using decision tables: Case studies in churn prediction (Lima et al., 2009)		
Analysis of coefficient signs in logistic regression with the monotonicity analysis of Decision trees can be used to check if the knowledge contained in data mining models matches with domain knowledge, and ways and means to correct any discrepancies found.	ANN with software development has better performance than Neuro-Solution Infinity software.	Use varying data sets from same business domain to analyze the similarity of the behavior
Predicting credit card customer churn in banks using data mining (Kumar and Ravi, 2008)		
Multilayer Perceptron (MLP), decision trees (J48), Random Forest (RF), Radial Basis Function (RBF), Logistic Regression (LR) and Support Vector Machine (SVM)	Decision tree J48 performs as an early warning expert system.	

2.2 Usage of Hybrid or ensemble methods

There's an old proverb "**Unity is Strength**" which expresses the concept of "ensemble methods" in machine learning. Ensemble methods have recorded in achieving top ranking in machine

learning competitions. Thus there's a hypothesis that combination of multiple models to one model has the ability to yield much more power of predicting.

According to Odusami et al the research that was carried in the telecommunication industry, he developed a model using K Nearest Neighbor, Logistic Regression, Random Forest and Decision Tree. Then the developed combinatory model was named as K_LoRD hybrid model. It is assessed on the performance metrics such as accuracy and the receiver operating Curve values. Then the efficiency of the built model yielded prediction accuracy of 91.85% and Area under curve value of 95.9%. The conclusion from the research work was the hybrid model has superceded the ordinary KNN, Logistic Regression, Random Forest and Decision Tree classifiers.(Odusami et al., 2021)

According to the (Ahmad et al., 2019), the experimental outcomes in terms of prediction accuracy displayed that the hybrid models superceded the single neural network baseline model.

In contrast to the age old machine learning models novel machine learning models are continuously developing and evolving. Mostly the advancement is done by introducing hybrid or ensemble learning models. It is proven that they empower in the means of performance such as accuracy, with the computational power, in terms of functionality and as well as the robustness superceding the single machine learning models. At present there are various models which belong to these hybrid or ensemble machine learning categories. (Choubin et al., 2019)

Moreover (Zhang and Mahadevan, 2019) have concluded that in order to achieve more accurate and reliable machine learning methods ensemble and Hybrid ML are competent when compared to the baseline machine learning models.

Integration of different ML methods or either with other soft computing optimization methodologies results in Hybrid models whereas ensemble models are resulted with bagging or boosting techniques which group more than one classifier.(Singh et al., 2018)

According to Khagi et al, they suggested the development of novel hybrid and ensemble methods determine the future of the Machine Learning.(Khagi et al., 2019).

Instead of single baseline machine learning algorithm, series of ML classification algorithms are grouped in ensemble technique hence the accuracy is substantially enhanced from that of the individual baseline models. Moreover ensemble ML belongs to the supervised algorithms and

they are mostly benefited leading to higher training accuracy reaching a higher testing accuracy as well (Ardabili et al., 2019)

What are ensemble methods?

It is a machine learning model which is a combination of multiple models to achieve better results than the individual models. These multiple models are also known as weak learners. It is proven to reach higher accuracy and robustness when these weak learners are combined properly.

There are three main types to combine the weak learners. They are Bagging, Boosting and Stacking.

Bagging : here we consider about the weak learners which are homogenous. And these weak learners learn independently and parallelly and then followed up with combining considering some specific process of deterministic averaging

Boosting : This also deals with homogenous weak learners, but contrast to bagging the boosting technique enable the weak learners to learn in an adaptive sequential manner.

Stacking : Contrast to the above two techniques stacking combines heterogeneous weak learners and learns in parallel manner.

Moreover stacking enables in combining multiple classifications or regression model. The main concept lies behind stacking is we can outbreak a learning problem with various types of models . Then these different models are capable to learn some part of the problem, without approaching the whole problem space. So, it enables in building multiple different learners and use them to build an intermediate prediction, one prediction for each learned model. Then addition of a new model which learns from the intermediate predictions the same target. This final model is supposed to be stacked on the top of the others, justifying the name stacking . Thus, it improves the overall performance, and often resulted with better results when compared to the individual intermediate model.

Using the stacking model we can learn several different weak learners and then combine them together via a meta model to achieve the target of the predictions which is based on the predictions of the rest of the weak learners. SO first we have to identify what are the weak learners we are going to use and the meta learner based on them to combine into a single model.

Here the meta model 's input is the outputs retrieved from the weak learners and the final prediction which is built on this aggregation is the output from the meta model.

2.3 Presentation of Scientific Material

For the analyzing results, it is necessary to have a clear idea about the evaluation criteria that is going to be applied.

TRUE POSITIVE (TP) - A customer who is actually churned and classified as churned -

TRUE NEGATIVE (TN) - A customer who is actually not churned (negative) and classified as not churned (negative).

FALSE POSITIVE (FP) - A customer who is actually not churned(negative) and classified as churned (positive).

FALSE NEGATIVE (FN) - A customer who is actually churned (positive) and classified as not churned (negative). -

Accuracy

This is a mostly used metric. It is calculated by the number of correctly classified data instances upon the total number of data instances.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Equation 1: Accuracy Calculation

But in cases where the dataset is not balanced i.e. both negative and positive classes with different number of data instances, accuracy is not considered as a good measure

Precision

When classifying the data instances, the Positive predictive value is called as the precision. A classifier with precision of 1 can be considered as a good classifier.

$$Precision = \frac{TP}{TP + FP}$$

Equation 2: Precision Calculation

Recall

Recall is also known as true positive rate or sensitivity. A classifier with recall of 1 can be considered as a good classifier.

$$Recall = \frac{TP}{TP + FN}$$

Equation 3: Recall Calculation

F1-score

It is metric where both precision and recall are taken into consideration

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Equation 4: F1 Score Calculation

Since both metrics of precision and the recall is used, it is considered as a measure which is better than accuracy.

Matthews correlation coefficient (MCC)

MCC metrics comes in handy when the dataset is unbalanced as when the accuracy cannot be guaranteed in such instances as it provides an overoptimistic estimation of the ability of the classifier on the majority class. (Sokolova et al., 2006)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Equation 5: MCC calculation formula

Area under the Curve (AUC)

This metric enables in distinguishing between classes and is used as a summary of the ROC curve. When AUC value is higher it represent that the performance of the model is better at distinguishing between the positive and negative classes.

2.4 Tools and Technologies

Python



Figure 1: Python logo

Python is used as the programming language in the research work and easy implementation of machine learning algorithm is the main reason for choosing it. Availability of numerous libraries and frameworks in python aids in easy and time saving implementation.

Table 1 Python Libraries

Data Analysis and Visualization	NUMPY, SCIPY, PANDAS, SEABORN, MATPLOTLIB, PLOTLY
Machine Learning	SCIKIT-LEARN

Weka



Figure 2: Weka logo

For data mining purposes such as association rule mining weka tool is used. Easiness with direct application of ML algorithms is the main reason for selecting the weka tool.

Streamlit

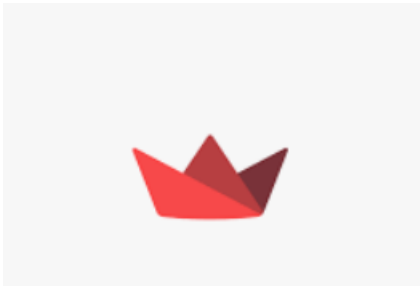


Figure 3: Streamlit logo

Streamlit is an open-source python framework for building web apps for Machine Learning and Data Science. We can instantly develop web apps and deploy them easily using Streamlit. Streamlit allows you to write an app the same way you write a python code. Streamlit makes it seamless to work on the interactive loop of coding and viewing results in the web app.

CHAPTER 3

METHODOLOGY

This chapter aims on the design and the methodology of the research work. The methodology comprises with the solutions to the research objectives that were stated in the introduction chapter.

Cross-industry process for data mining(CRISP-DM) process (Schröer et al., 2021) , is adopted as the basic framework for the methodology. With the literature different models are investigated which are already used and a new approach will be implemented and discussed via forthcoming sections of the methodology

The whole procedure or the workflow of the research is depicted aligning with the CRISP-DM. Here a 6-phase approach is considered.

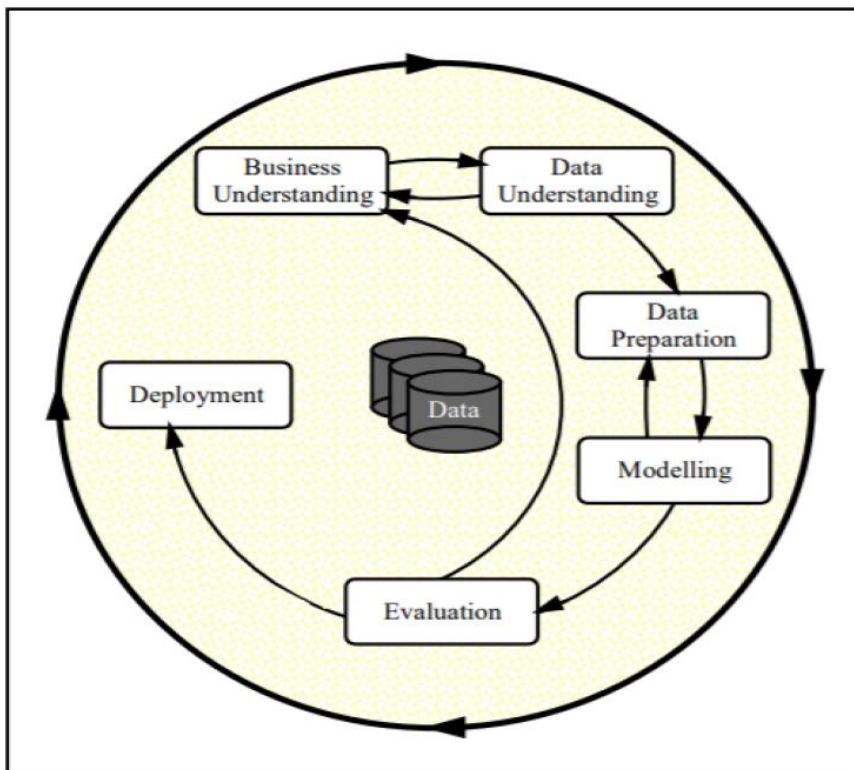


Figure 4: CRISP_DM Process

The main phases are

1. Business understanding
2. Data understanding
3. Data preprocessing
4. Modeling
5. Evaluation

As per the above-mentioned steps, first three steps help in achieving the first three research objectives.

3.1 Business understanding

With the research being predicting the retention or the churn, it can be considered as a great asset in the eye of business strategy as it can be applied, and steps can be taken before the exit of customers. It is vital for the domain experts in decision making and also a proper communication strategy can be implemented to retain the potential churners by understanding their needs. And as a result of that the revenue loss due to churning can be mitigated.

3.2 Data understanding phase

With regards to the methodology, according to the data set, we have to select the correct target data and need to prepare the data for the model. Then the entire data set was split into two data sets. One is taken as the training data set and it covers the 80% of the whole data set, The rest of the data which is 20 % is used as the testing data set. The testing data set is used for calculationg the performance metrics such as accuracy and etc.

- Source of Data : <https://www.kaggle.com/sakshigoyal7/credit-card-customers>
- 10127 Customer Data
- 1627 - Churned Clients
- 8500 - Existing Client

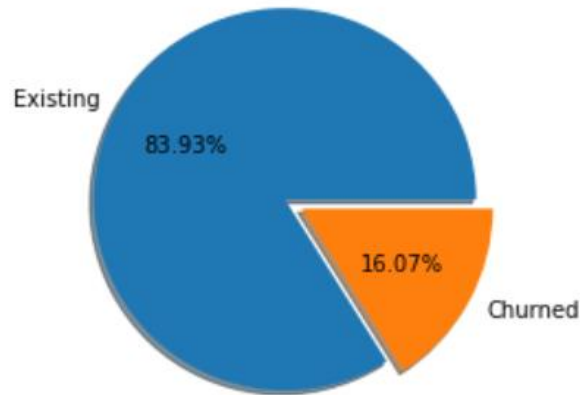


Figure 5: Distribution of churners and non-churners of the Credit card churn dataset

This dataset consists of around 10,000 customers who are the credit card users of a particular financial company. Moreover the dataset covers nearly 18 attributes

Pandas Library of python is used to load the data. Afterwards, the number of records, the data types are investigated using the `info()` function. Then the basic quantitative analysis was done with the data. Central tendency, range, standard deviation, mean, max and min values are calculated using descriptive statistics. Exploratory Data Analysis (EDA) is performed along with the data visualization with the aid of python libraries such as Matplotlib and Seaborn. Histogram and box-plot is created and displayed for each variable to identify the data distribution and to detect the outliers as well.

The spearman correlation is used to identify the correlation between the dependent and independent variables.

1. Clientnum

This attribute indicates the Client number. This is the unique id for each credit card user.

2. Customer_Age

A Demographic variable – Age of the credit card user in years

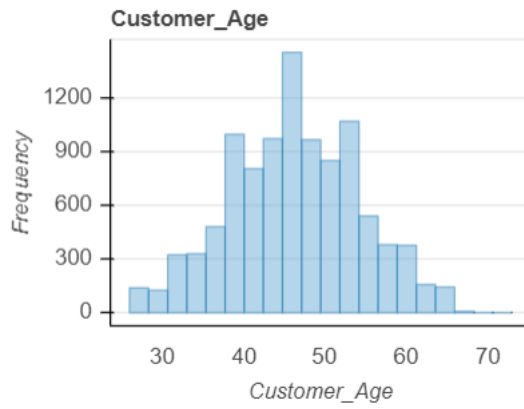


Figure 6: Customer Age Distribution

3. Gender

A Demographic variable - M=Male, F=Female

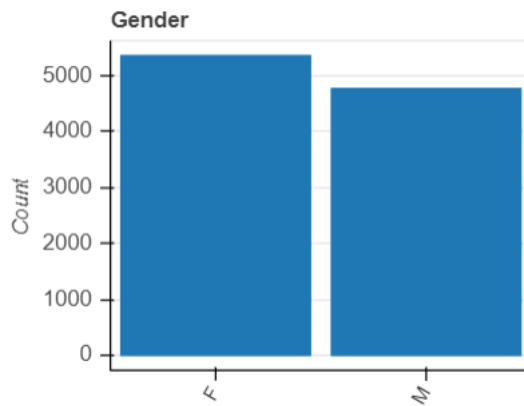


Figure 7: Gender Distribution Chart

4. Dependent_count

Demographic variable - Number of dependents on the account holder

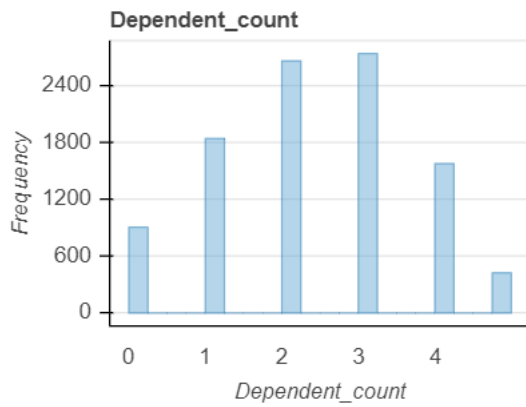


Figure 8: Dependent count Distribution

5. Education_Level

Demographic variable - Educational Qualifications of the credit card policy holder

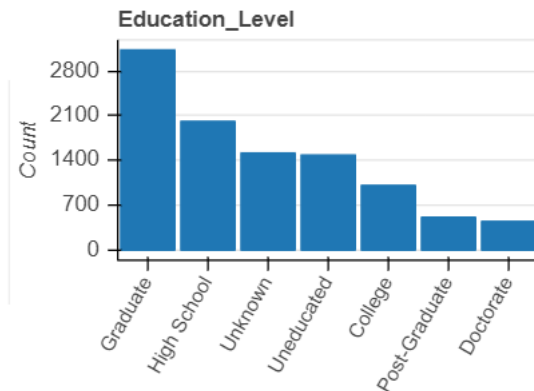


Figure 9: Education Level Distribution

6. Marital_Status

Demographic variable – The marital status as Married, Single, Divorced or Unknown

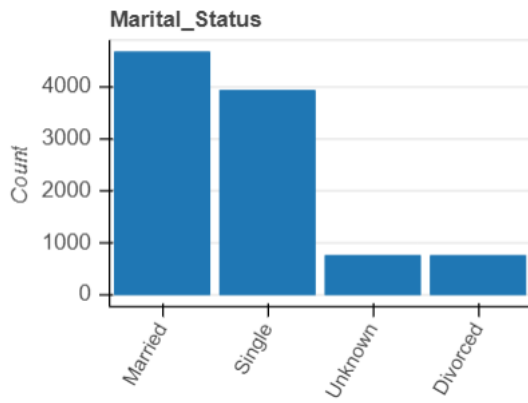


Figure 10: Marital Status Distribution

7. Income_Category

Demographic variable - Annual Income Category of the account holder

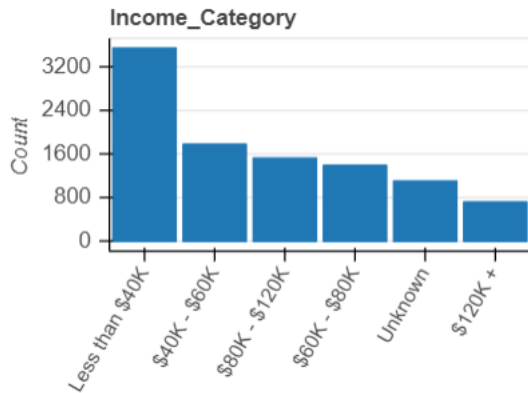


Figure 11: Income Category Distribution

8. Card_Category

Product Variable - Type of Card (Blue, Silver, Gold, Platinum)

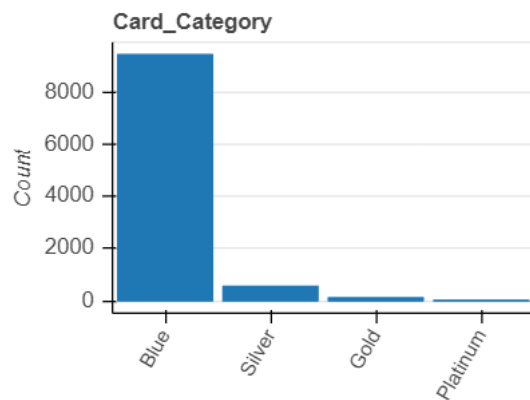


Figure 12: Card Category Distribution

9. Months_on_book

Period of relationship with bank



Figure 13: Months_on_book Distribution

10. Total_Relationship_Count

Total no. of products held by the customer such as cards, accounts and etc.

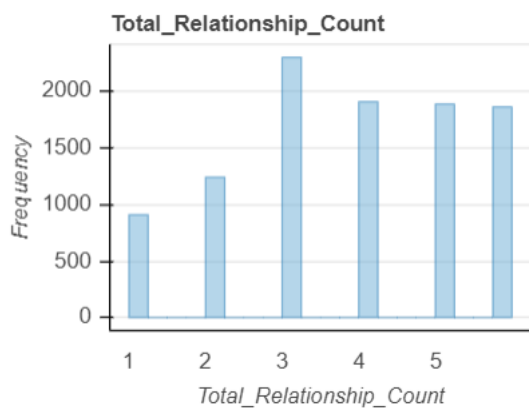


Figure 14: Total_Relationship_Count Distribution

11. Months_Inactive_12_mon

Inactive period denoted using months during the last 12 months

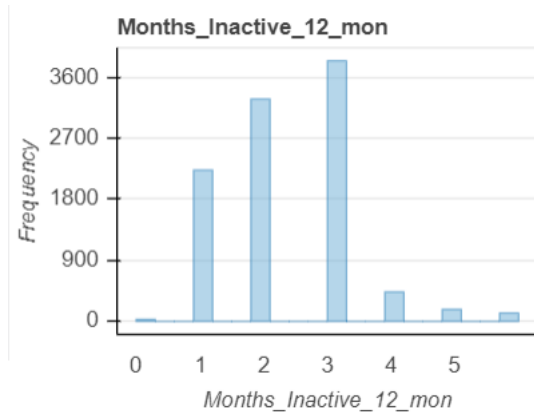


Figure 15: Months_Inactive_12_mon Distribution

12. Contacts_Count_12_mon

No. of Contacts in the last twelve (12) months or the number of times the bank contacted the customer and/or vice versa.



Figure 16: Contacts_Count_12_mon Distribution

13. Credit_Limit

Credit Limit on the Credit Card

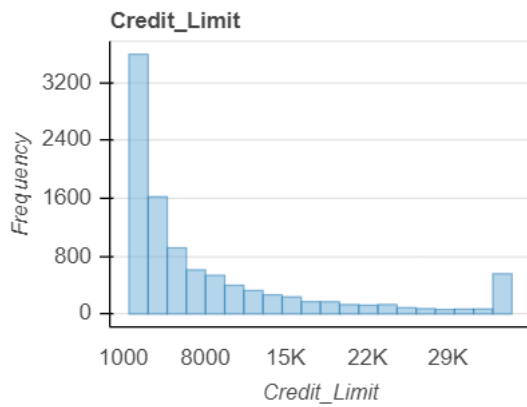


Figure 17: Credit Limit Distribution

14. Total_Revolving_Bal

Total Revolving Balance on the Credit Card and it's the unpaid amount that carries off on the customer's next credit card's cycle.

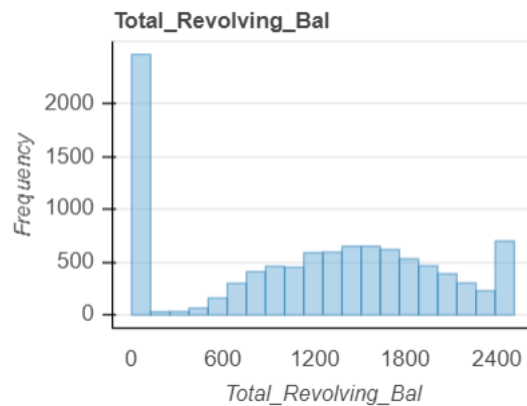


Figure 18: Total_Revolving_Bal Distribution

15. Avg_Open_To_Buy

This is the Open to Buy Credit Line (The difference between the **credit limit** assigned to a cardholder account and the present balance on the account) another way it is the average amount left in the credit card to use. (Average of last 12 months),

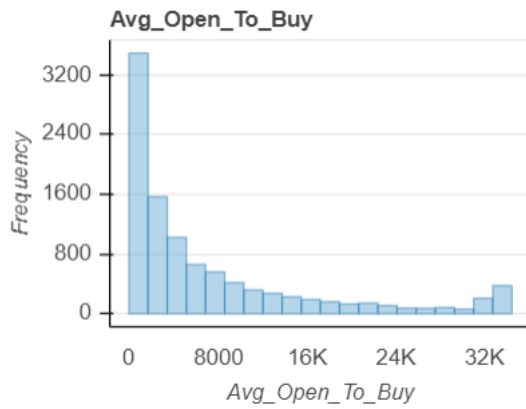


Figure 19: Avg_Open_To_Buy Distribution

16. Total_Amt_Chng_Q4_Q1

Change in Transaction Amount (Q4 over Q1). Represents how much the customer increased their expenditure when comparing the 4th quarter against the 1st.

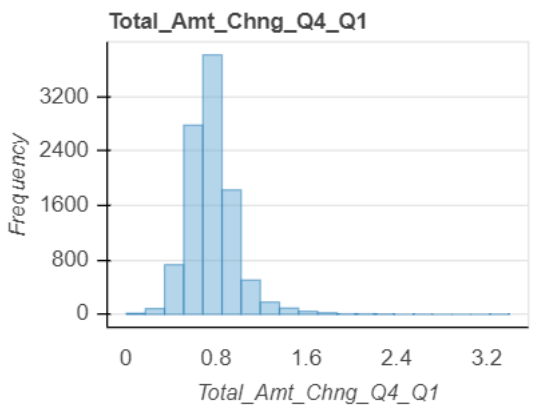


Figure 20: Total_Amt_Chng_Q4_Q1

17. Total_Ct_Chng_Q4_Q1 - Change in Transaction Count.

Represents the number of transactions that the customer has increased in his expenditure comparing with the 4th quarter against the 1st quarter.

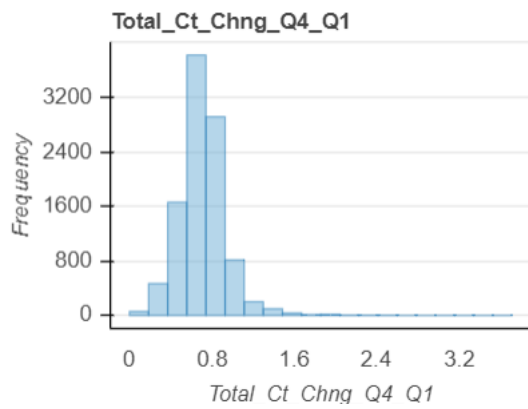


Figure 21: Total_Ct_Chng_Q4_Q1

18. Total_Trans_Amt

Total transaction amount during the last 12 months. Total_Trans_amt is the sum of transactions one has done in the last 12 months. This basically tells us the total usage of a credit card by the user.

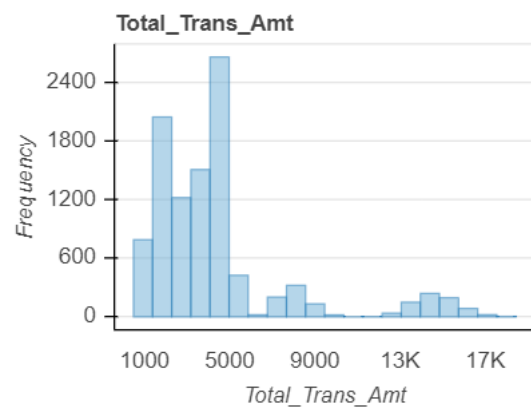


Figure 22: Total_Trans_Amt Distribution

19. Total_Trans_Ct

Total transaction count during the last 12 months.

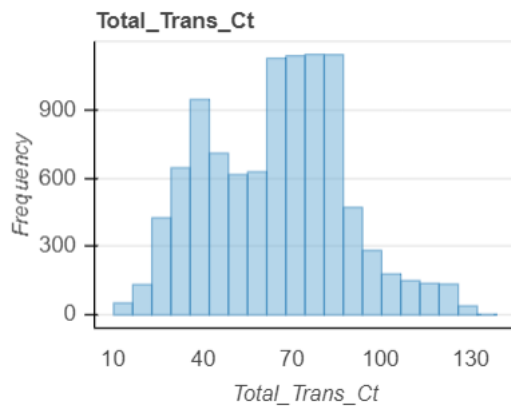


Figure 23: Total_Trans_Ct Distribution

20. Avg_Utilization_Ratio - Average Card Utilization Ratio.

It is the ratio of (credit card spent + money withdrawal)/(Total available limit for credit card spends + Total money withdrawal limit)

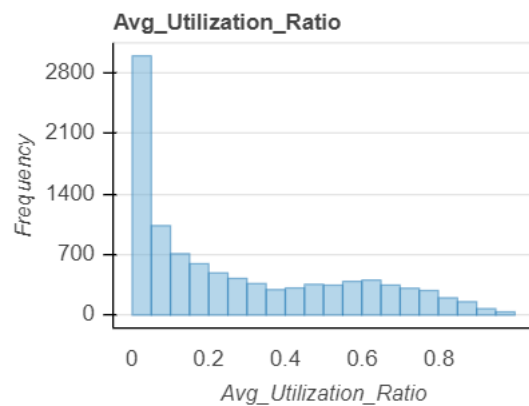


Figure 24: Avg_Utilization_Ratio Distribution

21. Attrition_Flag

This implies whether the customer has left or an existing customer.

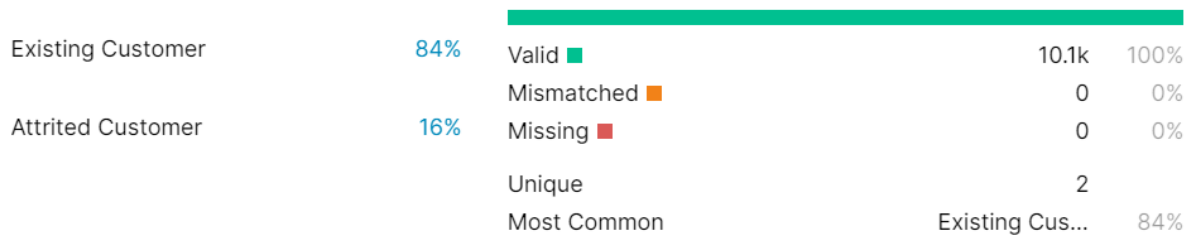


Figure 25: Attrition Distribution

Considering above variables, the Attrition Flag is the dependent variables where the rest of the variables are the independent variables.

3.3 Data Preparation

It is very important phase as it comprises with the activities of conversion of raw data into the final dataset. The resultant data set is then capable to be fed to the modeling algorithms and the output is the anticipated model.

Data cleaning, Removing outliers, Imputing missing values, Creation of new variables, Feature selection and the data transformation are the activities performed during the data preparation phase.

3.3.1 Handling Missing Values

Handling missing values is crucial. If a variable has missing values up to 60% , then it is advised to remove from the final dataset . Many machine learning algorithms do not support the data sets with considerable amount of missing values for a particular attribute. (Kelleher et al., 2015) Luckily the dataset has no missing values reported

```

Data shape: (10127, 21)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 21 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   CLIENTNUM                            10127 non-null  int64
1   Attrition_Flag                       10127 non-null  object
2   Customer_Age                        10127 non-null  int64
3   Gender                              10127 non-null  object
4   Dependent_count                     10127 non-null  int64
5   Education_Level                     10127 non-null  object
6   Marital_Status                      10127 non-null  object
7   Income_Category                     10127 non-null  object
8   Card_Category                       10127 non-null  object
9   Months_on_book                      10127 non-null  int64
10  Total_Relationship_Count             10127 non-null  int64
11  Months_Inactive_12_mon               10127 non-null  int64
12  Contacts_Count_12_mon               10127 non-null  int64
13  Credit_Limit                        10127 non-null  float64
14  Total_Revolving_Bal                 10127 non-null  int64
15  Avg_Open_To_Buy                     10127 non-null  float64
16  Total_Amt_Chng_Q4_Q1                10127 non-null  float64
17  Total_Trans_Amt                     10127 non-null  int64
18  Total_Trans_Ct                      10127 non-null  int64
19  Total_Ct_Chng_Q4_Q1                 10127 non-null  float64
20  Avg_Utilization_Ratio                10127 non-null  float64
dtypes: float64(5), int64(10), object(6)
memory usage: 1.6+ MB
Information of data columns:None

```

Figure 26: Information of the data columns of credit card churn dataset

3.3.1 Association rule Mining

In order to find the correlations and the co-occurrences between the datasets, the association rule mining can be used. Support and the confidence are two metrics that is used to calculate the strength of a particular association rule. Support refers as the frequency of a given rule and confidence refers to the number of times that the particular association rule is truth

Here for this credit card churn dataset, the Aprori Algorithm is used.

3.3.2 Feature Selection

Feature selection is a fundamental concept in machine learning which largely affects the performance of the model created

Model's performance can be adversely impacted if irrelevant or partially relevant features are available in the data set. Two Feature selection techniques are widely used and they are Univariate Selection, Feature Importance and Correlation Matrix with Heat map (Shaikh, 2018). Moreover Principal Component Analysis is another technique which is used mostly among the researchers. It extracts the linear feature for unsupervised feature selection based on eigenvectors analysis to identify critical original features for principal component (Parveen et al., 2012). From the above techniques, I have used PCA, Feature Importance and the Correlation Matrix techniques to distinguish the most important features of the Credit Card Churn Dataset.

3.3.3 Encoding

The data set which is used consists of data with different data types where some of them are numerical, categorical and etc. But when we are using machine learning or deep learning techniques all the categorical data needs to be encoded it to numeric form before we can fit and evaluate a model. Hence one hot encoding will be used instead of label encoding as it neglects the concept that the higher the magnitude higher the importance.

Table 3 Categorical Attributes and the Categorical value of the data set

	Categories
Attrition_Flag	[Existing Customer, Attrited Customer]
Gender	[F, M]
Education_Level	[Graduate, High School, Unknown, Uneducated, College, Post-Graduate, Doctorate]
Marital_Status	[Married, Single, Unknown, Divorced]
Income_Category	[Less than 40K, 40K - 60K, 80K - 120K, 60K - 80K, <i>Unknown</i> , 120K +]
Card_Category	[Blue, Silver, Gold, Platinum]

Some examples are shown in the below table.

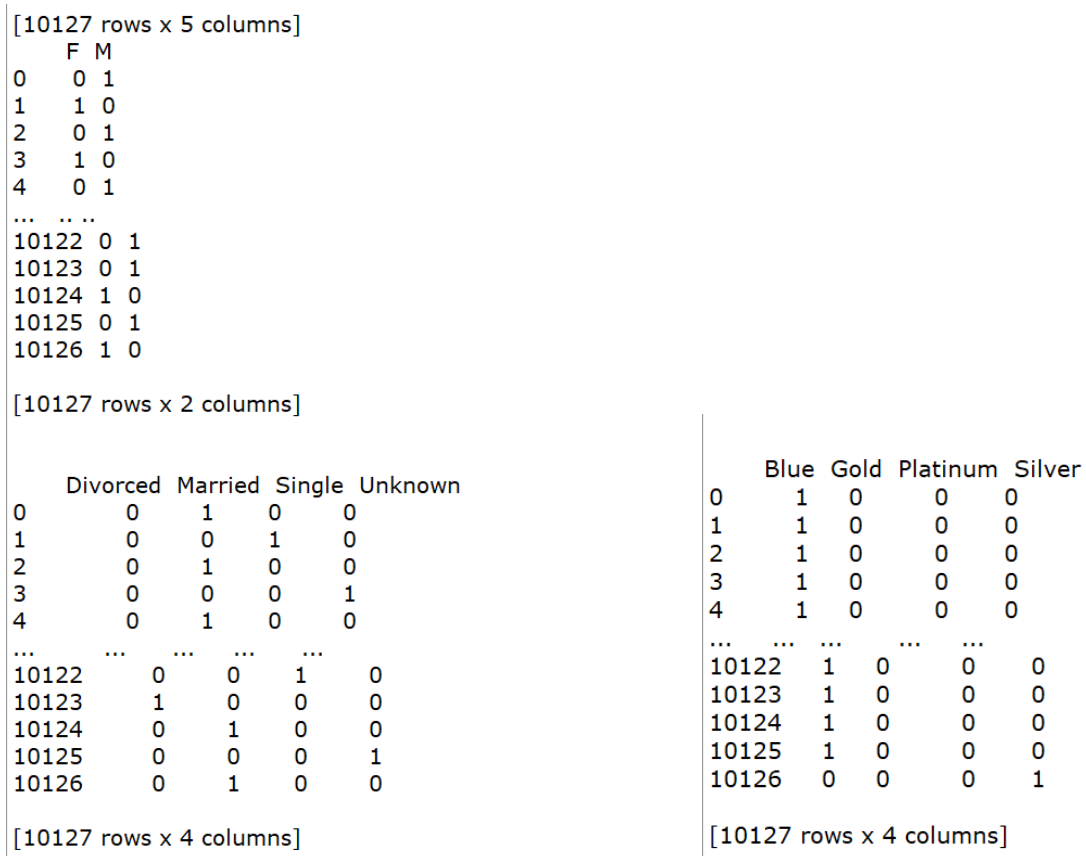


Figure 27: One hot encoded examples

3.3.4 Sampling

Class Imbalance is the most common problem related to the data set when it is related to the churning. This is due to the reason that among two classes one class comprises with more samples than the other. In such situations the focus of the most of the algorithms merely considers the majority class neglecting or misclassifying the minority sample. Here the main fact is there can be crucial knowledge hidden in the minority class hence that shouldn't be ignored.(Longadge and Dongre, 2013)

Related to the data set I have used it also have become a victim of this class imbalance issue. Below diagram shows the distribution of the Class variable related to the Credit Card Churn data set.

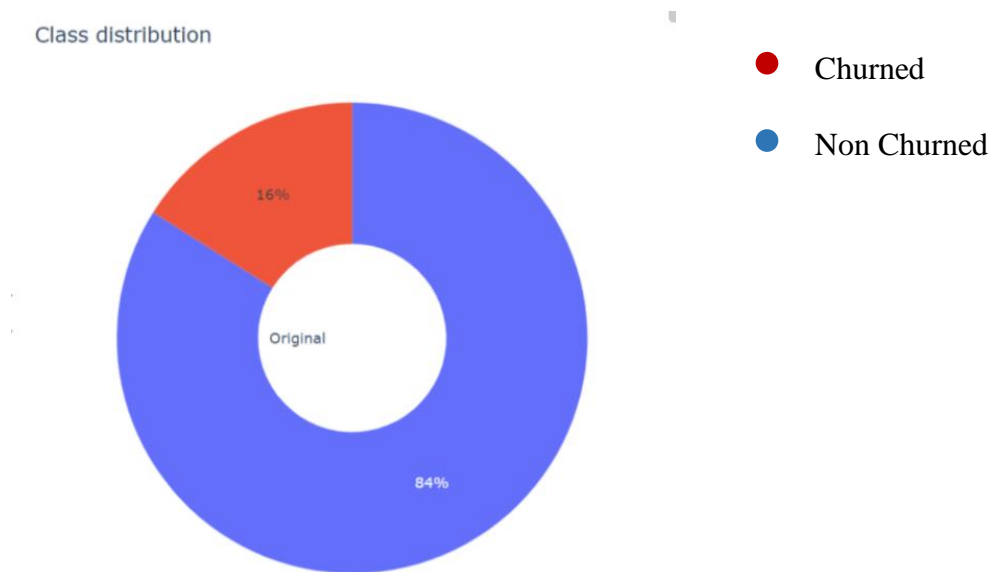


Figure 28: Class distribution

So in order to address the class imbalance oversampling the minority class can be implemented. Here it simply duplicates the samples that hold the class variable of minority class. But the disadvantage is they do not add any new data but merely a duplication of existing minority data. As a solution to above cause generation of new data samples from the existing data can be applied. This can be implied as a data augmentation technique of the minority class. This technique is called as the Synthetic Minority Oversampling Technique, or SMOTE for short. (“SMOTE for Imbalanced Classification with Python,” n.d.)

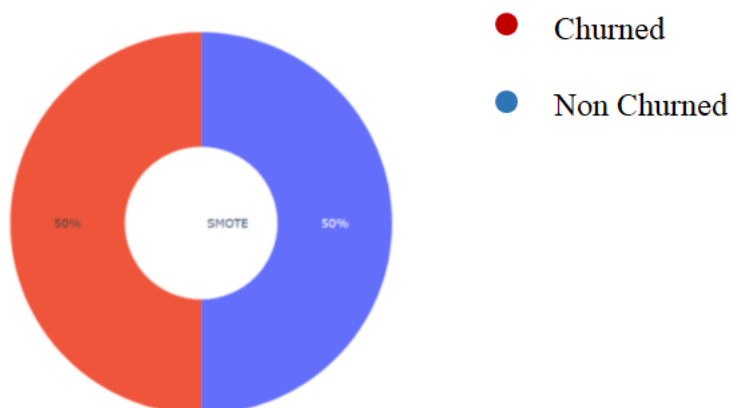


Figure 29: Class distribution with SMOTE

I



Figure 30: Imbalanced dataset vs. SMOTE and Undersampled dataset

According to (Chawla et al., 2002) they found rather than plain under sampling technique, the combination of SMOTE with under sampling has yield better results. Therefore the credit card users dataset is balanced with the SMOTE and under-sampling and it is imported from `imblearn.over_sampling`.

3.3.5 Data Splitting

Final data set was created after going through the above processes and then the obtained data set was splitted into train and test dataset. Based on the most of the existing researchers the dataset was splitted 80% of training data and the 20% of testing data. For better result comparison the research was conducted for both imbalance and balanced datasets as well.

3.4 Model Building

After investigating the literature, it was identified the different researches have used different machine learning algorithms to build models for churn prediction for the credit card domain. Hence a novice model from all those techniques was a major focus in the research.

Even though much of the researches are carried out with the customer attrition or the employee attrition predictions, the research work on financial data such as credit card domain is found to be less due to the scarcity of data with the privacy concerns. The accruing of such datasets due to company policies seems to be the reason but it is actually a vast area that should be given highest priority for the credit card facilitating financial bodies. Given below is a list of research works carried recently

Via literature the researchers have concluded that some machine learning algorithms have superceded the other models and different machine learning algorithms were concluded based on the different data sets they have used in the Credit card domain.

Machine Learning Models used in previous work according to the Literature related to the credit card domain in churn prediction are as follows.

- Logistic Regression,
- Random Forest,
- MLP
- k-Nearest Neighbor
- Naïve bayes
- Decision Tree
- Ada Boost
- XGBoost
- LightGBM

With the above findings the credit card data is subjected to above machine learning model inorder to identify the performance of each machne learning model for the credit card churners dataset. Below architectural diagram is used to achieve the results for each identified machine learning model from the research

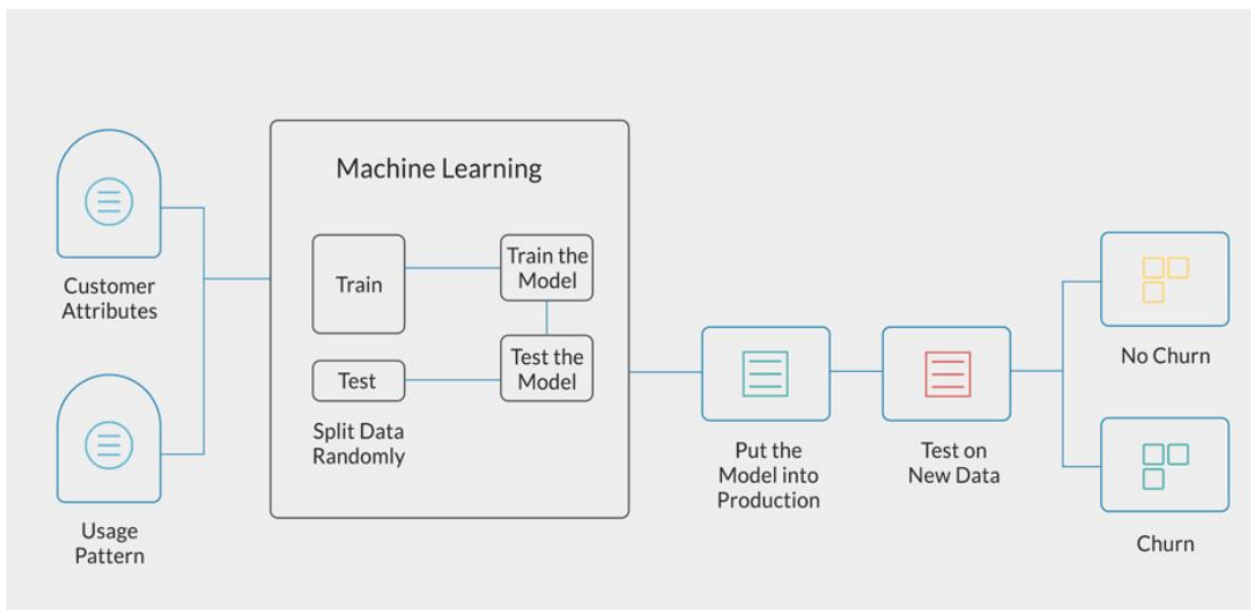


Figure 31: Architectural Diagram for single Machine Learning Model

Here the machine learning model is applied twice separately on both imbalanced data and the SMOTE and under sampled data for a better comparison. The evaluation criteria such as Accuracy, F1 Score, Area Under Curve, Precision, Recall and Mathews Correlation Coefficient are used. Main objective of applying identified machine learning technique is to get an idea about the performance of these techniques and to come up with the technique which outperforms the rest of the machine learning techniques.

Considering the findings of previous research works, it persuaded me to go for an ensemble machine learning algorithm which uses a machine learning algorithm to learn with the means of finding the best way to combine the predictions from two or more base machine learning algorithms. Hence by using stacking technique which is an approach of ensemble modeling the research work anticipated in achieving the capabilities of a range of well-performing models and make predictions that have better performance than any single model in the ensemble

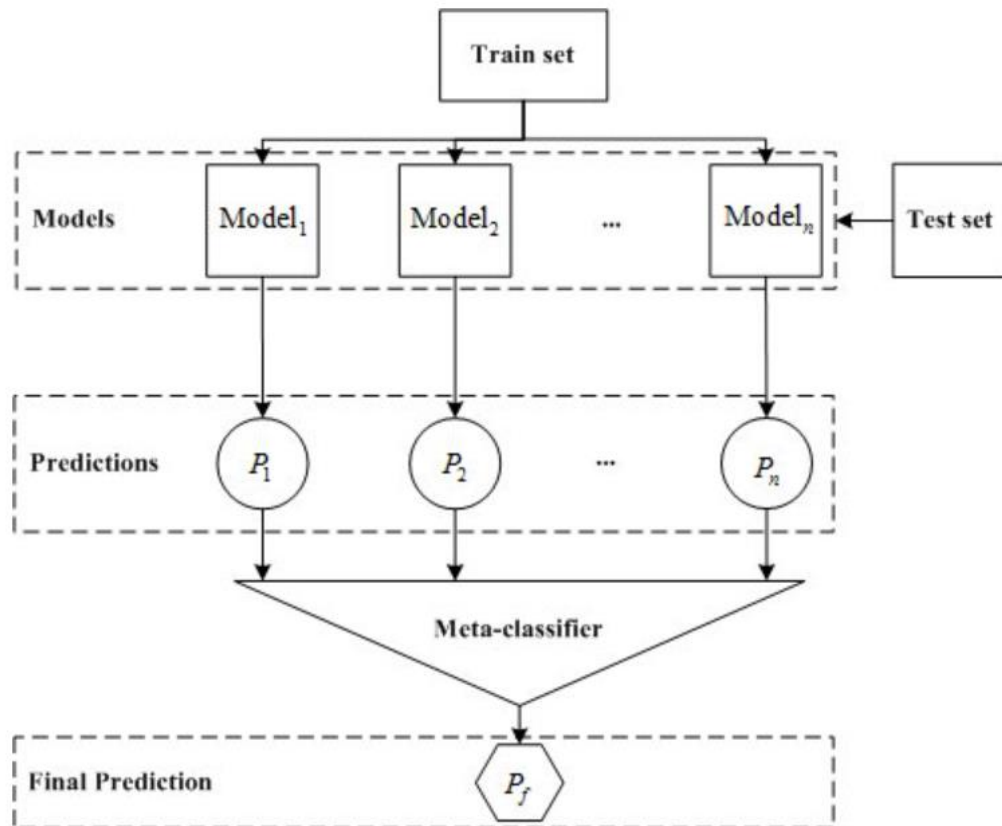


Figure 32: Architecture diagram for Stacking model

So with the results I achieved Logistic Regression, Random Forest, MLP, k-Nearest Neighbor, Naïve bayes, Decision Tree, Ada Boost, XGBoost were considered as weak learners and the one which outperformed above models which is LightGBM is taken as meta model in constructing the Stacking model.

All the Python scripts run using the Jupyter notebook and the web app is developed by Streamlit tool

Considering the higher performance of the ensemble technique this model is embedded in a web app to predict whether the potential customer is a churner or non-churner as below. The Streamlit open source tool has been used when developing this user friendly web app with python.

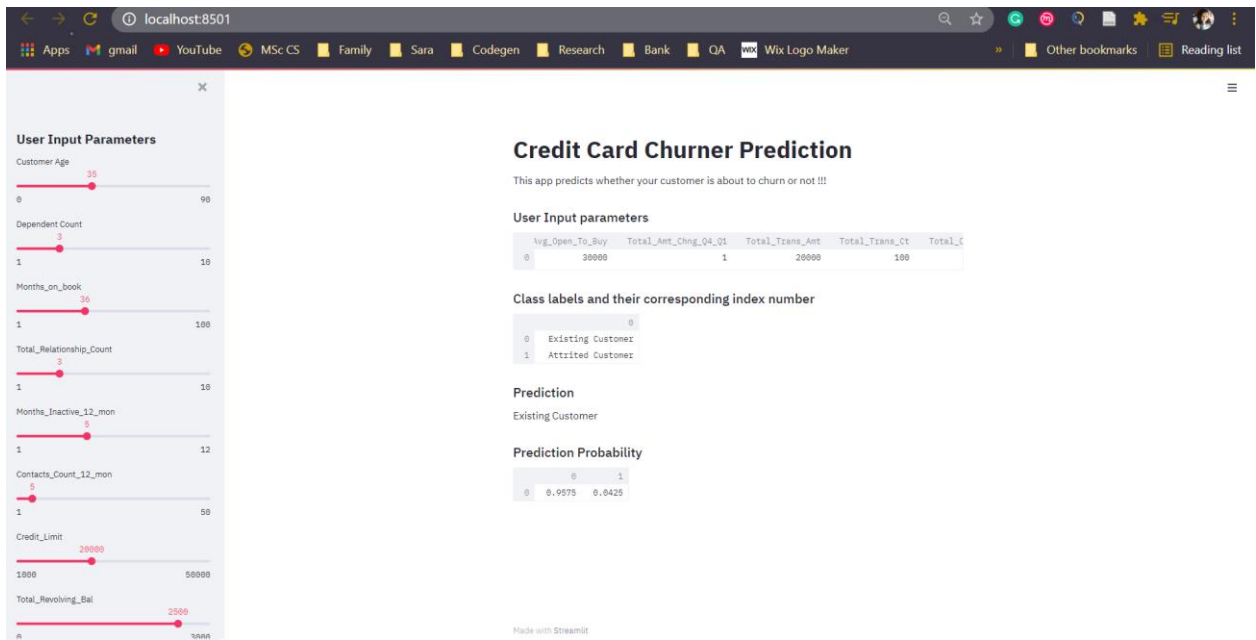


Figure 33: Credit Card Churner Prediction Web Application using Ensembler model.

The domain experts can use this app when predicting the churning of their customers just by entering already available data through the sliders and then the results are retrieved in the interface with its probability metrics as well.

CHAPTER 4

EVALUATION AND RESULTS

In Evaluation and Results section analysis of the results retrieved from the research work are discussed. It evaluates the predictive power of the built ensemble model over the individual machine learning models that were taken into the consideration in chapter 3. Accuracy is the main criterion that is used to compare the models. Moreover the comparison is also considered with the imbalanced data and sample data using SMOTE and under sampling technique.

4.1 Evaluation of the Results

4.1.2 Association rule mining results

Hence the rules that are generated do not mean they are based on individual credit card user's behavior but focuses in finding relationship among the factors considering the behaviors of all the credit card users.

An association rule has two metrics named antecedent and a consequent, both of which are the features or the attributes considered in the credit card data set.

The support value aids in recognizing the association rules that is necessary to focus for further analysis. For example, 0.5 of minimum support is considered as we want to consider only the attribute sets which occurs at least 5063 instances of all the instances. But if there's a low support, we cannot arrive into conclusions as it indicates that there's less information on the relationship between the particular items.

	Attribute 1	Attribute 1 Value	Attribute 2	Attribute 2 Value	Attribute 3	Attribute 3 Value	Attribute 4	Attribute 4 Value	Attribute 5	Attribute 5 Value	Class Variable	Confir: Level
1	Total_Trans_Amt	Less than \$5K	Total_Trans_Ct	75 - 100							Existing	1
Almost all the credit card holders whose total transaction amount is less than \$5K and the total transaction count is between 75-100 are existing.												
2	Total_Trans_Amt	Less than \$5K	Total_Trans_Ct	75 - 100	Total_Ct_Chng_Q4_Q1	Less than 1					Existing	1
Almost all the credit card holders whose total transaction amount is less than \$5K and the total transaction count is between 75-100 an the ratio of Total transaction count change of Q4 and Q1 is less than 1 are existing users.												
3	Total_Amt_Chng_Q4_Q1	Less than 1	Total_Trans_Amt	Less than \$5K	Total_Trans_Ct	75 - 100					Existing	1
Almost all the credit card users whose ratio of Total transaction count change of Q4 and Q1 is less than 1 and total transaction amount is less than \$5K and the total transaction count is between 75 - 100 are existing users.												
4	Total_Amt_Chng_Q4_Q1	Less than 1	Total_Trans_Amt	Less than \$5K	Total_Trans_Ct	75 - 100	Total_Ct_Chng_Q4_Q1	Less than 1			Existing	1
Almost all the credit card users whose ratio of Total transaction count change of Q4 and Q1 is less than 1 and total transaction amount is less than \$5K and the total transaction count is between 75 - 100 are existing users.												
5	Avg_Open_To_Buy	Less than \$5K	Total_Trans_Amt	Less than \$5K	Total_Trans_Ct	75 - 100					Existing	1
Almost all the credit card users whose Average open to buy amount is less than \$5K and whose Total transaction amount is less than \$5K and whose total transaction count is 75 -100 are existing users												
6	Avg_Open_To_Buy	Less than \$5K	Total_Amt_Chng_Q4_Q1	Less than 1	Total_Trans_Amt	75 - 100					Existing	1
Almost all the credit card users whose Average open to buy amount is less than \$5K and whose ratio of total amount change over Q4 and Q1 is less than 1 and whose total transaction amount between 75 -100 are existing users												
7	Avg_Open_To_Buy	Less than \$5K	Total_Trans_Amt	Less than \$5K	Total_Trans_Ct	75 - 100	Total_Ct_Chng_Q4_Q1	Less than 1			Existing	1
Almost all the credit card users whose Average open to buy amount is less than \$5K and whose Total transaction amount is less than \$5K and whose total transaction count is 75 -100 and and whose ratio of total amount change over Q4 and Q1 is less than 1 are existing users												
8	Credit limit category	Less than \$5K	Total_Trans_Amt	Less than \$5K	Total_Trans_Ct	75 - 100					Existing	1
Almost all the credit card users whose Credit limit category is less than \$5K and whose total transaction count is 75 -100 are existing users												
9	Credit limit category	Less than \$5K	Avg_Open_To_Buy	Less than \$5K	Total_Trans_Amt	Less than \$5K	Total_Trans_Ct category	75 - 100			Existing	1
Almost all the credit card users whose Credit limit category is less than \$5K and whose Average open to buy amount is less than \$5K and whose total transaction amount is less than \$5K and whose total transaction count is 75 -100 are existing users												
10	Gender	F	Total_Trans_Amt	Less than \$5K	Total_Trans_Ct	75 - 100					Existing	1
Almost all the credit card users who are female and whose total transaction amount is less than \$5K and whose total transaction count is between 75 to 100 are existing users												

Figure 34: Association rules by Apriori Algorithm with Class Variable as Existing Customers

	Attribute 1	Attribute 1 Value	Attribute 2	Attribute 2 Value	Attribute 3	Attribute 3 Value	Attribute 4	Attribute 4 Value	Attribute 5	Attribute 5 Value	Class Variable	Confir: Level
1	Total_Trans_Ct	25 - 50	Total_Ct_Chng_Q4_Q1	Less than 1							Attrited	0.42
42% of the credit card users whose total transaction count is between 25-50 and whose ratio of total transaction count change of Q4 to Q1 is less than 1 are churned users												
2	Total_Trans_Amt	Less than \$5K	Total_Trans_Ct	25 - 50	Total_Ct_Chng_Q4_Q1	Less than 1					Attrited	0.42
42% of the credit card users whose total transaction amount is less than \$5K and whose total transaction count is between 25-50 and whose ratio of total transaction count change of Q4 to Q1 is less than 1 are churned users												
3	Total_Trans_Ct	25 - 50									Attrited	0.37
37% of the credit card users whose total transaction count is between 25-50 are churned users												
4	Total_Trans_Amt	Less than \$5K	Total_Trans_Ct	25 - 50							Attrited	0.37
37% of the credit card users whose total transaction amount is less than \$5K and whose total transaction count is between 25-50 are churned users												
5	Total_Ct_Chng_Q4_Q1	Less than 1	Avg_Utilization Ratio	Less than 0.25							Attrited	0.22
22% of the credit card users whose ratio of total transaction count change of Q4 to Q1 is less than 1 and whose avg utilization ratio is less than 0.25 are churned users												

Figure 35: Association rules by Apriori Algorithm with Class Variable as Attrite Customers

4.1.3 Feature Importance

Simply the feature importance value represents the attributes which are more prominent considering the entire data set. If the magnitude is higher, then such attributes are more important where the lesser values are less important towards predicting the class variable (Shaikh, 2018). I have used Decision Tree Classifier for extracting the top 10 features for the dataset. (Geurts et al., 2006)

Out[48]: <AxesSubplot:>

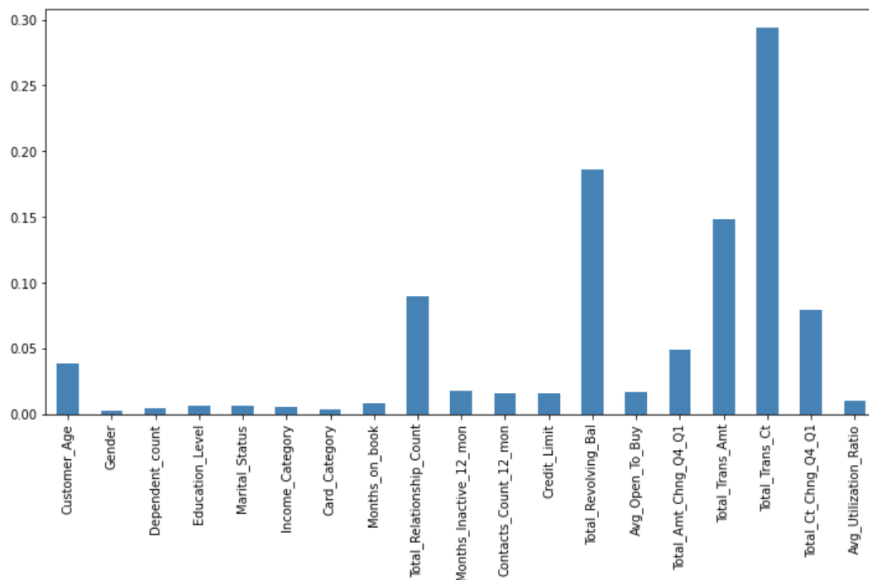


Figure 36: Feature Importance Graphical Representation

Out[50]:

	Feature	Feature importance
0	Total_Trans_Ct	0.293657
1	Total_Revolving_Bal	0.186161
2	Total_Trans_Amt	0.148754
3	Total_Relationship_Count	0.089648
4	Total_Ct_Chng_Q4_Q1	0.078924
5	Total_Amt_Chng_Q4_Q1	0.049055
6	Customer_Age	0.038234
7	Months_Inactive_12_mon	0.017428
8	Avg_Open_To_Buy	0.017312
9	Credit_Limit	0.016306
10	Contacts_Count_12_mon	0.015715
11	Avg_Utilization_Ratio	0.010171
12	Months_on_book	0.008140
13	Education_Level	0.006915
14	Marital_Status	0.006601
15	Income_Category	0.005811
16	Dependent_count	0.004934
17	Card_Category	0.003915
18	Gender	0.002319

Figure 37: Feature Importance with the values of importance

Top 10 features are only considered in model building

4.1.3 Correlation Matrix with Heat map

In order to identify the best predicting attributes related to the customer churn Correlation technique can be used. The correlation between the independent and dependent variables are shown as a correlation matrix

.

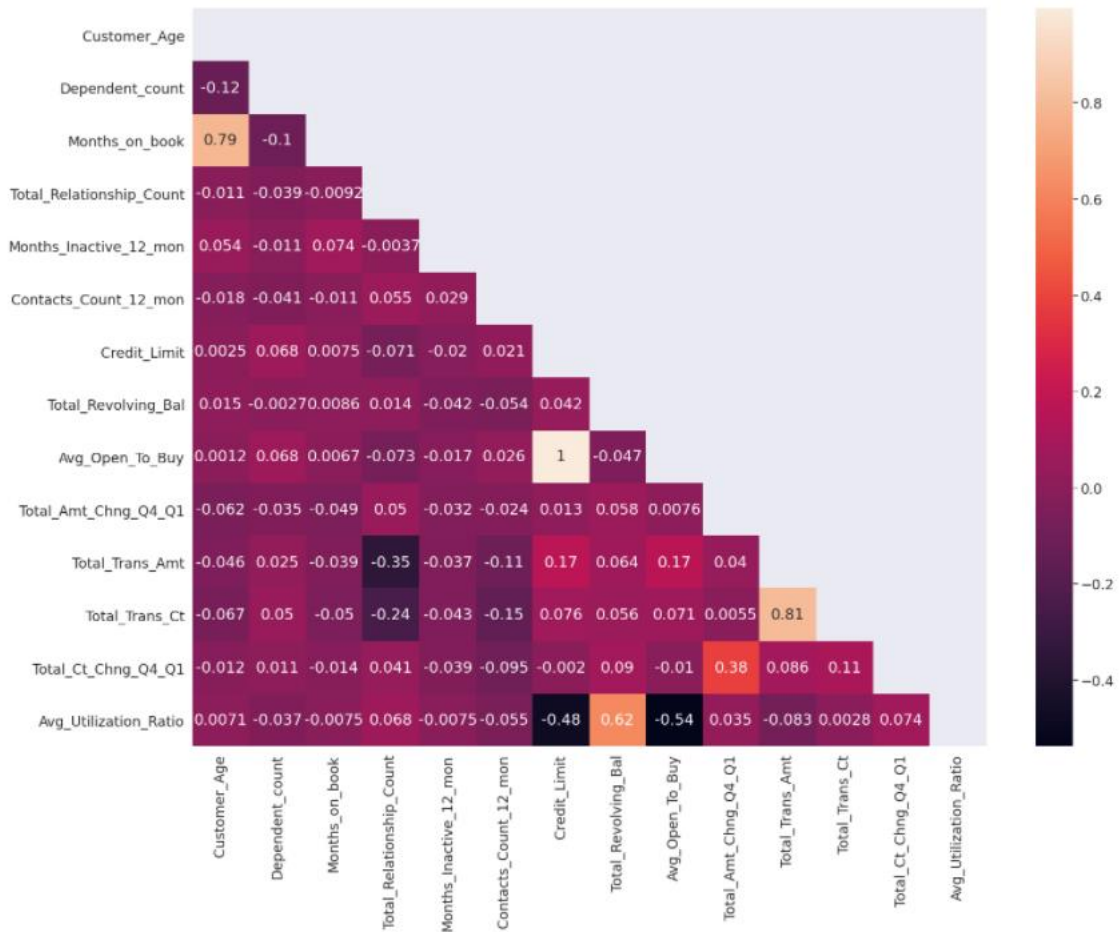


Figure 38: Correlation Matrix with Heat map

The value in each square represents the correlation between the particular variables and the value ranges from -1 to +1. If there's no strong relationship then the values go closer to the 0. If the values are closer to 1 it represents that there's a strong relationship between the given variables.

From the above Heat map below conclusions can be extracted

Positive Relationship

Avg_open _to_Buy and Credit Limit →	1
Total_Trans_Amount and Total_Trans_Ct →	0.81
Customer_Age and Months_on_Book →	0.79
Total_Revolving_Balance and Avg_Utilization_Ratio →	0.62

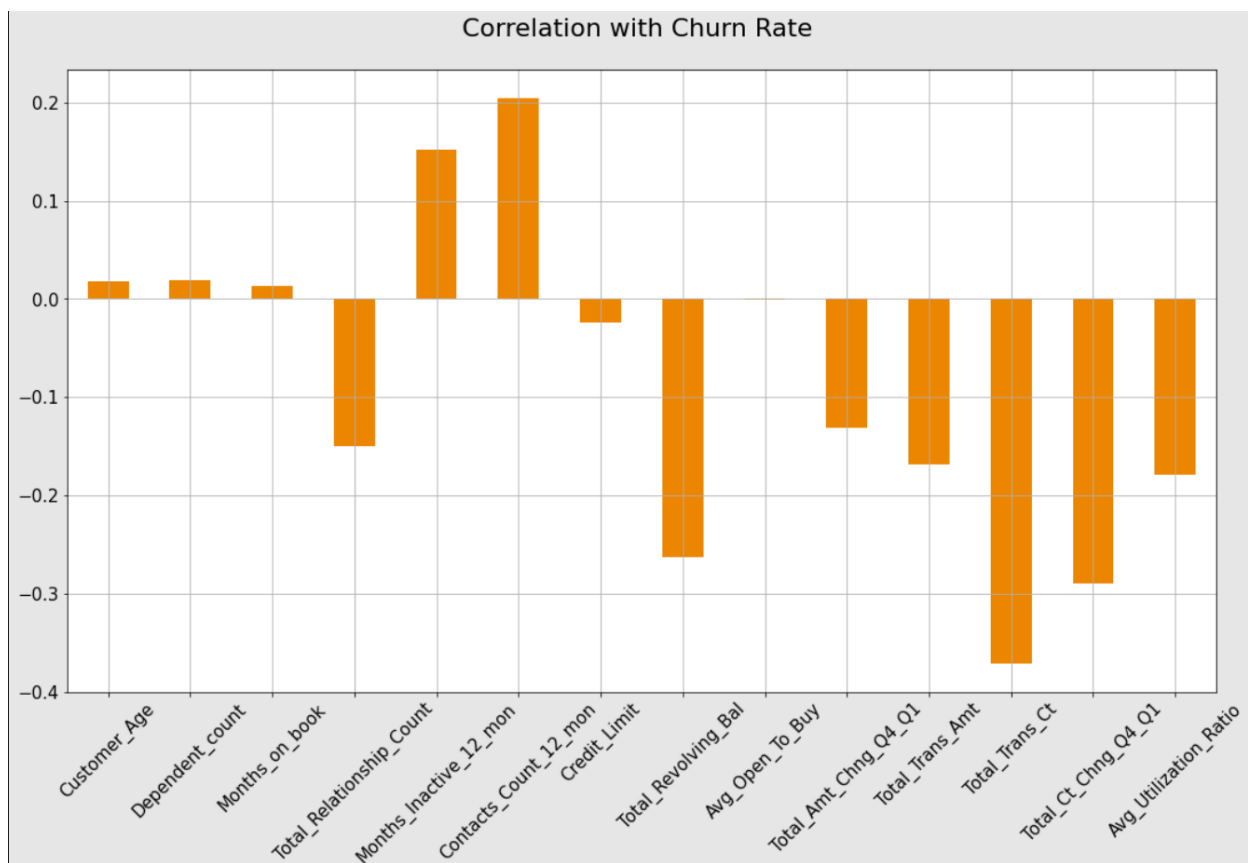


Figure 39 Visualization for important features: Churned vs Non Churned Customers

From the above diagram it shows that Attrition rises when the credit card user stays inactive for more months. Furthermore when the credit card user has many products such as savings accounts, fixed deposits, debit cards and credit cards with the bank, such users tend to churn more with the increase of the product count.

Meanwhile when the Total transaction counts increases there's a less chance for a credit card user to churn. When Total Revolving Balance increases then there's a less chance to churn. When the total transaction count difference of Q4 and Q1 (Total ct change Q4 Q1) increases there's a less chance to churn are the conclusions arrived from the above diagram.

After the correlation analysis it is identified that when the below attributes magnitude increases there's a less chance to churn

- i. Total_Trans_Ct
- ii. Total_Ct_Chng_Q4_Q1

- iii. Total_Revolving_Bal
- iv. Avg_Utilization_Ratio
- v. Total_Trans_Amount
- vi. Total_Relationship_Count
- vii. Total_Ct_Chng_Q4_Q1
- viii. Credit Limit

While the churning is increased with the increase of below attributes

- i. Contatcts_Count_12_Months
- ii. Months_Inactive_12_Months
- iii. Customer Age
- iv. Dependent Count

Hence from above data we can classify the policy holders who are likely to continue or terminate their credit card. Then these customers who are at high risk of churning can then be aimed for promotions to reduce the rate of attrition

4.1.4 PCA Analysis

Principle Component Analysis is performed to recognize the prominent attributes and as a means of dimensionality reduction for the large data set.(Parveen et al., 2012)

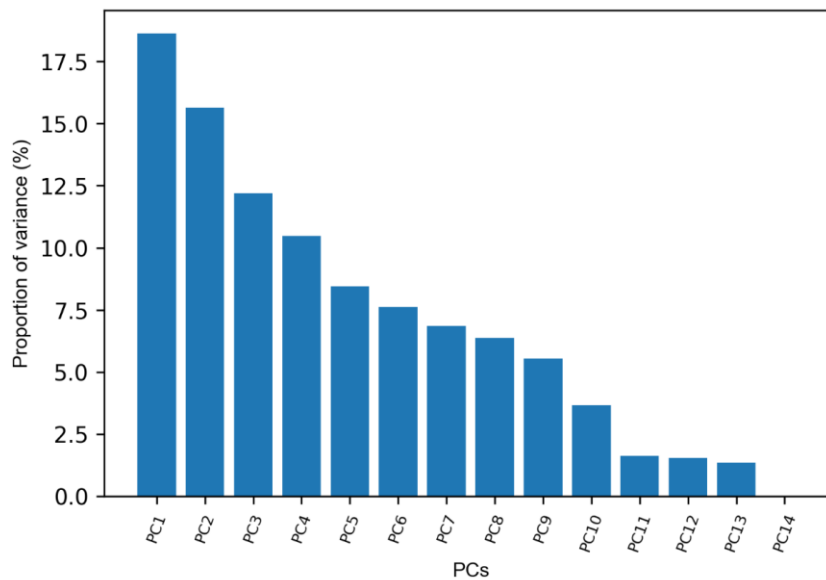


Figure 40 Scree Plot

With the given data set the above scree plot aids in identifying the number of components that explain the most of the variation in the data.

[2 rows x 14 columns]

Proportion of Variance (from PC1 to PC14)

```
[1.86319886e-01 1.56479936e-01 1.21975166e-01 1.04877879e-01
8.45250447e-02 7.61042019e-02 6.85976892e-02 6.37110156e-02
5.54528957e-02 3.65994570e-02 1.62304017e-02 1.55143478e-02
1.36120806e-02 2.75843105e-33]
```

Cumulative proportion of variance

```
[0.18631989 0.34279982 0.46477499 0.56965287 0.65417791 0.73028211
0.7988798 0.86259082 0.91804371 0.95464317 0.97087357 0.98638792
1. 1. ]
```

Figure 41 Eigen Analysis on Credit card churns data

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Proportion	0.186	0.157	0.122	0.105	0.085	0.076	0.069	0.064	0.056	0.037	0.016	0.016	0.014	0.028
Cumulative	0.186	0.343	0.465	0.57	0.654	0.73	0.799	0.863	0.918	0.955	0.971	0.987	1	1

Figure 42: Cumulative Values for each Principal Component

We have to decide the acceptable level of variance; here I have considered a variance of 80% which extends up to PC8

From the above table it depicts that the first feature explains roughly 18.6% of the variance within the given data set while first two explain 34.3% and so on.

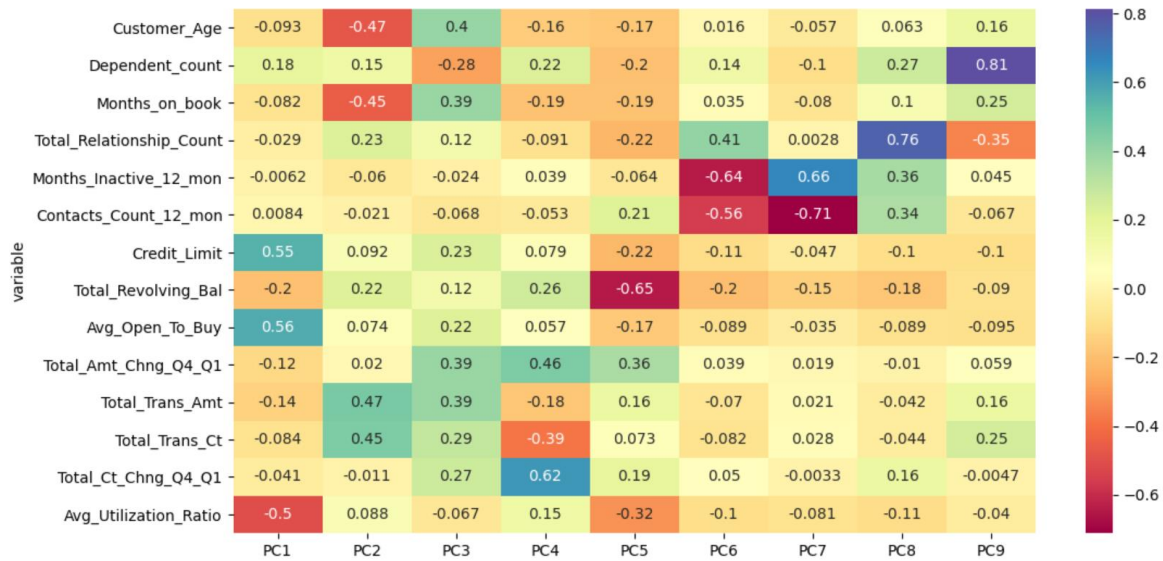


Figure 43 Correlation matrix plot for loadings

Table 2 Principle component and the relevant attributes for each component

Attribute	PC1	PC2	PC3	PC4	PC5
Avg_Open_To_Buy	0.56				
Credit_Limit	0.55				
Total_Trans_Amt		0.47			
Total_Trans_Ct		0.45			
Total_Relationship_Count		0.23			
Total_Revolving_Balance		0.22			
Customer_Age			0.4		
Months on Book			0.39		
Total_Amount_Chng_Q4_Q1			0.39		

Total_Ct_Chang_Q4_Q1			0.27		
Avg Utilization Ratio				0.15	
Dependent Count				0.22	
Contacts_Count_12_Months					0.21

The first principal component increases with increasing Avg_Open_To_Buy and the Credit Limit. This suggests that these two criteria vary together. If one increases, then the remaining ones tend to increase as well.

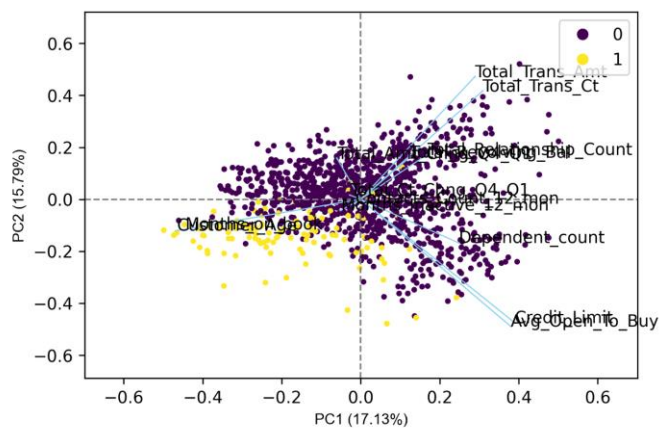


Figure 44 PC1 against PC2

4.1.5 Model Evaluation

After the data mining approach then as per the methodology the performance of single machine learning models are calculated for the credit card churn data set.

Below tables display the results achieved in each single model

4.1.5.1 Results for the Decision Tree Classifier for both Imbalanced and SMOTE and Undersampled data

Table 3: Decision Tree Classifier

Decision Tree for Imbalanced Data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.96	0.96	0.96	1699	0.9309	0.8711	0.74411
Churned	0.79	0.78	0.79	327			
Decision Tree for SMOTE and Undersampled data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.97	0.92	0.95	1699	0.9117	0.7126	0.8905
Churned	0.68	0.86	0.76	327			

4.1.5.2 Results for the K-Nearest Neighbor Classifier for both Imbalanced and SMOTE and Undersampled data

Table 4: K-Nearest Neighbor Classifier

KNN for Imbalanced Data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.92	0.95	0.94	1699	0.8894	0.5619	0.7575
Churned	0.79	0.78	0.79	327			
KNN for SMOTE and Undersampled data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.95	0.88	0.91	1699	0.8578	0.5554	0.8139
Churned	0.68	0.86	0.76	327			

4.1.5.3 Results for the Naive Bayes Classifier for both Imbalanced and SMOTE and Undersampled data

Table 5: Naive Bayes Classifier

Naive Bayes for Imbalanced Data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.92	0.94	0.93	1699	0.8845	0.5527	0.762
Churned	0.66	0.58	0.62	327			
Naive Bayes for SMOTE and Undersampled data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.94	0.82	0.88	1699	0.805	0.4668	0.7837
Churned	0.44	0.75	0.55	327			

Table 6: Multi-layer Perceptron classifier

4.1.5.4 Results for the Multi-layer Perceptron classifier for both Imbalanced and SMOTE and Undersampled data

MLP for Imbalanced Data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.91	0.7	0.79	1699	0.691	0.2703	0.6762
Churned	0.29	0.65	0.41	327			
MLP for SMOTE and Undersampled data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.95	0.57	0.71	1699	0.6106	0.2978	0.7023
Churned	0.84	0.61	0.66	327			

4.1.5.5 Results for the Logistic Regression for both Imbalanced and SMOTE and Undersampled data

Table 7: Logistic Regression

Logistic Regression for Imbalanced Data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.87	0.97	0.91	1699	0.846	0.2812	0.5933
Churned	0.56	0.22	0.32	327			
Logistic Regression for SMOTE and Undersampled data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.91	0.84	0.88	1699	0.7981	0.3562	0.7006
Churned	0.41	0.56	0.47	327			

4.1.5.6 Results for the Random Forest Classifier for both Imbalanced and SMOTE and Undersampled data

Table 8: Random Forest Classifier

Random Forest for Imbalanced Data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.96	0.99	0.97	1699	0.9526	0.8168	0.8768
Churned	0.93	0.76	0.84	327			
Random Forest for SMOTE and Undersampled data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.97	0.0.97	0.0.97	1699	0.9531	0.8262	0.9115
Churned	0.86	0.85	0.85	327			

4.1.5.7 Results for the Ada Boost Classifier for both Imbalanced and SMOTE and Undersampled data

Table 9 Ada Boost Classifier

Ada Boost for Imbalanced Data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.96	0.98	0.97	1699	0.9521	0.8181	0.8949
Churned	0.88	0.81	0.85	327			
Ada Boost for SMOTE and Undersampled data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.98	0.95	0.96	1699	0.9423	0.8052	0.9298
Churned	0.77	0.91	0.84	327			

4.1.5.8 Results for the XGB Classifier for both Imbalanced and SMOTE and Undersampled data

Table 10: XGB Classifier

XGB for Imbalanced Data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.97	0.98	0.98	1699	0.963	0.8613	0.9236
Churned	0.9	0.87	0.88	327			
XGB for SMOTE and Undersampled data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.99	0.97	0.98	1699	0.9595	0.8575	0.9449
Churned	0.84	0.92	0.88	327			

4.1.5.9 Results for the LightGBM Classifier for both Imbalanced and SMOTE and Undersampled data

Table 11: LightGBM Classifier

LightGB for Imbalanced Data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.96	0.99	0.97	1699	0.9526	0.8168	0.8768
Churned	0.93	0.76	0.84	327			
LightGB for SMOTE and Undersampled data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.99	0.97	0.98	1699	0.96	0.8606	0.949
Churned	0.84	0.93	0.88	327			

Hence a summary of above classifiers are concluded considering the SMOTE and undersampled data

Table 12 Summary of the Classifiers for Credit Card churn Dataset

ML Algorithm	Accuracy	Mathews CC	AUC
MLP	0.6106	0.2978	0.7023
Logistic Regression	0.7981	0.3562	0.7006
Naive Bayes	0.8050	0.4668	0.7837
KNN	0.8578	0.5554	0.8139
Decision Tree	0.9117	0.7126	0.8905
Ada Boost	0.9423	0.8052	0.9298
Random Forest	0.9531	0.8262	0.9115

XGBoost	0.9595	0.8575	0.9449
LightGBM	0.9600	0.8606	0.9490

After going through the literature considering the machine learning techniques that were used by most of the researchers, identified the weak learners and then the model which performed better out of all is used as the meta model. The ensembler model is built as per the chapter 3 and below table shows the final result achieved by using the ensembler model. Further the same ensembler is applied twice, one with the imbalanced dataset and the other with the balanced data set using SMOTE and under sampling technique.

From the above results it is identified that the LightGBM classifier has the highest accuracy 0.9600, highest MCC value 0.8606 and the highest AUC value 0.9490. Hence Light GBM is taken as the meta model and the rest of the models are taken as the weak learners.

Then the combination of these weak learners and the meta model is introduced in the ensemble model using stacking technique. The result of the stack Ensemble model is listed below

4.1.5.10 Results for the Stack Ensemble model for both Imbalanced and SMOTE and Undersampled data

Table 13: Stack Ensemble

Stack Ensembler for Imbalanced Data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.98	0.98	0.98	1699	0.9645	0.8687	0.9344
Churned	0.93	0.76	0.84	327			
Stack Ensembler for SMOTE and Undersampled data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.98	0.97	0.98	1699	0.9645	0.872	0.9455
Churned	0.87	0.92	0.89	327			

According to the above results the same accuracy is reported for both imbalanced data and the SMOTE and undersampled data which is 0.9645.

But in terms of Precision, Recall, F1 Score, Mathews CC and AUC values the SMOTE and undersampled data has superseded the results of the imbalanced data.

Table 14 Comparison of Stack ensembler with the rest of the invidual models

ML Algorithm	Accuracy	Mathews CC	AUC
MLP	0.6106	0.2978	0.7023
Logistic Regression	0.7981	0.3562	0.7006
Naive Bayes	0.8050	0.4668	0.7837
KNN	0.8578	0.5554	0.8139
Decision Tree	0.9117	0.7126	0.8905
Ada Boost	0.9423	0.8052	0.9298
Random Forest	0.9531	0.8262	0.9115
XGBoost	0.9595	0.8575	0.9449
LightGBM	0.9600	0.8606	0.9490
Stack Model	0.9645	0.8720	0.9455

With the above results among other individual classifiers the ensembler has outperformed with the highest accuracy, Mathews CC, AUC, Precision, Recall and F1 Score values. The SMOTE and undersampled data of the ensembler has retrieved the higher values rather than the values for the imbalanced data.

Since the SMOTE and undersampled data has performed well when compared with the unbalanced data, the comparison is considered with the SMOTE and undersampled data.

4.2 Strengths of the Research

In this research the major strengths is the precise identification of credit card churn prediction as the result suggested that the built ensemble model is the best predictor of customer churn or the attrition when compared to the other individual models. Moreover this research is considered with the most of the transaction related attributes such as Months_on_book, Credit_Limit, Total_revolving balance, Total_Transaction_Count , Avg_Utilization_Ratio and etc and these features are prominent in identifying credit card churners. The main strength of the research is accruing such financially related data which largely contributes with the relationship of the Credit card service providers as there were not much researches had performed in this area considering such data

CHAPTER 5

CONCLUSION

Final chapter focuses on giving an overview of the research and summarizes the final outcomes aligning with the research objectives which are introduced at the beginning of the research.

5.1 Research Overview

The goal of this research is to address the problem of customer attrition in the credit card domain by attaining the following research objectives:

- Identify factors that contribute most to the customer's decision of a credit card service termination.

Above objective is achieved by the feature importance, correlation matrix and PCA analysis which are performed at feature selection.

Hence below attributes are considered to be the factors which mostly contribute to the customer's decision of credit card usage

Total_Trans_Ct, Total_Revolving_Bal, Total_Trans_Amt, Total_Relationship_Amt, Total_Ct_Change_Q4_Q1, Total_Amt_Change_Q4_Q1, Months_Inactive_12_months, Avg_Open_To_Buy, Credit_limit and Contacts_Count_12_months are considered to be the factors which mostly contribute to the customer's decision of credit card usage according to the feature importance with decision tree classifier.

After the correlation analysis it is identified that when the below attributes magnitude increases there's a less chance to churn

Total_Trans_Ct

Total_Ct_Chng_Q4_Q1

Total_Revolving_Bal

Avg_Utilization_Ratio

Total_Trans_Amount

Total_Relationship_Count

Total_Ct_Chng_Q4_Q1

Credit Limit

While the churning is increased with the increase of below attributes

- I. Contatcts_Count_12_Months
- II. Months_Inactive_12_Months
- III. Customer Age
- IV. Dependent Count

- Using these factors, perform data mining techniques to understand customer retention patterns by classifying policy holders who are likely to continue or terminate their credit card.

Above objective is achieved by performing apriori algorithm as per the chapter 3.

- Building a model to predict the credit card holders who are about to churn

Above objective is achieved via the ensemble model with the highest accuracy of 0.9645, F1 score and AUC values superseding the rest of single machine learning models

Stack Ensembler for SMOTE and Undersampled data							
	Precision	Recall	F1 Score	Support	Accuracy	Mathews CC	AUC
Non Churned	0.98	0.97	0.98	1699	0.9645	0.872	0.9455
Churned	0.87	0.92	0.89	327			

Figure 45 Conclusion from the ensemble model

- From the model, predict whether the specific customer will churn or will continue the service with the model prediction and then the potential churners can then be targeted for promotions to reduce the rate of attrition

A web application is implemented to be used by the domain experts where they have given the ability to predict whether the potential credit card user may churn or not churn with the probability measurement. Once the potential churners are identified these customers can be targeted on promotion campaigns which motivates them to retain in the service

Here the Chapter two described literature survey related to the domain and as well as technical aspects that could be applied to reach the research objectives. Design and methodology used along with the exploratory data analysis of the research is detailed in the chapter three. Chapter four outlined the evaluation of the model designed.

5.2 Limitations of the Research

The selected dataset comprised with customer details and the transactional details only. But it did not comprised with the macro economic factors. Moreover the bank specific details such as interest rate, annual fee and other offers given by the credit card service provider are not included in this data set. It would be ideal to do a broader research with the above mentioned factors as well. Further the dataset comprised data for a period of 1 year, if there's a possibility to have a data values for couple of years much advanced predictions could be achieved.

5.3 Future Work and Recommendations

Here in this research the various attributes can be explored not being limited to the transactional data but also other socio economic factors which affect the churn of the credit card users.

Different types of ensemble models can be used to develop the model and further investigation can enhance a new area of knowledge as well.

Further recommendation is to build the time-series model to predict customer churn.

APPENDICES

- URL for data set

https://drive.google.com/file/d/1cr0ByaE_6jfcv7hylaeqsmgyLhzuljtQ/view?usp=sharing

- URL for results

<https://drive.google.com/file/d/1cheUFXEjRPhoTYADQvXiRdL0Ko-V2NbC/view?usp=sharing>

- URL for Source codes

<https://drive.google.com/file/d/1cheUFXEjRPhoTYADQvXiRdL0Ko-V2NbC/view?usp=sharing>

Bibliography

- Ahmad, A.K., Jafar, A., Aljoumaa, K., 2019. Customer churn prediction in telecom using machine learning in big data platform. *J Big Data* 6, 1–24. <https://doi.org/10.1186/s40537-019-0191-6>
- Ardabili, S., Mosavi, A., Varkonyi-Koczy, A., 2019. Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods. <https://doi.org/10.20944/preprints201908.0203.v1>
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *jair* 16, 321–357. <https://doi.org/10.1613/jair.953>
- Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., Mosavi, A., 2019. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Science of The Total Environment* 651, 2087–2096. <https://doi.org/10.1016/j.scitotenv.2018.10.064>
- Clemente, M., Giner-Bosch, V., Matías, S., 2012. Assessing classification methods for churn prediction by composite indicators.
- Cohen, D., Gan, C., Yong, H., Chong, E., 2007. Customer Retention by Banks in New Zealand. *Banks and Bank Systems* 2, 40–55.
- Customer Churn Analysis In Banking Sector Using Data Mining Techniques | Semantic Scholar [WWW Document], n.d. URL <https://www.semanticscholar.org/paper/Customer-Churn-Analysis-In-Banking-Sector-Using-Oyenyi-Adeyemo/0fd0067775eef2aa52547229c17e06128fdc0633> (accessed 5.21.21).
- Dutta, S., Bandyopadhyay, S., 2020. Employee attrition prediction using neural network cross validation method. *International Journal of Commerce and Management* 6, 80–85.
- Gajadeera, H., n.d. Developing a model for knowledge management practices in SMEs of software development industry 1.
- Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. *Mach Learn* 63, 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Gladly, N., Baesens, B., Croux, C., 2009. Modeling churn using customer lifetime value. *European Journal of Operational Research* 197, 402–411. <https://doi.org/10.1016/j.ejor.2008.06.027>
- Hassan, H., Bin-Nashwan, S., 2017. Impact of customer relationship management (CRM) on customer satisfaction and loyalty: A systematic review. *Research Journal of Business Management* 6, 86–107.
- Hudaib, A., Dannoun, R., Harfoushi, O., Obiedat, R., Faris, H., 2015. Hybrid Data Mining Models for Predicting Customer Churn. *International Journal of Communications, Network and System Sciences* 08, 91–96. <https://doi.org/10.4236/ijcns.2015.85012>
- Jamalian, E., Foukerdi, R., 2018. A Hybrid Data Mining Method for Customer Churn Prediction. *Eng. Technol. Appl. Sci. Res.* 8, 2991–2997. <https://doi.org/10.48084/etasr.2108>
- Kelleher, J.D., Namee, B.M., D’Arcy, A., 2015. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press, Cambridge, MA, USA.
- Khagi, B., Kwon, G.-R., Lama, R., 2019. Comparative analysis of Alzheimer’s disease classification by CDR level using CNN, feature selection, and machine-learning

- techniques. *International Journal of Imaging Systems and Technology* 29, 297–310. <https://doi.org/10.1002/ima.22316>
- Kumar, D.A., Ravi, V., 2008. Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies* 1, 4–28.
- Lejeune, M.A.P.M., 2001. Measuring the impact of data mining on churn management. *Internet Research* 11, 375–387. <https://doi.org/10.1108/10662240110410183>
- Lima, E., Mues, C., Baesens, B., 2009. Domain knowledge integration in data mining using decision tables: case studies in churn prediction. *Journal of the Operational Research Society* 60, 1096–1106. <https://doi.org/10.1057/jors.2008.161>
- Longadge, R., Dongre, S., 2013. Class Imbalance Problem in Data Mining Review. *arXiv:1305.1707 [cs]*.
- Ma, Y., He, H. (Eds.), 2013. *Imbalanced Learning: Foundations, Algorithms, and Applications*, 1st edition. ed. Wiley-IEEE Press, Hoboken, New Jersey.
- Nie, G., Rowe, W., Zhang, L., Tian, Y., Shi, Y., 2011. Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications* 38, 15273–15285. <https://doi.org/10.1016/j.eswa.2011.06.028>
- Odusami, M., Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., Sharma, M.M., 2021. A Hybrid Machine Learning Model for Predicting Customer Churn in the Telecommunication Industry, in: Abraham, A., Sasaki, H., Rios, R., Gandhi, N., Singh, U., Ma, K. (Eds.), *Innovations in Bio-Inspired Computing and Applications, Advances in Intelligent Systems and Computing*. Springer International Publishing, Cham, pp. 458–468. https://doi.org/10.1007/978-3-030-73603-3_43
- Ohny, M., Mathai, M.P.P., 2017. Customer Churn Prediction: A Survey. *International Journal of Advanced Research in Computer Science* 8, 2178–2181. <https://doi.org/10.26483/ijarcs.v8i5.4079>
- Parveen, A.N., Inbarani, H.H., Kumar, E.N.S., 2012. Performance analysis of unsupervised feature selection methods, in: 2012 International Conference on Computing, Communication and Applications. Presented at the 2012 International Conference on Computing, Communication and Applications (ICCCA), IEEE, Dindigul, Tamilnadu, India, pp. 1–7. <https://doi.org/10.1109/ICCCA.2012.6179181>
- Pendharkar, P.C., 2009. Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services. *Expert Systems with Applications* 36, 6714–6720. <https://doi.org/10.1016/j.eswa.2008.08.050>
- Rico-Poveda, C.A., Galpin, I., 2020. Forecasting Credit Card Attrition using Machine Learning Models. *undefined*.
- Schröer, C., Kruse, F., Gómez, J.M., 2021. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science, CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020* 181, 526–534. <https://doi.org/10.1016/j.procs.2021.01.199>
- Shaikh, R., 2018. Feature Selection Techniques in Machine Learning with Python [WWW Document]. Medium. URL <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e> (accessed 8.29.21).
- Singh, H., Rana, P.S., Singh, U., 2018. Prediction of drug synergy in cancer using ensemble-based machine learning techniques. *Mod. Phys. Lett. B* 32, 1850132. <https://doi.org/10.1142/S0217984918501324>

- SMOTE for Imbalanced Classification with Python [WWW Document], n.d. URL <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/> (accessed 8.30.21).
- Sokolova, M., Japkowicz, N., Szpakowicz, S., 2006. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation, in: Sattar, A., Kang, B. (Eds.), *AI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, pp. 1015–1021. https://doi.org/10.1007/11941439_114
- Umayaparvathi, V., Iyakutti, K., n.d. A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics 03, 7.
- Universitat Politècnica de València, E., 2014. Universitat Politècnica de València. *ing.agua* 18, ix. <https://doi.org/10.4995/ia.2014.3293>
- Zhang, X., Mahadevan, S., 2019. Ensemble machine learning models for aviation incident risk prediction. *Decision Support Systems* 116, 48–63. <https://doi.org/10.1016/j.dss.2018.10.009>
- Zhao, Y., Hryniewicki, M.K., Cheng, F., Fu, B., Zhu, X., 2019. Employee Turnover Prediction with Machine Learning: A Reliable Approach, in: Arai, K., Kapoor, S., Bhatia, R. (Eds.), *Intelligent Systems and Applications, Advances in Intelligent Systems and Computing*. Springer International Publishing, Cham, pp. 737–758. https://doi.org/10.1007/978-3-030-01057-7_56