S	
E1	
E2	
For Office Use Only	

Masters Project (MCS)

Final Report

2019

Project Title	Accelerate Selling to Zebras selling process using machine learning techniques
Student Name	E.U.I. Sumanarathne
Registration No. & Index No.	14440751 & 2014MCS075
Supervisor's Name	Dr. Manjusri Wickramasinghe

For Office Use ONLY

Declaration

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

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

Student Name: E.U.I. Sumanarathne

Registration Number: 14440751

Index Number: 2014MCS075

_____

Signature: Date:2019/05/31

This is to certify that this thesis is based on the work of Mr. E U I Sumanarathne under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name:

_____

Signature: Date:

Accelerate Selling to Zebras selling process using machine learning techniques

A dissertation submitted for the Degree of Master of Computer Science

E U I Sumanarathne
University of Colombo School of Computing
2019

Acknowledgements

First and foremost, I would like to offer my sincere thankfulness to my supervisor Dr. Manjusri Wickramasinghe, lecturer at University of Colombo School of computing, for the guidance and support he had given me throughout this research. He had been a great listener, a great advisor and a great teacher.

Abstract

Sales opportunity outcome prediction is the foundation for effective and productive sales management. Selling is easy if sales people can identify prospects who is willing to buy from them. Selling to zebras is a proven methodology for complex sales process which implemented as an application as well. This method could identify prospects which are best fit for an organization. So that, the organizations don't need to waste time on prospect that are not fit for them. The Selling to Zebras has sales stages which represented the state of a sales opportunity. Besides from that, Selling to Zebras method uses a scoring method call zebra scoring to identify qualifying deals. Sale opportunities which have higher zebra score considered as opportunities that the sales people should pay more attention. This research proposed a method which uses machine learning techniques with Selling to Zebras data set to predict sales opportunities outcome. The proposed method uses a machine learning driven classification model to predict the outcome. In summary, based on the predicted result sales people can detect the most potential opportunities in their pipeline. Then they can apply selling to zebras methodology on that. The research demonstrates the applicable machine learning method on anonymized data set provided by Selling to Zebras Inc.

Table of Contents

List of Figures

List of Tables

Chapter 1

1. Introduction

Selling products and services is indisputably one of the most challenging tasks for an organization. Most organizations have separate sales teams to sell their products. These sales teams are consisted with sales representatives and sales managers, sales engineers, customer service people and marketing people. Each individual of a sales team has assigned targets which they need to achieve within a given period of time. Ability to sale products may differ on experience and skills of a sales person. Usually sales reps are using a customer relationship management system as known as CRM to keep track of their customer related information. Sales representatives find people, who could be their potential customers, but they don't have enough information about them that sales reps don't have relationship with yet. These people are called as sales leads. Sales people also have contacts those who have previously done business together. CRM consists with Sales Leads, Contacts, Accounts which are business entities where Contacts work for Account, Opportunities which are sales events associated to an Account and one or more Contacts. These lists of possible sales leads create a sales pipeline.

When salespeople need to close more sales what they do is try to pursue every sales leads and contacts they have. This type of approach will fail because most of the time salespeople waste time on the prospect that they will not buy from them. Traditional sales management process philosophy is that all sales activity is a good activity and worth spending time on it. This will result talk to 100 prospects to get 25 appointments so you can do 12-13 prospect surveys which will lead to about 6-7 proposals and ultimately all this effort ended up with only one sale. Since salespeople pursue every possible opportunity in their sales pipeline, they have exhausted their limited number of resources with poor judgement. Once they have nothing left to spend on the best opportunities in their pipeline, those deals might be evaded. Unfortunately, from this traditional process most of the organizations only close 15% of their sales pipeline. However, nowadays companies tend to use automatic sales analysis tools to improve their sales process. McAfee and Brynjolfsson [3] said, the companies those using data-driven decision support systems, in the top third of their industry are, on average, 5% more productive and 6% more profitable than their competitors. In this study we try to accelerate the selling process of the selling to zebras method using machine learning techniques.

1.1 Selling to Zebras

There is a conclusion of the above paragraph is 85% of the time salespeople spend their time on prospects who don't want to buy from them. It could have been great if sales people spend more time on pursuing prospect that they know they could win. They are the customers who are perfectly matched to the salesperson's product. This is where the Selling to Zebras [1] comes from. Zebras are the animals which can be easily recognized from others. You can't mistakenly recognise a Zebra for any other animal. In the sales context, Zebras are the prospects that salesperson could sell easily, and they are the perfect fit for your organization. The good salesperson has an instinct for Zebras .As the same as the Zebra's stripes salesperson can identify these prospects by looking at their objective characteristics, those are perfectly aligned with what they sell. Organizations or individuals never buy a product, that doesn't line up with what they require an even skilful salesperson try thousands of times. Therefore, salespeople should pursue only Zebras. Selling to zebras [1] is a selling process, which help salespeople to identify the perfect prospect for their organization and it will help them to close the deal 90% of the time.

There can be an argument if someone pursues only Zebras, he or she is risking selling less than they are currently at. Since they are focusing only on Zebras, total value of their sales pipeline will be decreased. But at the end of the day you will be closing 90% of your sales pipeline which is far better than 15%. Using this method sales people can save their important time to sale their products to quality accounts where you are adding unique value. While you are selling, you could identify your strengths and how you could diverse, in order to mitigate your weaknesses.

The most skillful and experience salespeople in the world have a common attribute, which they can sense their critical sales issues and how to address them. Most common critical issues are, sales cycles are getting longer and often end with no decision, qualified prospects are hard to find since the sales pipeline is blocked, products or services becoming increasingly similar, deals are becoming smaller and profits are diminishing since giving heavy discounts and access to high-level decision makers are hard to get and number of competitors in the market has been raised. Every organization faces at least one of these issues.

Overcoming sales issues more important than focusing on increasing sales activities. If an organization considers they could get more sales by increasing sales activity quantity over quality generally yields a destitute return on investment (ROI). In the traditional sales process, the people in a sales department

working towards different directions to achieve their sales targets. The outcome of this is salespeople might lose focus on activities where it truly matters. They might have not enough time left to focus on deals that they should win. With the Zebra model people will have enough information to take out prospects that has unfavorable sales cycles and respond with fullest capacity towards the deals that perfectly fit to the organization as a Zebra profile.

The Zebra model has a sales cycle call zebra buying cycle. The traditional method sales people hunt for some to sell to. The zebra buying cycle process helps salespeople to get on board a customer who buys the product or service. The buyer should be a person who is defining the company's critical business issues, who has the power of decision making and who is taking the responsibility to get the project approval and is responsible for the result. This person is called as the Power.

The solution that is presented to the Power should address their business's key pain points that are related to the business issues. Then the Power has the authority to make the decision to buy. The Zebra buying cycle enables sales person to realize the pain points that they need to reach to Power and make the most of their time and resources worth. This will avoid the needless time wasting on people that will never buy from you.

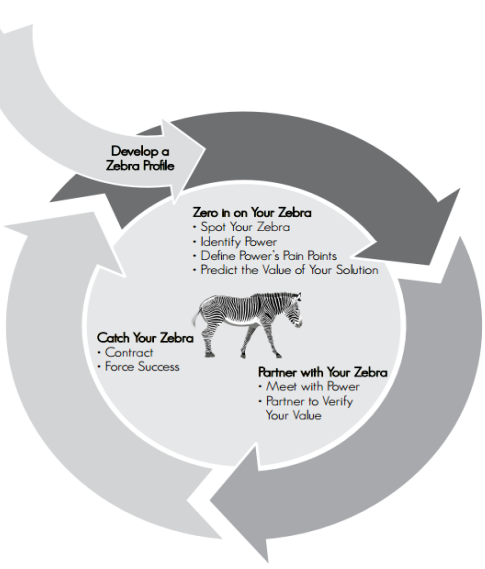

Figure 1: Zebra buying cycle

During the process, Power is presented with the information of a predicted value that they can expect to achieve, make their decision making easier and suggest how to progress on the solution and what it takes

in terms of resources and time and finally clearly state what will they lose if you are not going to buy this product.

The Zebra and Zebra Buying Cycle have been modeled to answered following three most important questions.

- Will this prospect buy anything?
- Will this prospect buy from me?
- Will this prospect buy now?

If a deal unable to answer "yes" to one of the above questions is not even considered as a prospect.

Creating a zebra profile involves a scoring method call Zebra Scoring. This has been designed using organizations, business characteristics and project characteristics. Business characteristics are Company Characteristics, Operational Characteristics, Technology Characteristics and Service Characteristics. Project characteristics are Access to Power, Funding and Return on Investment. These seven characteristics also known as zebra attributes. Each of these seven characteristics assigns a score between 1 – 4. If a salesperson scoring correctly if it is a Zebra it will be more likely to have a higher zebra score. Based on the zebra score salespeople can decide whether the prospect is zebra or not. A Zebra profile depicts the characteristics of those organizations that has your best opportunities. If the total zebra score for an opportunity is 0 – 16 it is said to be a high-risk prospect which isn't a good fit for your company. If the total zebra score is 17- 22 it has some risk and it may or may not be worth of pursuing. If the zebra score is 23-28, most probably the prospect is a Zebra and so it has a higher chance of winning and salespeople should chase it.

The Selling to Zebras process defines sales stages for represent status of an opportunity.

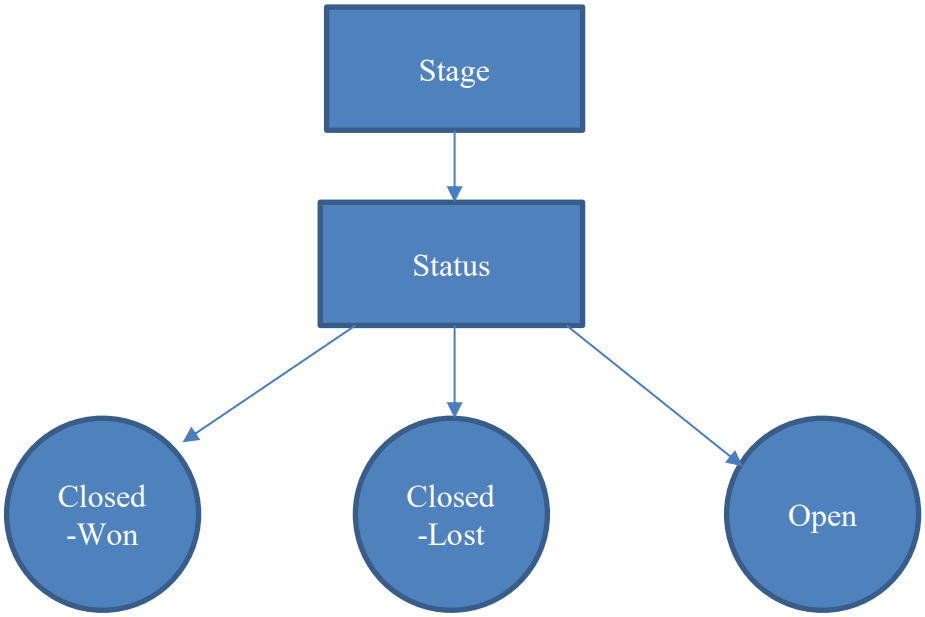

Figure 2: Sales stage features

Sales stage has sales status, a status can be Closed-Won, Closed-Lost or Open. The Closed-Won represented as "W", Closed-Lost represented as "L" and Open represented as "O". The Closed-Won and Closed-Lost status are final status, where opportunity has an outcome. The Open status means opportunity is in an in progress state where opportunity has not done yet. Figure 3 shows the sales stages of Selling to Zebras.

Sales Stages
1 - Zero in your Zebra - O
2- Identify Power - O
3- Meet with Power - O
4- Partner to Verify Value - O
5- Co-Present Findings - O
6- Negotiation - O
7- Perception Analysis - O
8- Negotiate Contract - O
9- Closed Deal - W
10- Force Success - W
11- Lost - L

Figure 3: Selling to Zebras sales stages

If an opportunity is in a 'W' stage the winning probability is 1 if the stage is 'L' the winning probability is 0. If we can predict winning probability of an opportunity based on its sales stage, management decisions can be made effectively.

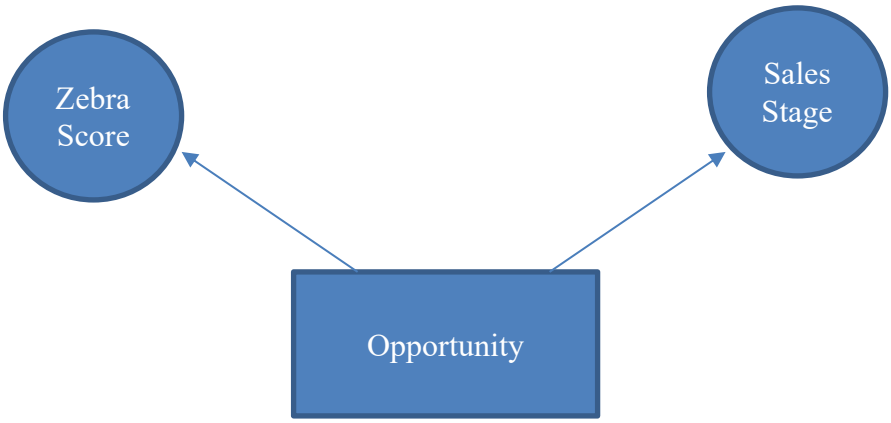

Figure 4: Selling to Zebras opportunity, two most important attributes

As we discussed in the above paragraphs, for a given instant opportunity has a Zebra Score and a Sales Stage. We will discuss more on these attributes when we explain about the dataset.

1.2 Motivation

Usually, the organizations keep their customer information in a CRM. It is possible to find out best customers from a CRM database using machine learning techniques, which is known as customer classification [2]. Let's assume we have found out such customers from our CRM database. Then we need to find a way to sell our product to them. The selling to zebras is a proven sales process for accelerating selling process. This method has few steps which need to perform by a sales person or a team.

- Identify your perfect prospect
- Research the Zebra to identify its critical business issues
- Create persuasive value-based presentations
- Obtain executive-level buy-in early in the process

During this process opportunity can be won or lost after any open sales stage which can be an early stage like "Partner to verify value" or latter stage like "Negotiate contract". But sales people don't know the

exact sales stage, where they know an opportunity is going to win or lose. They could predict the outcome for an opportunity based on their past experiences. To do that, they need to go through their past activities. It might take a considerable amount of sales persons' valuable time. This is where we can take the advantage of machine learning. We can use machine learning techniques to analyse sales people's past activities and based on that we can give valuable insights to the business.

1.3 Research Objective

The main aim of this research is to discover adequacy of machine learning based approach to predict outcome of an opportunity, which based on a selling to zebras sales data set. Sales people develop common patterns of updating and managing opportunities over time. Using machine learning, we can predict each opportunities likelihood of becoming Closed-Won (a win) .

Hypothesis - For a given customer, we can auto-calculate the outcome of a given opportunity based on the organizations past performance.

1.4 Research Scope

The study will focus on how selling to zebras' past activities influenced the outcome of an opportunity. Supervised classification methods have been used in the data set is being collected from Selling to Zebras during 2013 January to 2013 December.

1.5 Organization of the Dissertation

The first part of the chapter will be discussed about selling to zebras' sales methodology and the motivation for this study. The second chapter is dedicated to discuss about the background information including theoretical and domain. The chapter 3 forms a design for the proposed method, where chapter 4 introduced informative discussion about experimental setup. The fourth chapter of the research has reserved for the evaluation of the model. The conclusion will be discussed in the final chapter of the study.

Chapter 2

2. Literature Review

In the first chapter we have discussed the background about selling to zebras. This chapter highlights the previous studies and results justified as the foundation of this research. In first part of this chapter we will focus on some of the basic concepts and tools use in sales industry. CRM is the fundamental tool used in the sales process. In this chapter we will examine some of the previous studies have published in the context of data mining in CRM. Then, we discussed about few selected applications of machine learning in sales domain.

2.1 Sales Process and Opportunity Pipeline

Organizations sales teams use a sales pipeline method to handle a stream of leads through various sales stages of the sales process, at which some opportunities are eliminated, while others extend through to successful closure, resulting in revenue for the firm [4], as described in Figure 5. There is a sales model, which considers three sales stages with static results. Those are generation of new leads, the conversion of these leads into appointments, and the subsequent conversion of the opportunities to closed sales with the sales close rate. The close rate is a basic measure of sales people performance [5].

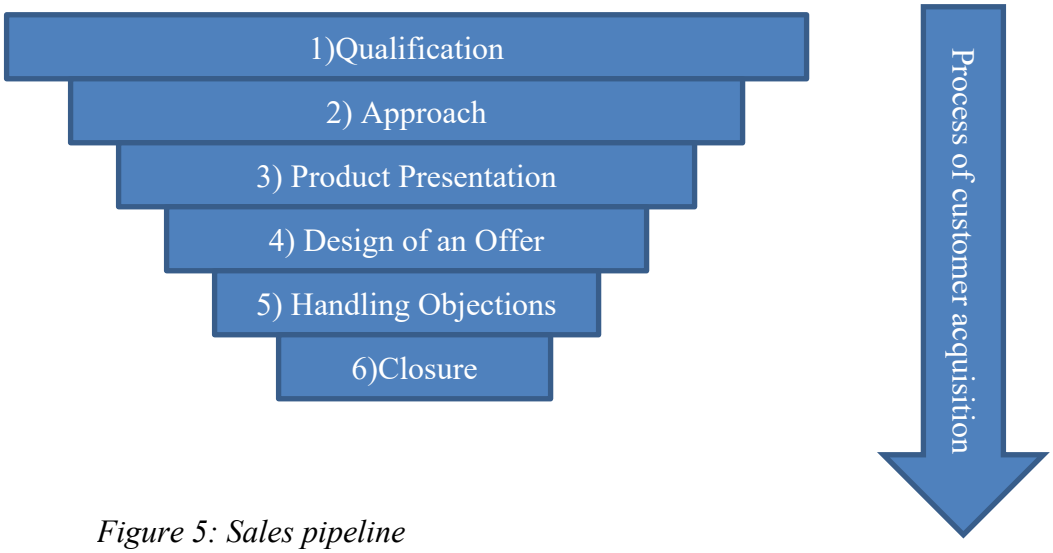

Figure 5: Sales pipeline

2.2 Sales forecasting

A research states data mining techniques are extensively applied in to customer relationship management(CRM) systems. However, changes in the market, unorganized information, be in need of academic application frequently has an effect on the sales forecasting.

Sales forecasting requires informative knowledge about historical sales activities and domain knowledge. As in any other data mining tasks, the very first and most challenging task of sales forecasting is knowledge representation. In general, features in a data set have been assembled from theoretical application and sales professionals inputs. Sales experts can be influential in feature selection from their past experience, which creates a base to describe sales context both won and lost sales opportunities.

A study[6] has suggested that Business to Business(B2B) sales forecast will be represented as a decision-making method, which is predicated on historical information, formalized rules, subjective judgment, and implicit structure data. Further it proposed a machine learning driven classification model based on insights from sales professionals in B2B domain.

2.3 Data mining (DM)

Data mining is a process of discovering hidden patterns and the information from the existing data.

2.4 Customer Relationship Management (CRM)

Organizations use Customer Relationship Management to support sales people to keep track of their customer information, help sales people to identify and grow sales prospects [6]. The CRM constructed with the fundamentals such as customer relationship proposals and managing a customer, including its termination [6]. It enhances teamwork and performance of an organization by improving task management and inter organizational communication [6]. Also, CRM improves effectiveness of sales. It is a management philosophy and a strategy which enables an organization to optimize revenue and increase customer value and service quality through understanding and satisfying the individual customers' needs [6].

2.5 CRM in Data mining

In today's context, data mining has gained a great deal of attention, mainly in the information industry. This method is very popular in data analysis, which has been considered as a newly emerging analysis tool [38]. CRM stored large amounts of complex data. DM, a successfully proven technique to extract useful insights from complex data, is facilitate identifying customer demand precisely and endorse customer value effectively [36]. CRM data contain useful, valid, novel, potentially useful hidden patterns and correlations which can be identified by using data mining techniques. Data mining considered as an important aspect in CRM, such as generating more profit for customers, retaining and making the valued customers and so on. CRM becomes stronger with the use of DM. DM based CRM identify the customer's need for the full as possible. It increases the level of customer satisfaction, provide insights which help to gain more quotes in the market and promote the profitability. Therefore, it boosts organization's competitive advantage. In this section, some of the cutting-edge systems in DM based CRM is discussed [40].

Chang, Lin and Wang [37] have studied the extraction of textual information to optimize customer relationship management. They have collected data from the customer service centre from three sources: (1) electric news; (2) the customer service hotline; and (3) data from various conferences. Usually these customer data were not clearly structured, and content was diversified because, the data were mainly imported from emails. Thus, the quantitative analysis has been done using the content analysis in this research. Content analysis was used to examine the text data. The model suggested marketing strategies for the CRM using customer insights. These customer insights were discovered using OLAP and decision tree analysis. However, the accuracy of the system required a good information system otherwise, structured data was scattered. Data mining did not produce noticeable discoveries, so the data analysis was undeniably sketchy. As a result, recommendations of this research concentrate on the complete CRM execution process and on establishing a complete system cycle to enhance customer interactions. They also used established categories by categorizing text data in follow-up studies. In addition, the process of using content analysis to transform text data into structured data is used to enhance the training of the operating staff of the system. The study included an implementation as well. Main characteristic of this model is the staff involvement in CRM concepts. The advantage of this study is developing a data mining based CRM model easily. Though, the research has been conducted to a poor number of companies.

Hosseni et al. [38] have implemented a data mining approach in SAPCO organization from the 20th March to the 21st November in 2008. They have used expanded WRFM model into K-means algorithm. As a result, they have achieved an impressive classification accuracy. Generally, many researchers use cluster distance and integrated rate as two separate parameters. However, for this study they have combine these two parameters and treated as one. They found out there is a greater ability to cater the customer loyalty when a sales firm creating a marketing strategy. Statistically they have shown that their data mining technique for CRM has adequate result to prove, their methodology has a higher level of confidence with compare to the techniques that other organizations used. However, they had the advantage of customer loyalty in financial sector in comparing to others.

As another research, Khan et al. [39] introduced a method, which uses CRM and Data Warehousing. They extracted and analysed the facts and information existing to make a critical assessment. Deepening the study and evaluation of the available information allowed to determine several benefits that CRM and data warehousing integration can bring to the industry as well as to customers. The study has listed benefits of this integration in CRM domain. It reduces the cost and time in customer acquisition, which shorten the sales cycle length. The method also enhances, customer satisfaction, customer retention rate, customer close rate, return on investment, competitive edge over other organization, revenue per customer. They have considered only one company for this evaluation, which is the main limitation of this study.

The research of Wei et al. [40] investigate customer classification of hairdressing industry in Taiwan. Their study based on a previous study, where it proposed RFM model, which uses two stage clustering method. They have use K-means and SOM together to adopt this RFM model in order to semantically analyse customer details of a salon data dataset. The result of this RFM model outlined there are four type of customers in the dataset, which are loyal customers, potential customers, new customers and lost customers. The study has helped to identify different customer groups and helped to implement distinct marketing strategies for each customer groups. Hair salon can be benefited, because they can develop different marketing strategies by targeting specific customer profiles. The main limitation of this research is, it focused only on specific country and specific set of customer data.

Jiang et al. [41] has analysed customer information based data mining techniques. GIS can be used to obtain customers' geographical location. Also, customer information like geographical location, age, gender, tuition time, school time are analysed by DM techniques to discover knowledge for salespeople

to understand present trend in the market. The study has conducted on more than 50000 of customer data extracted from numerous training departments. Geographical information of this customer data has been gathered using Google Maps API [https://developers.google.com/maps/]. Customer information analysis has been done using an improved clustering algorithm. The improved clustering algorithm, based on traditional K-means and K-centroids, which extracted out clustering features of geographical information from above mentioned dataset. This geographical location-based approach helps to determine market trends in specific areas, how the promotion campaigns precisely conducted and improve cost effectiveness of such campaigns. This can be used as a decision-making tool in the process of developing marketing strategies. The most noticeable limitation of this method is, it didn't consider the population size and structure of different communities.

Bahari and Elayidom [42] proposed a system to help retain customers by predicting their behaviours. They have used a dataset of a direct banking marketing campaigns for this study. The experimental data setup by pre-processing the 10% from above mentioned dataset. Sixteen input variables and two classification algorithms have been used to predict the customer behaviour. They have compared and outlined the performance of classifiers in terms of accuracy, sensitivity and specify. This CRM data mining framework with two classification models. Which are Naïve Bayes and Neural network. They have shown that the Neural Network give accurate results compare to Naïve Bayes classifier. Based on the customer behaviour prediction, organization can made strategies to retain the customers. The limitation of this technique is they focus only on Neuro-fuzzy classifier.

Table 1: Advatages and Disadvantages of recent studies on data mining in CRM

Authors	***Advantages***	***Disadvantages***
Chang, Lin and Wang (2009)	• Easily executable • Fully implemented in CRM • Foundation of buiding CRM	• The research was limited to certain companies. It could have been used wide range of company data
Hosseni et al. (2010)	• Customer loyalty evaluation in service industry	• Research was limited to only one organization
Wei et al. (2013)	• Develop marketing strategies for unique customer groups	• Implemented only in one country by focusing on particular customer data
Jiang et al. (2013)	• Provide information to develop marketing strategies for training institutes	• Used no information about population structure and population size within different communities
Bahari and Elayidom (2015)	• Predict the customer behaviour	• Didn't used different AI techniques. Focus was only on neural network call Neuro-fuzzy classifier

2.6 Machine learning application in sales domain

Mostly the machine learning models are used in decision support process and sales forecasting process in sales domain. In this section we will be discussing about few machine learning models in sales context. Merkert, Mueller, and Hubl [43], stated the usefulness of a machine learning model depends on the task, applied technologies and the decision making stage. Their study has based on 52 research papers from 1993 to 2013. Our approach also supported to these three phases evidently.

Meyar et al. [44] proposed a method of improving dynamic decision making. The system uses Prediction of Control Errors in Dynamic Context (PRCEDO) methodology to improve feedback control strategies. It does help to make decisions in a complex dynamic context. Factors for chance of getting an unwanted or suboptimal result is identified by ML and necessary improvements are suggested. They have proven their method by experimenting on a medical and two manufacturing use cases. They have used decision tree method as their predictive model. The decision tree fulfils the requirement of outcome interpretability. Since the choice of ML model is limited, use of a powerful ML predictive model listed under future works. Furthermore, Florez-Lopez and Ramon- Jeronimo [45] study proposed more powerful ML predictive model by integrating both decision tree and correlated adjacent decision forest (CADF). This model provides both interpretable and highly accurate prediction results in a much complex environment.

The review of Armstrong, Green and Graefe [46], reviewed 105 papers with the comparisons and introduced a set of guidelines to follow in the sales forecasting. The study concluded with advices to make forecasting to any situations. Our data driven ML approach has introduced without violating the guidelines mentioned in this method.

Yan et al. introduced ML based sales pipeline win-propensity score for a given period of time, which based on sellers past sales activities. The dataset contains past opportunity information list with their outcome and sellers' activities on that, where outcome can be win or lose. The system predicts win-propensity score for a new opportunity to a given future time frame. The results shown human based rating outdone by ML based method. The method helps the higher management to make their resource allocation decisions efficiently, which increases the productivity and cost effectiveness. D'Haen and Van der Poel [47] introduced a three phase ML model, which helps in the customer acquisition process. Their aim is to produce a top grated list of potential clients, which are most likely to become new sales opportunity, finally clients. The potential clients have a rank. The sales representative can pick an opportunity from the top half of the potential clients list. Since most of the time sales representatives pursuing top rated prospects will improves the conversion rate of sales opportunity to client. The limitation of this study was, it is not possible to run this model in a real world scenario.

M. Bohanec et al. [48] introduces a B2B sales forecasting method, which uses novel approach of general explanation methodology. The approach is to bring high performing ML models, like random forest, support vector machines (SVM) into action where it requires transparency and comprehensibility.
The system goes further than the numerical win-propensity score method to further improve the results, by offering contextual advice to develop opportunities to sell an opportunity to an established customer. This method can be slow for large dataset. Therefore, precomputation mechanism should be implemented in order to use in a real-time decision-making scenario. The explanations going after prediction mode accordingly. If the prediction model giving wrong results or perform badly, the explanation will be affected.

2.7 Machine learning algorithms

There are three major different categories in machine learning algorithms based on the nature of data. Those are supervised learning, unsupervised learning and reinforcement learning [11]. Supervise machine learning compose with set of inputs and desired outputs which is known as training dataset. This method produces an inferred function by evaluating training dataset's inputs and outputs, which can be used to predict output of a future event [12]. Unsupervised learning in the other hand doesn't contain any structure or any labelled data, where the algorithm has the responsibility to find the structure based on the given data [13]. The third category reinforcement learning involves a computer program, which interact with a dynamic environment in order to learn using trial and error [14, 15].

This thesis target to build a model to come up with a predictor which correctly prove the given dataset. The given dataset has been labelled and outcome also known. Hence, it is quite fitting to study this case on supervised learning algorithms. The proposed model's output can have discrete number of values. The outcome for a given opportunity can only be won and lost, which can be labelled as 0 or 1 (More discussion on the dataset will be introduced in the next chapter). Apparently the model become a binary classification problem. Next part of this chapter will describe more about algorithms involved in this model.

2.8 Logistic Regression

Logistic regression is a statistical model, which used to predict binary outcome of a dependent variable with one or more descriptive independent variables. It constructed based on a linear function, which maps log ratios of probability of different outcomes and easily applicable to a dataset [16]. The outcome of the logistic regression is always a quantifiable value, which can be considered as the main advantage of this model. This quantifiable value is probability of the predicted outcome given a set of features [17].

In this thesis logistic regression has been used as a classifier with the intention of identifying most potential opportunities from the opportunity pipeline. Specifically, the classifier assigns an essential measure to each opportunity, which is the probability of win of a given opportunity. It gives insight to the sales teams whether to pursue or give up. Since the output is a probability, the measure always be within 0 and 1 [18]. To this end, within a logistic regression framework, the hypothesis or predictor function $h\theta(\cdot)$ is chosen to be the logistic or sigmoid function described as *Figure,* Which always return a value between 0 - 1

$$h_\theta(x) = \frac{1}{1 + e^{-\theta T_x}}$$

2.9 Support Vector Machines (SVMs)

Support Vector Machines(SVMs) can be found as a newest supervised learning technique which revolves around the margin of two data classes which is separated by a hyperplane. Main idea of SVMs is to determine the largest possible distance to the hyperplane by maximising its margin which reduces the upper bound of generalization error.

Assume that data set is linearly separable and,(w,b) exist where,

$\mathbf{w^T x_i + b \geq 1 \text{ for all } x_i \in P}$
$\mathbf{w^T x_i + b \leq -1 \text{ for all } x_i \in N}$

Decision rule : $\mathbf{f_{w,b}(x) = sgn(w^T x + b)}$ w - weight vector , b - bias or threshold

Above formula shows that if the data is linearly separable, we can minimize the squared norm of hyperplane. Therefore, next is to find how to set up minimization. It can be achieved by convex quadratic programming (QP)

Minimize : $\underset{w,b}{\Phi}(w) = ½||w||^2$

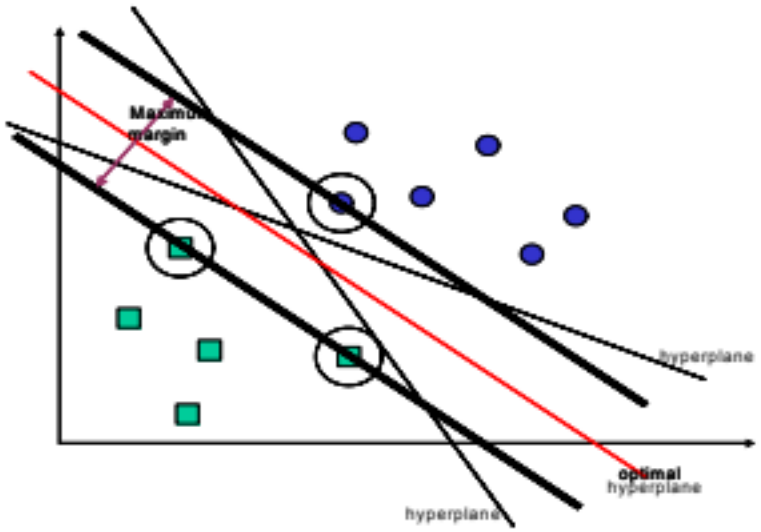

Figure 6: Maximum margin

Once we find the optimized hyperplane, the points on the margin is defined as support vector and the rest of the points can be ignored since those points are not relevant to the decision rule. The complexity of the model does not depend on the number of features in the training set [19].

2.10 K-Nearest Neighbours(kNN)

The kNN [20] is a one the oldest techniques that can be used to pattern classification. It can produce competitive results if someone could analyse and smartly mixed previous knowledge. It can be used to label unlabelled data records with majority appears in the k nearest neighbours in the training dataset. Performance of this algorithm largely depends on the distance matrix which use to identify nearest neighbours [21]. kNN should have example dataset, which is the training set, all the data should be labelled. That means we knows all the clusters that data should fall into. When the kNN is receive new data piece, it matches with every piece of existing data. Then it looks in to the most similar piece of data which are the nearest neighbours and look in to their labels. Then it looks at the top k most similar pieces of data from the known dataset. Finally, depending on the majority of votes for the k-most similar piece of data it will decide the new class.

2.11 Decision Trees

Decision trees are machine learning technique where the decisions can be visualized easily in a tree structure to understand for humans. It uses the data set which contains classification attributes and the set of class attributes to assign the data in to a certain classes. The data set is iteratively spitted according to the attributes until it reached its stopping criteria. Finding a decision tree is a NP-complete problem and researchers have worked to find the heuristics to get nearly - optimal decision tree. Therefore, each method should apply on the data set and the same procedure should apply on the sub trees as well until the training data set is classified into same classes [50].

Iterative Dichotomiser[51] 3 known as (ID3) and its successor C4.5 algorithms were used as the first algorithms for decision tree training and later those were the basis algorithms for further developments. Decision trees contain set of nodes called root nodes, inner nodes and leaf nodes where root nodes do not have any incoming edges, inner nodes have one incoming edges with one or more outgoing edges and leaf nodes have only incoming edges. Each node in the tree is a feature and the branch is represented a value which can be obtained by a node and also the edge can be interpreted as a decision made from

previous node. In a decision tree, the starting point is a root node and the rest of the nodes are presented in a sorted manner.

ex: Do the customer buy a pencil case from the shop while he is buying pencils? This question is called a decision problem where the answer is lying in leaf nodes. So that , the class predictions can be identified as ‘yes’ and ‘no’

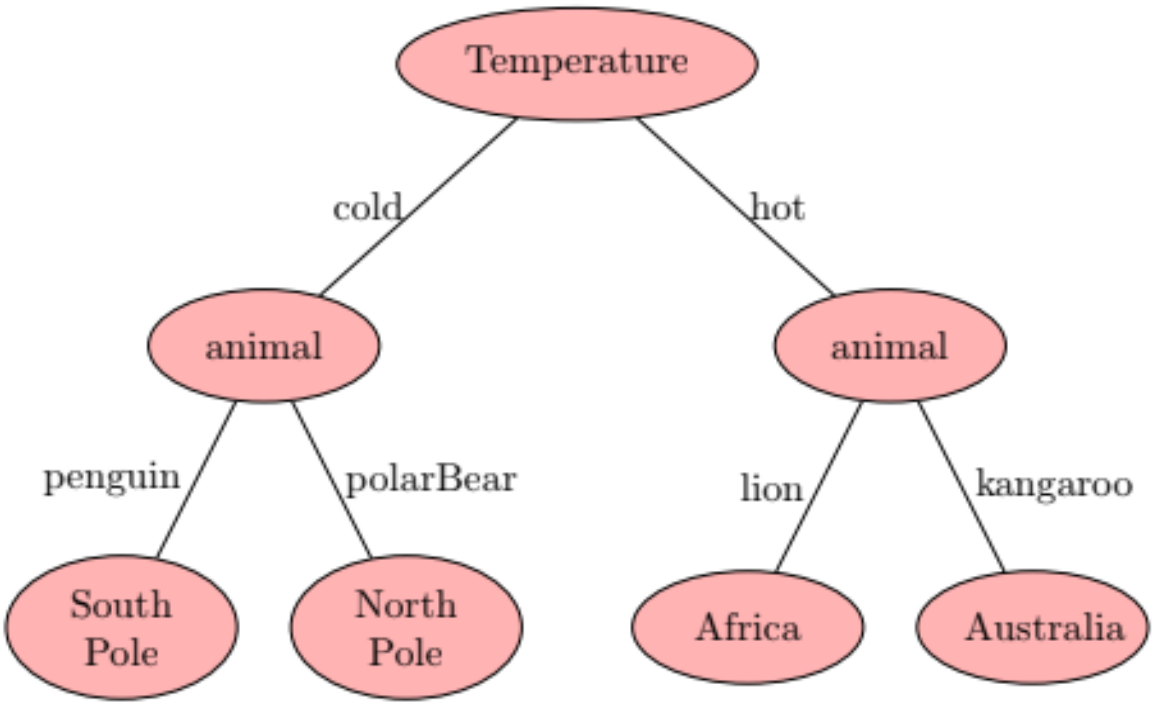

Figure 7: Decision tree example

Above is the example for a decision tree, for the instance where having the attribute cold and penguin is going down onto left subtree and being classified to “South Pole” with the corresponding label.

Training a decision tree is used as supervised learning technique where we use to predict the value of the target attribute value by looking at the decision tree which is built according to the data set patterns. Afterwards, it helps to predict the target attribute values of an unseen scenarios. To train a decision tree, we need set of prerequisites such as a data set which has a target attribute and input attributes, split criteria and stopping criteria. Then the process to find the best split should be recursively done through all the sub trees until it reaches the stopping criteria.

Common stopping criteria can be identified as below [49].

1. Maximum height has been reached (depth of a node has been exceeded the pre-specified limit)
2. Per node, there is no enough records (according to the pre-defined limit)
3. No rule can be obtained since all prediction values are identical.

Sometimes, we encounter the situation where the decision tree is overfitted for the training set too tightly. Therefore, on other hand it might be a disadvantage since we are unable to make a decision for unseen data. Later to overcome such kind of situation, the concept called pruning has been applied for the data set where it removes the non- productive, noisy or erroneous data.

2.12 Random Forrest (RF)

Random forest classification has high performance in classification. It is quite different from other learning methods since it uses ensemble learning [22]. RF uses multiple independent decision trees and provide composite prediction as the final result. Once the test data is feed in to the RF, all decision trees makes its own classification decision. The final classification class will be decided based on the majority of votes received for each tree [23]. RF helps to overcome noise and performance issues compare to single tree based models since its variance is very low.

2.13 Scikit-learn

The implementation for this thesis has been implemented using python scikit-learn library. It provides very detailed data science API with many options.

2.14 Evaluation of machine learning

In machine learning models it is very important how a computer program choose which results are fitting which contains more errors. In general evaluation of a machine learning model has been done by splitting the dataset in two sets as training set and test set. The training set has been used to train the model and test set has been used to validate the model. The main issue of this method occurs when the algorithm has limited access to the actual data where it can lead to the overfitting problem. However, the following are some of the widely used performance measures of classification machine learning models.

Miss classification rate

This represents the rate of incorrectly classified data in a dataset. Miss classification can be defined as following equation, where data point i and y_i are actual values and ý is the predicted value.

$$misc_n = \frac{1}{n} * \sum (y_i \neq \acute{y})$$

The problem of misclassification is when the number of classes is increases expected misclassification goes high. Benchmarking has been used to avoid this issue.

Benchmarking

In the context of machine learning benchmarking is the comparing the outcome of a model with another value which already been predicted using some other model. Then the misclassification can be compared with respect to a benchmark model.

The precision values

This method also known as positive prediction value, which can be described as amount of successful prediction for a particular class [27]. As an example, if there are two labels to represent won opportunity (W) and lost opportunity (L), precision 1 represents W labelled opportunity is actually won. This outcome has no effect on the amount of opportunity labelled L actually win.

The confusion matrix

The confusion matrix is one of the best approaches exists to demonstrate performance of a machine learning model, which differentiate concerning true positive, true negative, false positive and false negative predictions [27].

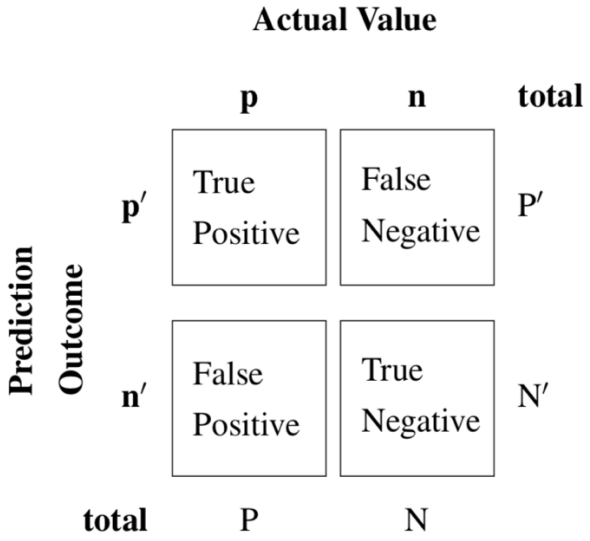

Figure 8: Confusion Matrix

In this chapter discussed the background in the context of machine learning and sales domain relevant to the study. Next chapter will be explained about the overall design of the proposed method.

Chapter 3

3. Design

The main objective of this chapter to discuss the design effort of the proposed method. The main objective of this model is to predict the outcome of a given opportunity, which are win or lose. This can be identified as a binary classification scenario. In this study, it will try different supervised machine learning algorithms and find out the performance of each algorithm.

3.1 Architecture

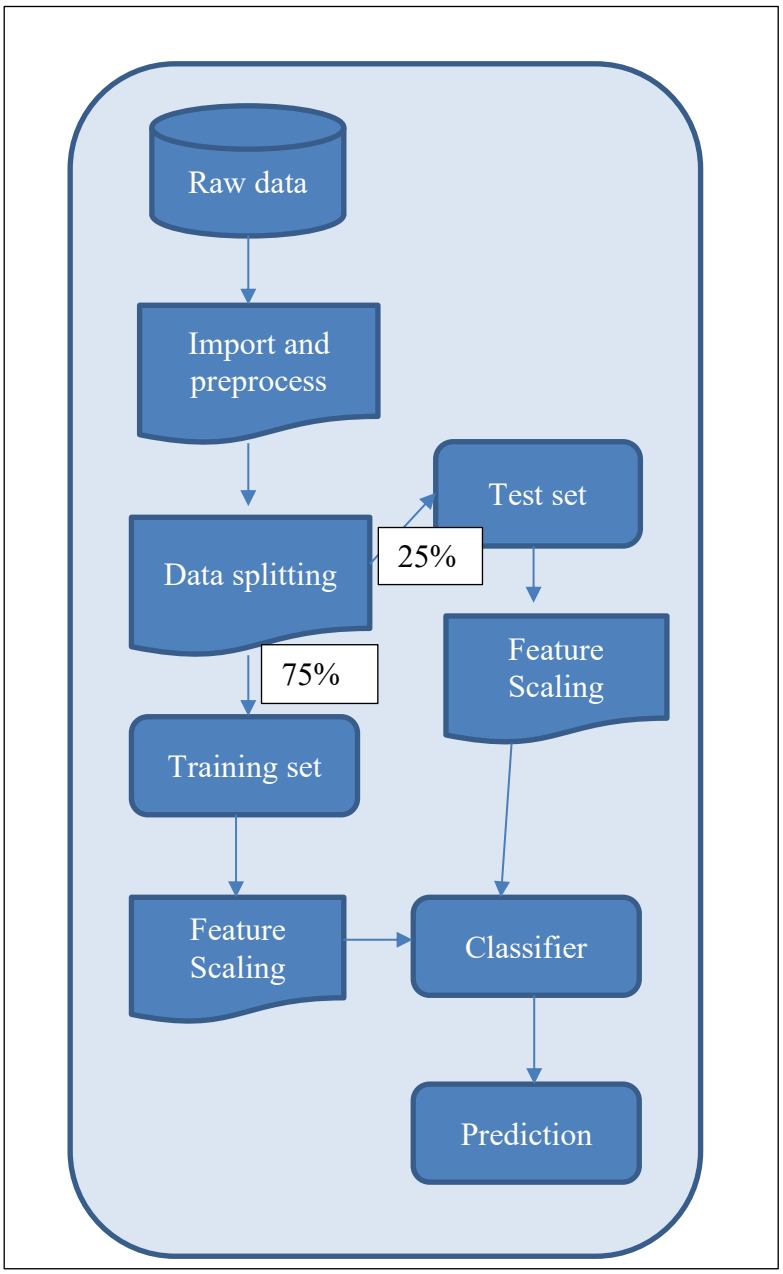

Figure 9: Overall Architecture of the Model

The figure shows the overall architecture of the proposed system. The design of this thesis is primarily consisted with selling to zebras captured historical dataset and supervise machine learning algorithms. The original data contains inside a csv file. That file first loaded in to a matrix. Pre-processing phase all the categorical variables will be converted to numerical values, which uses dummy encoding [28]. Then the data spitted in to training set and test set, where 25% goes to test set and 75% goes to train set. Then both the datasets have been gone through a feature scaling to normalize the data into a standard range. Throughout the study it uses different supervised machine learning classifiers, which will be plugged in to the classifier phase. Finally, the test data pump into the classifier to get the final prediction. Then the overall performance can be evaluated by analysing each algorithm's output.

In this chapter we have discussed about overall architecture of the system. In the next chapter we will discuss more about the experimental setup introduced to implement this design.

Chapter 4

4. Experimental Setup

This chapter mainly focuses on the characteristic of the dataset and implementation of the architecture described in the above chapter. All the algorithms related to this implementation have been discussed in the Chapter 2. All the tools and APIs we have used here are publicly available and free.

4.1 Dataset

This research is based on the sales data captured in Selling to Zebras Inc for the sales cycle from 2013 January to 2013 December. The dataset consists with 448 sales opportunities and for the confidentiality purpose customer names have been masked. Originally this data captured in different forms, emails and mainly in the selling to zebras database. Those information have been composed exported into single csv file for this experimental purpose. This section gives a in depth analysis on the dataset in the context of selling to zebras methodology and sales domain. Statistical distribution of the dataset also been included to this section.

Table summarizes the fields which captured for the experiment. The below fields have been identified as most important fields in selling to zebras.

Table 2: Fields in the selling to zebras dataset

Field Name	**Description**	**Values**
Access_to_Power	Indicate decision making power of the contact person of the organization	Low, Mid, High
Size	Indicate size of the company	Small, Mid, Big
Competitors	Indicate whether the opportunity has any competitors	Yes, No, Unknown
Is_Partner	Indicate whether the opportunity is a partner or not	Yes, No
Is_Budget_Allocated	Indicate whether the opportunity has enough budget allocation or not	Yes, No, Unknown
RFI	Indicate whether the opportunity requested any information or not	Yes, No

RFP	Indicate whether the opportunity requested any proposals	Yes, No
Com_Growth	Growing state of the company according to the current market trends.	Growth, Slowdown, Stable, Unknown
Positivity	Indicate whether the client's positivity on the deal	Yes, No, Unknown
Client_Type	Indicate the client type	Past, New, Current
Scope	Indicate whether the scope is clearly defined or not	Clear, Few Questions, Low
Is_Important	Indicate the importance of the client	Very important, Unimportant, Average important
Deal_Type	Type of the deal	Project, Maintenance, Solution
Clients_Clarity	Indicate whether client is clear about the requirement	Yes, No, Poor, Info Gathering
Attention	Attention to the client	First client, Strategic account, Normal, Bad Client
Status	Final outcome of the opportunity	Won, Lost

Below pie chart graphically represents the final outcome distribution of the sales opportunities. The total number of opportunities count is 448 where 227 have won and 221 have lost.

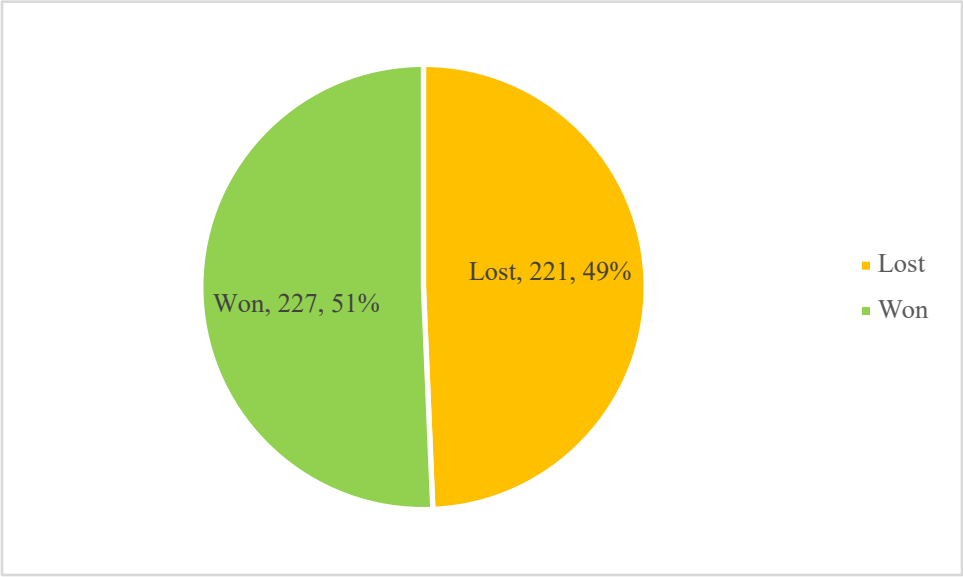

Figure 10: Opportunity distribution based on final outcome

The following figure 11 graphically represents the sales opportunity distribution based on access to power with the final outcome. In selling to zebras' methodology, sales teams always pursue opportunities which involves a person in the customer end, who has the power to make decision of buying. Here it marked as High, Mid and Low based on the contact persons' authority. Usually higher management likes

of CEO are very difficult to reach. That's why in the below table there are less record under High. As a practice selling to zebras sales team do not pursue opportunities which has 'Low' as the access to power. Thus, there are less records in Low category as well. Most of the active zone in this distribution is the Mid category.

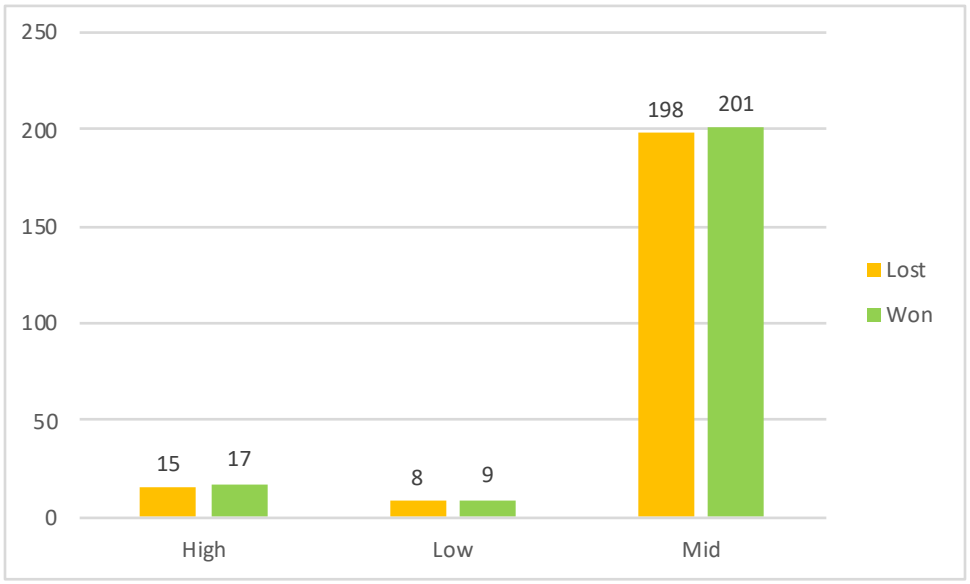

Figure 11: Opportunity distribution of access to power with the outcome

The following figure 12 graphically represents the sales opportunity distribution based on company size with the final outcome. Company type has been decided by annual revenue and number of employees in the company. Most of the time selling to zebras' sales teams were trying to close deals in organizations, which has big and mid-range of revenues. Only 10.71% of the deals in the sales pipeline are belong to the organizations are small size organizations. Sales teams are most active in big size companies. 58.93% sales opportunities belong to the big size companies where 30.36% are belong to medium size companies.

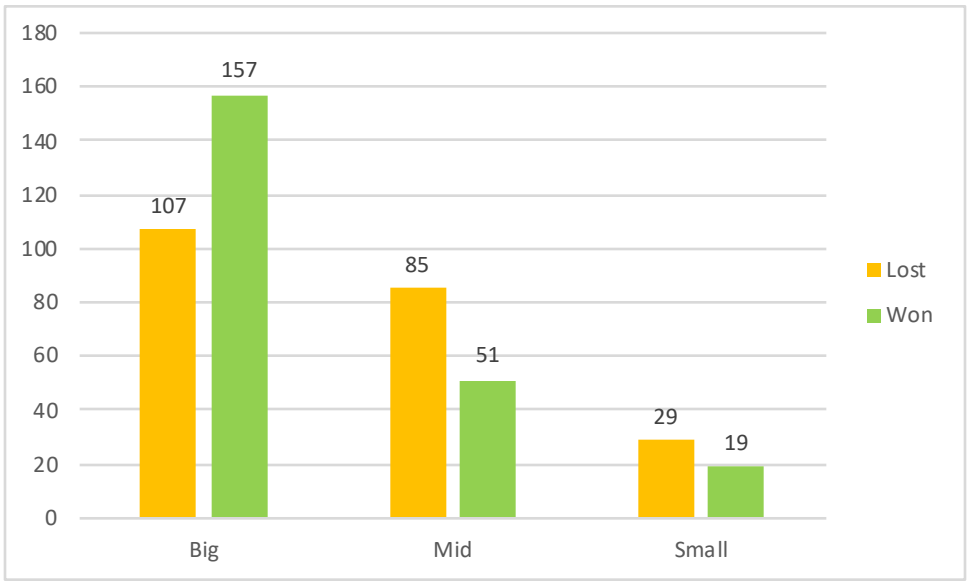

Figure 12: Opportunity distribution of company size with the outcome

The following Figure 13 graphically represents the sales opportunity distribution based on competitors' availability. It is confirmative when there are no competitors' sales is easy. Most of the deals have been won when competitors are none.

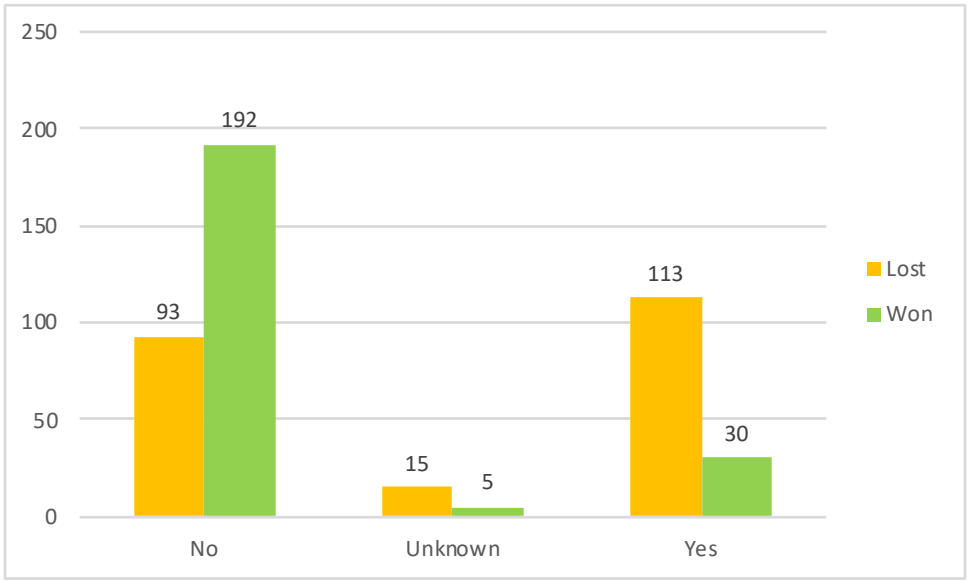

Figure 13: Opportunity distribution of competitors' availability with the outcome

The following Figure 14 graphically represents the sales opportunity distribution based on budget allocation with the final outcome. 54% of the cases have won when the budget is allocated where 53% of cases have won when the budget is not allocated.

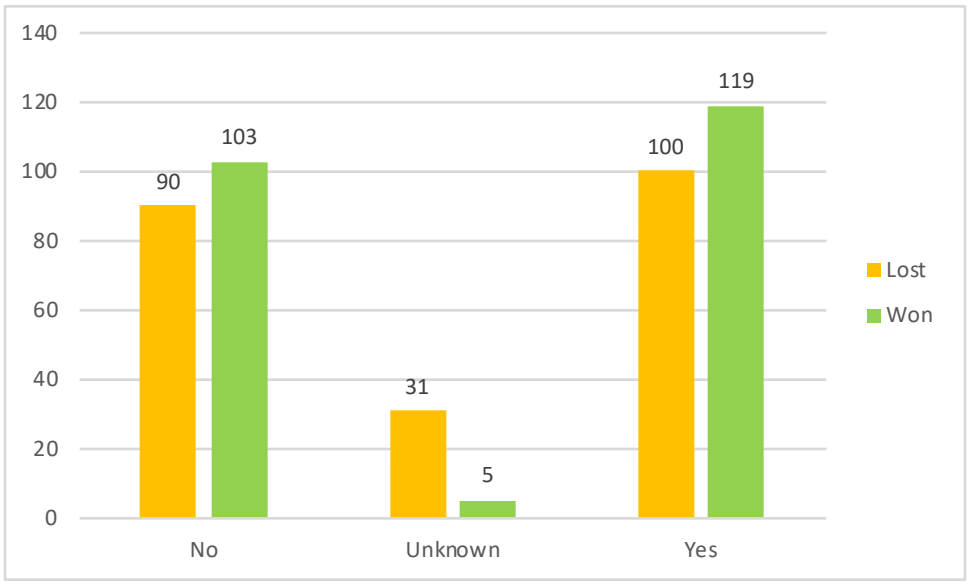

Figure 14: Opportunity distribution of budget allocation status with the outcome

The following Figure 15 graphically represents the sales opportunity distribution of RFI with the final outcome. It is evident from the chart more sales won when the client request for information.

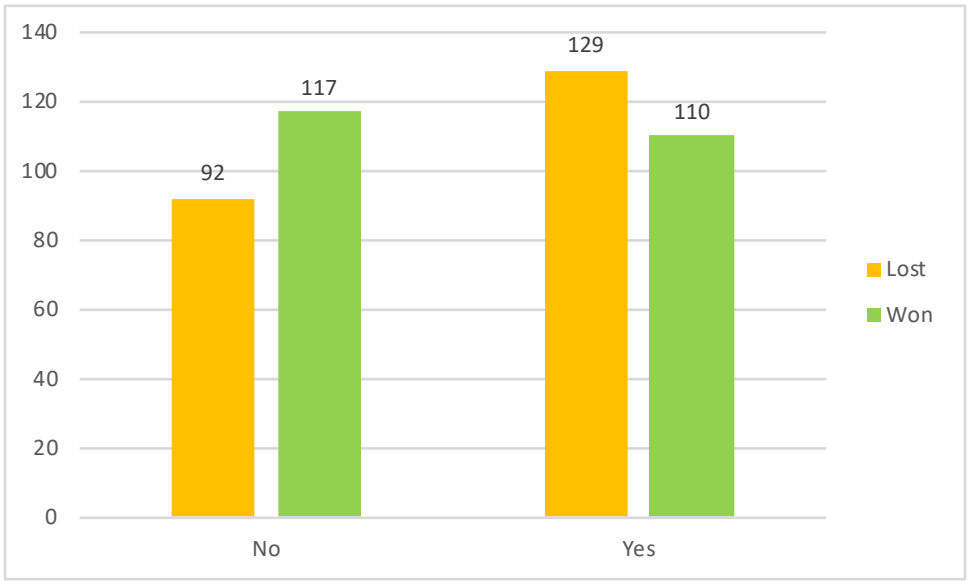

Figure 15: Opportunity distribution of RFI with the outcome

The following Figure 16 graphically represents the sales opportunity distribution based on RFP with the final outcome. The chart is confirming most of the time client request a proposal during the sales life cycle.

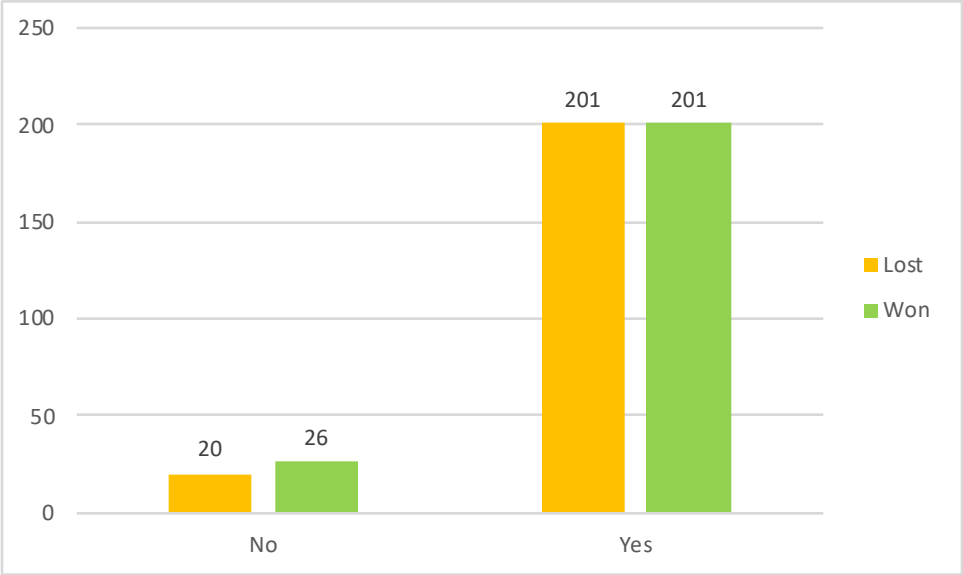

Figure 16: Opportunity distribution of RFP with the outcome

The following Figure 17 graphically represents the sales opportunity distribution based on company growth with the outcome. Usually any sales person try to sell their product to stable companies. The chart is evidently supporting that, by showing 403 opportunities belong to the stable companies.

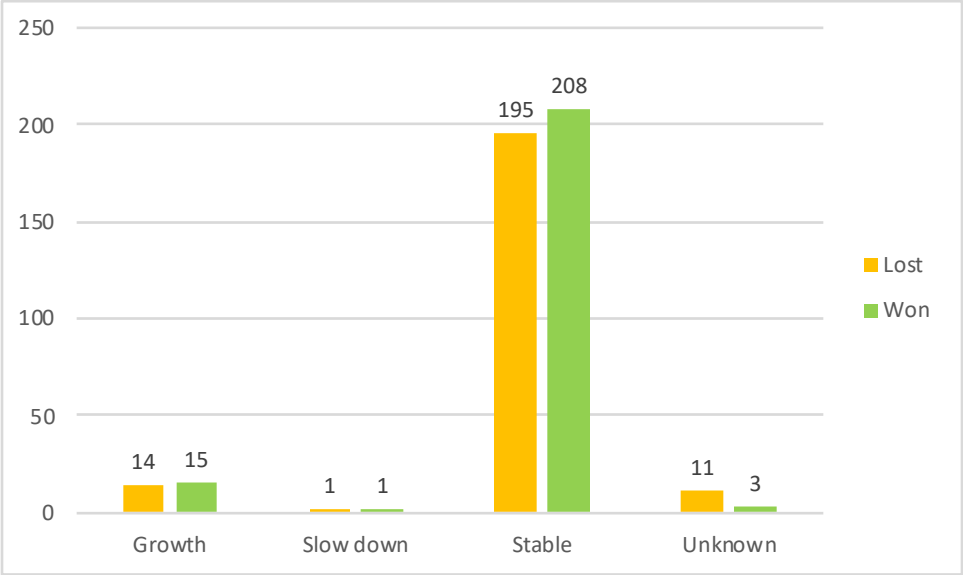

Figure 17: Opportunity distribution of company growth with the outcome

Figure 18 represents opportunities distribution of positivity of client with the final outcome of the opportunity. It shows that most of the time clients have been neutral rather than displaying positive attitude towards a product.

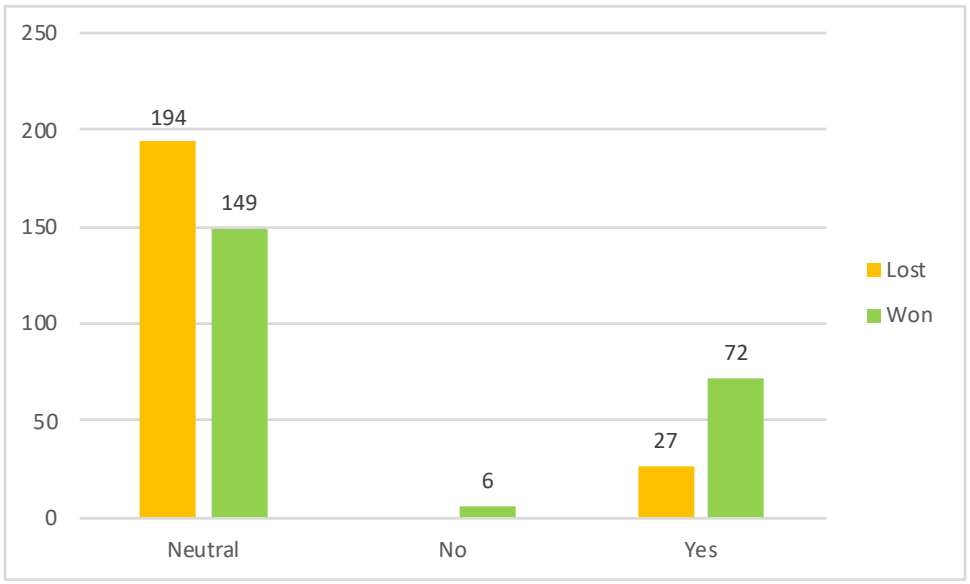

Figure 18: Opportunities distribution of positivity of client with the outcome

Figure 19 represents the opportunity distribution of client type with the final outcome. The chart demonstrates that, most of the wining deals are belong to the current customers category where most of the losing deals are belong to new customer category.

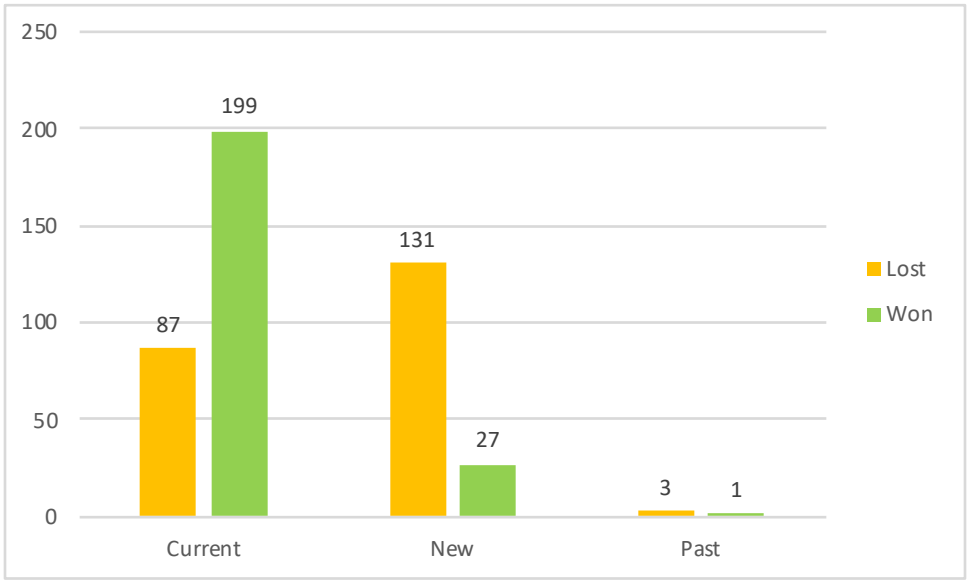

Figure 19: Opportunity distribution of client type with the outcome

Figure 20 shows the distribution of clearness of the scope with outcome. The chart demonstrates almost all the time sales teams are understand the scope clearly during the sales cycle.

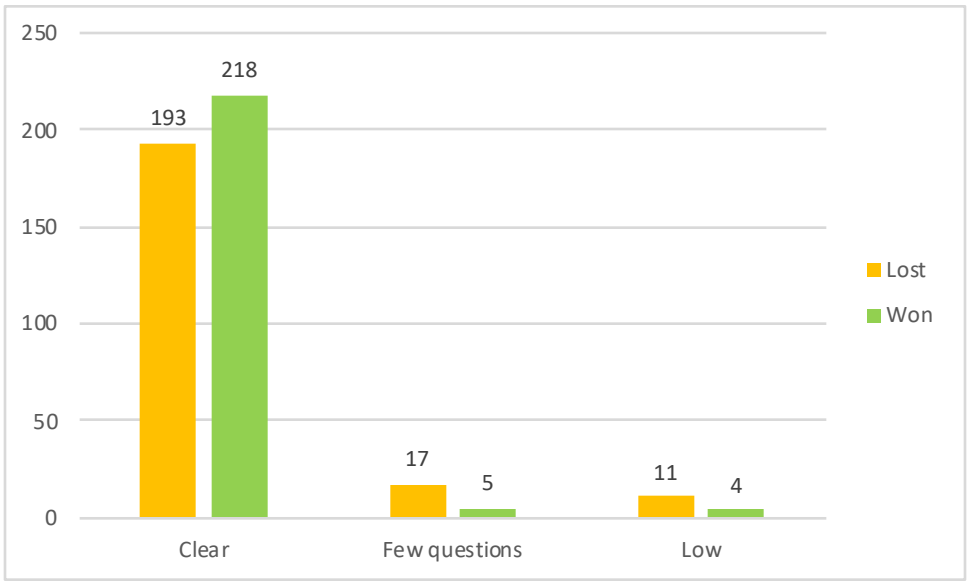

Figure 20: Opportunity distribution of clearness of the scope with final outcome

The figure 21 shows how selling to zebras sales teams assign importance to their customers and final outcome. Most of the time opportunities have been considered as average important. However, it is evident it shows higher change of winning rate when a deal mark as very important.

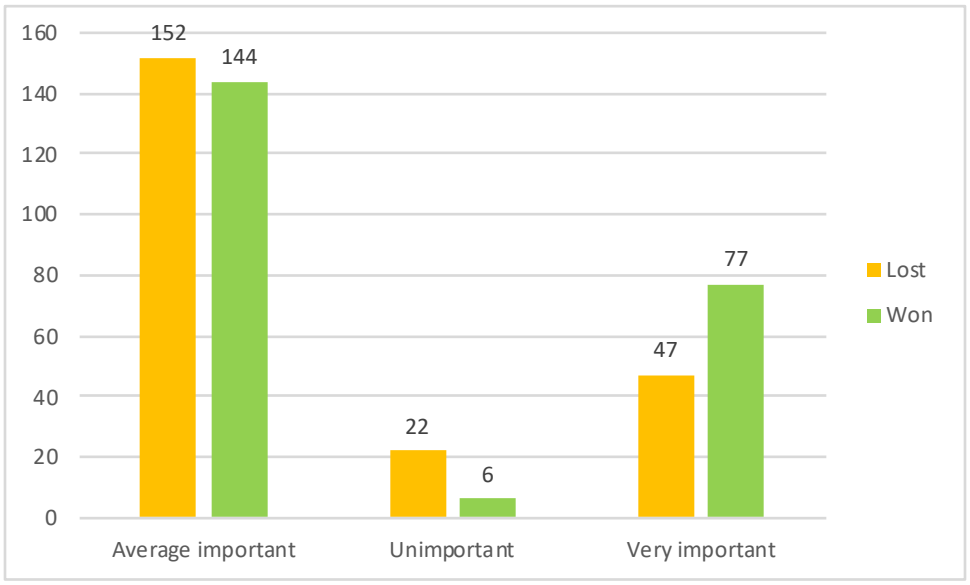

Figure 21: Opportunity distribution of importance with final outcome

Figure 22 shows opportunity distribution of client's understanding about the solution and outcome. Chart demonstrate most of the time client has been in a good understanding before completing the sales cycle.

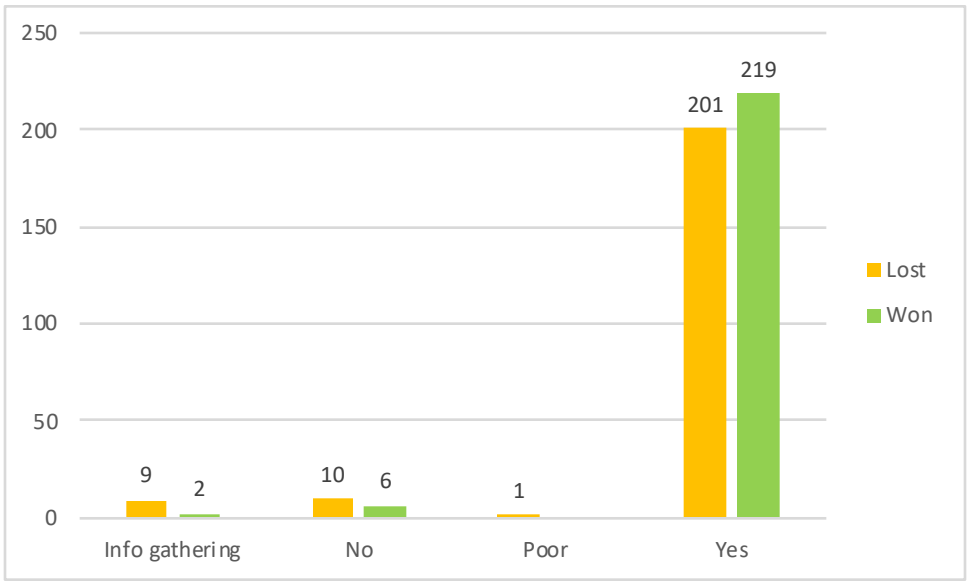

Figure 22: Opportunity distribution of client's understanding with final outcome

Figure 23 shows opportunity distribution of the attention to the deals with final outcome. Usually, sales teams don't treat all their opportunities in the same way. Here it is evident whenever sales teams have normal or strategic level attention to deals, the winning rate is high.

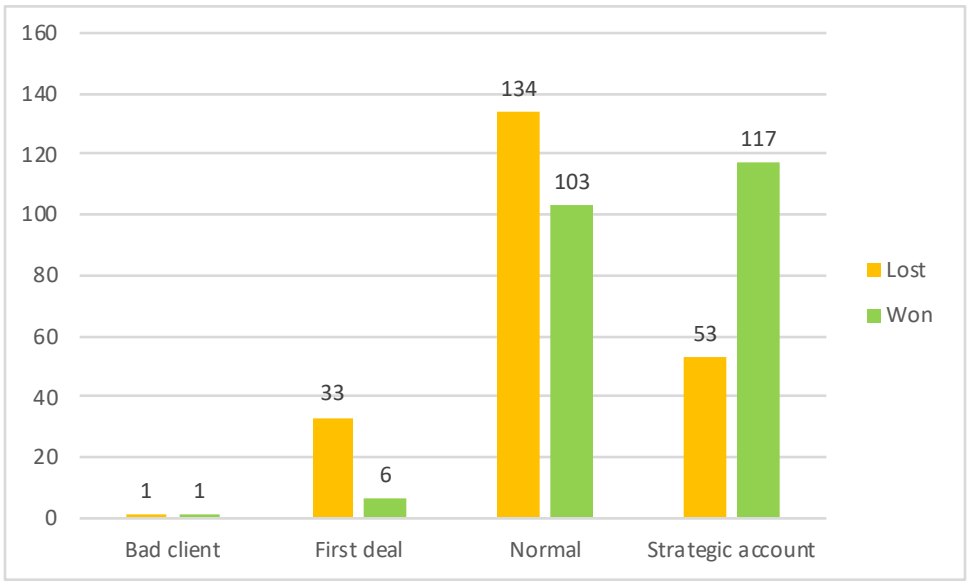

Figure 23: Opportunity distribution of attention to the customer with final outcome

In this section we have discussed about the data set of selling to zebras an its distribution. In order to identify the important features, statistical analysis is not enough. Therefore, we you a feature redundancy analysis to find out whether this contain any redundant features.

4.2 Feature redundancy analysis

Many researchers suggested that redundant features in a dataset could affect overall performance of a classification model [31, 32, 33]. If the same feature info appeared repeatedly, where single feature represents the same state indicates by another set of features identified as feature redundancy [37]. The original dataset contains 16 features as mentioned in the Table 2. Therefore, it is very important to find out whether it contains any redundant features.

Correlation matrix is one of the populate methods used in redundancy analysis [38]. In this study we used linear correlation coefficient which measure how two variables are tied together. The coefficient value can take values between -1 and 1. If two independent variables are strongly correlated when the coefficient value is -1 or 1, where 0 means no correlation at all. High correlation indicates that two variables are highly dependent on one another. If the correlation is strong, rather than keeping both the features in the dataset, we can remove one feature.

Before generates a correlation matrix the data set needed to be pre-processed. In this study we have used Rapid Miner tool to perform the pre-processing. Since all the attributes are categorical attributes, it is required to replace those with numeric values. There are several ways to encode categorical variable to numbered values. In this experiment we are encoding method call One Hot Coding.

One Hot Coding is the most widely used coding scheme [28]. In this method, If the variable has n number of possibilities a value will be represent in n number of binary variables. As an example, in this study we have a categorical variable call “Access_to_Power” which has values “Low”, “Mid” and “High”, Which can represent in 3 variables. Those are “Access_to_Power=Low”, “Access_to_Power=Mid” and “Access_to_Power=High”. If one instance has “Access_to_Power=Low” as 1 in other two would be zero filled.

Figure 24 represents the correlation matrix generation process for the given dataset.

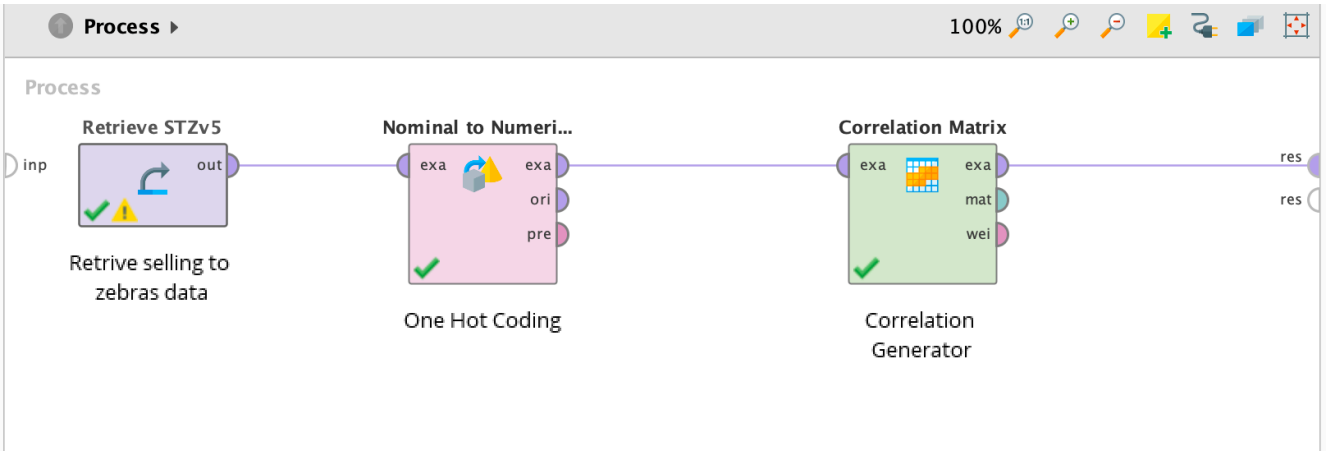

Figure 24: RapidMiner process for the correlation matrix

This process will output the following figure 25.

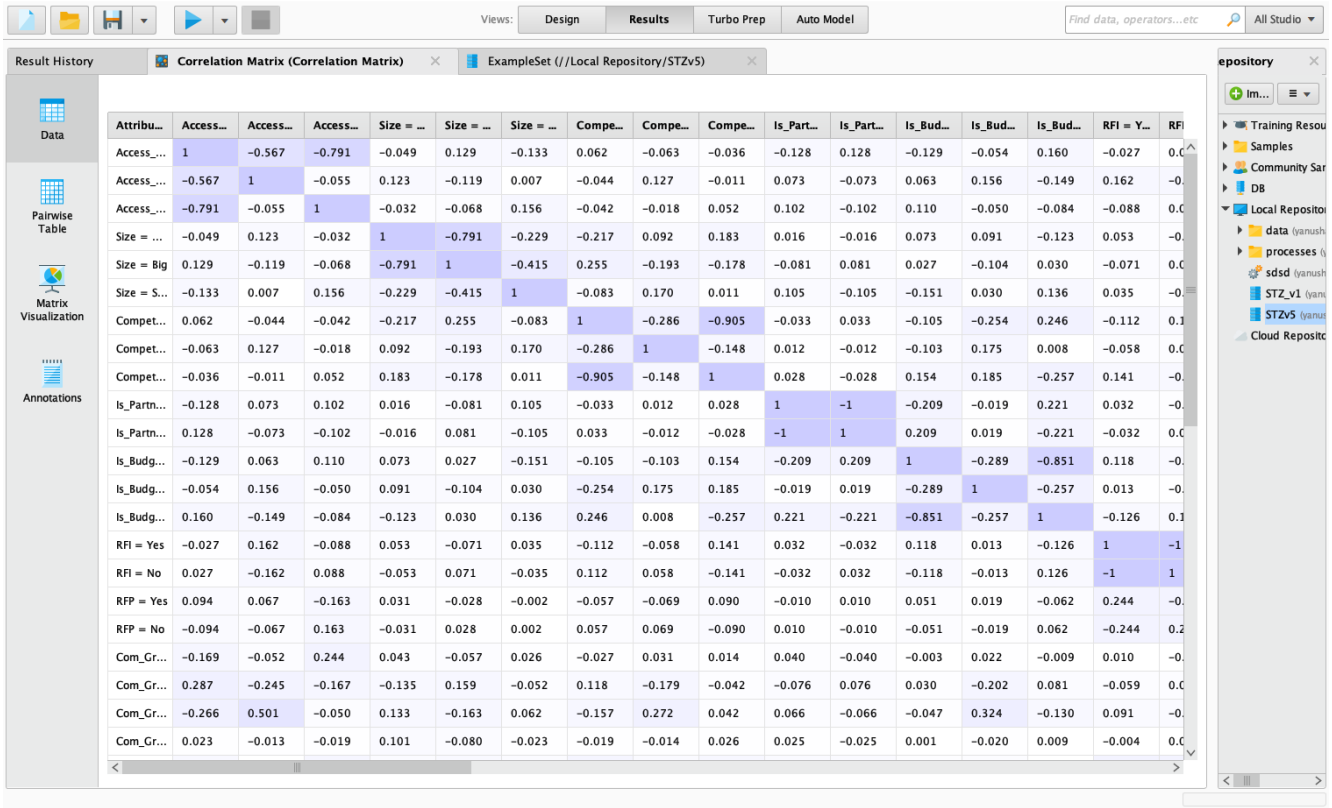

Figure 25: Correlation Matrix

After analysing the coefficient values following table illustrate the strongest relations between features. Here we have chosen the absolute values of correlation which greater than 0.75.

Table 3: Top correlation values of the dataset

First Attribute	Second Attribute	Correlation
Is_Partner = No	Is_Partner = Yes	-1
RFI = Yes	RFI = No	-1
RFP = Yes	RFP = No	-1
Status = Won	Status = Lost	-1
Client_Type = Current	Client_Type = New	-0.981
Positivity = Neutral	Positivity = Yes	-0.963
Competitors = No	Competitors = Yes	-0.905
Is_Important = Very Important	Is_Important = Average Important	-0.863
Is_Budget_Allocated = Yes	Is_Budget_Allocated = No	-0.851
Deal_Type = Project	Deal_Type = Solution	-0.830
Attention = Strategic account	Attention = Normal	-0.829
Access_to_Power = Mid	Access_to_Power = High	-0.791
Size = Mid	Size = Big	-0.791
Com_Growth = Growth	Com_Growth = Stable	-0.787
Scope = Clear	Scope = Few Question	-0.757
Scope= Clear	Clients_Clarity = Yes	0.794
Scope = Low	Clients_Clarity = No	0.767

The table 3 shows top correlation values of the dataset. First four values show strong correlation because those categorical variables are binary values. Therefore, we can keep only one field to show those. From our pre-processed data set we have removed one from each above-mentioned features pair. Table 4 illustrate the new features set which going to use in the experiment.

Table 4: List of columns after pre-processing

Access_to_Power = Low	Access_to_Power = High	Size = Big	Size = Small
Competitors = Unknown	Competitors = Yes	Is_Partner = Yes	Is_Budget_Allocated = Yes
Is_Budget_Allocated = Unknown	RFI = Yes	RFP = Yes	Com_Growth = Stable
Com_Growth = Unknown	Com_Growth = Slow down	Positivity = Neutral	Positivity = No
Client_Type = Current	Client_Type = Past	Scope = Clear	Scope = Low
Is_Important = Very important	Is_Important = Unimportant	Deal_Type = Project	Deal_Type = Maintenance
Deal_Type = Consulting	Clients_Clarity = Info gathering	Clients_Clarity = Poor	Attention = Strategic account
Attention = First deal	Attention = Bad client	Status = Won	

4.3 Tools and Programming language

This section provides brief summary on the tools and programming languages that have been used for the proposed method. Python is the programming language we have used in this project where Anaconda navigator is the primary tool. The Spyder has been used on top of the Anaconda to write python codes.

Anaconda navigator

Anaconda navigator is a desktop GUI application which supports applications to launch with Anaconda managed packages, environments and channels without the need of execution of command line tools.

Spyder

Spyder is a python IDE which specially introduced for the scientific python development.

Python

Python is a high-level programming language widely used in data science context, was developed by Guido van Rossum from National Research Institute for Mathematics and Computer Science in Netherlands.

4.4 Evaluation Procedure

The cross validation has been used for the evaluation of algorithms in this study. As we discussed in the Chapter 2, the programs use scikit-learn API to invoke machine learning algorithm. The scikit-learn contains helper function call train_test_split which helps the dataset to be spitted in to two parts as train set and test set. If a system train and verify using a same dataset, it will give perfect prediction for input which system already known. But in practical scenario it couldn't predict anything useful on newly introduced data. This situation called overfitting. Using cross validation method, it can avoid the overfitting. Following flow chart demonstrates the model training process of cross validation.

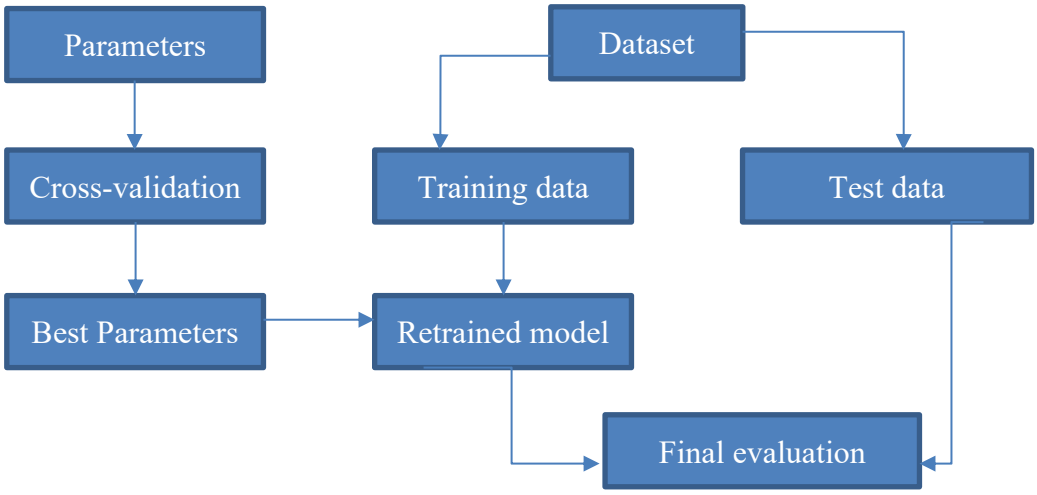

Figure 24: cross-validation model training

The main disadvantage of evaluating a model using a train set and test set is it can give different accuracy for a different test set due to variance of the data. Thus, judging a model performance based on a one test is not that static and strong. Then this is not the most relevant way of evaluating the performance of a model. There is a technique called k-field cross validation [30] which helps to overcome this variance issue by spinning the training set into k equal size subsets. One subset keeps as the validation set, where other k-1 used as training sets. The test data set can be kept for the final evaluation. Finally, average of the all estimated accuracies will be taking as the final estimate.

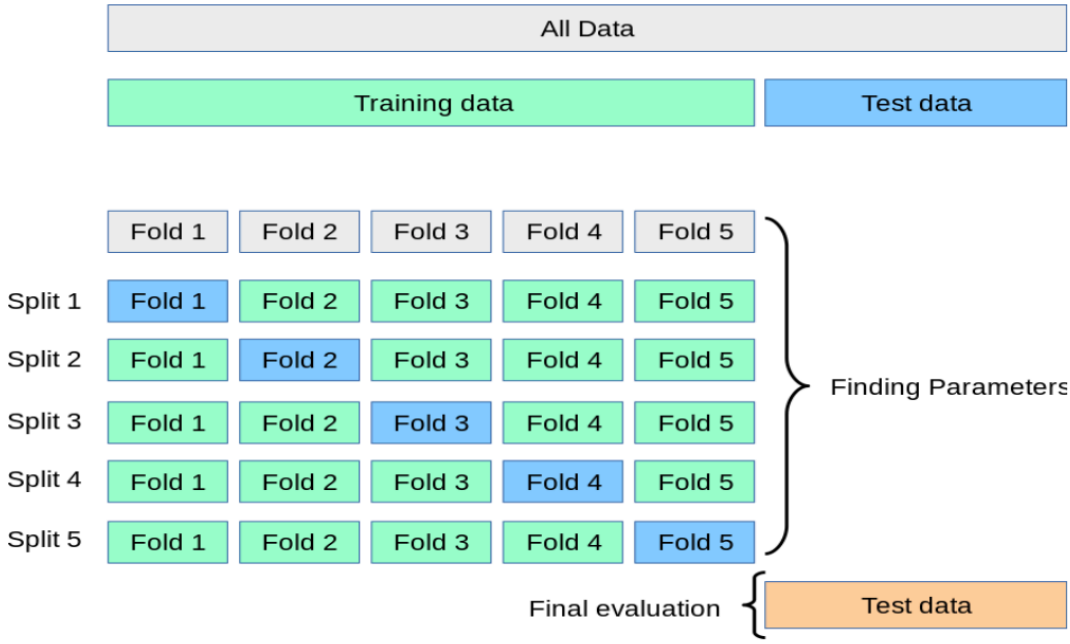

Figure 25: k-fold cross validation process

Figure 19 shows the process of k-fold cross validation process. In this study, the k value has been set to 10 which is most common and proven to be the best choice for the k [30].

Evaluation matrix

The study used a classification report for each algorithm to measure the quality of classification algorithms. This classification report includes Precision, Recall and F1 score. As it discussed in the chapter 2 confusion matrix, there can be four ways to check the prediction is correct or not. Which are,

- True Negative (TN): Actual value was negative and predicted negative
- True Positive (TP): Actual value was positive and predicted positive
- False Negative (FN): Actual value was positive but predicted negative
- False Positive (FP): Actual value was negative but predicted positive

The Accuracy, Precision, Recall and F1 score have been measured using above four attributes.

Accuracy

The Accuracy is the overall accurate prediction a given the model, which can be calculated using following equation,

$$Accuracy = TP + TN \;/\; (TP + TN + FN + FP)$$

Precision

The Precision is the accuracy of positive predictions, which can be calculated using following equation,

$$Precision = TP/(TP + FP)$$

Recall

The Recall is the ability of finding all positive instances, which can be calculated using following equation,

$$Recall = TP \;/\; (TP \; + FN)$$

F1 Score

The F1 score is the weighted harmonic mean of above mentioned two measures. The best value can F1 can have is 1 and worst is 0. The F1 Score equation has been given below and it is recommended to use weighted average of F1 for the comparison.

$$F1\ Score = 2 * (Recall * Precision)/\ (Recall + Precision)$$

In the program we have used classification report function in the sckit-learn API to retrieve all three values.

4.5 Environment

This study has been carried out on a computer system with below mentioned performances

Processor	2.6 GHz Intel Core i7
RAM	16 GB 2400 MHz DDR4
OS	Mac OS 10.14.3

Figure 26: Performance matrix of computer system

This chapter introduced the dataset and implementation methods that have been used in the proposed method. Next chapter will evaluate the actual result using the methods which have been introduced in this chapter.

Chapter 5

5. Results and Analysis

In this chapter will outline the results after execution of the proposed model. The used dataset, implementation techniques and evaluation methods have been explained in the previous chapter. Finally, we conclude the chapter with CAP curve analysis of each model.

5.1 Algorithm Evaluation

This section outlines the classification accuracy that has been achieved from each algorithm. This study experimented the classification prediction using five algorithms which are Logistic regression, Support Vector Machine, k-Nearest Neighbours, Decision tree and Random forests. The theoretical background of those algorithms has been discussed in chapter 2.

Parameter tuning

In this section briefly outlines parameter tuning, which required for the selected algorithms. It is very important to find out optimum parameters for each algorithm. SVM requires the optimal kernel type and Random Forrest algorithm requires the optimum number of trees. Hence, we change each parameter type and compare the mean squared error to find the applicable optimum parameter for this particular dataset.

In this study we have used C-Support Vector Machine, which requires to pass a kernel to classify a given dataset. Following table shows calculated mean squared errors calculated for linear, polynomial, sigmoid, radial basis function kernels. According to the table, the minimum mean squared error has given by sigmoid and radial basis function kernels. Thus, we have used radial basis function kernel since which is widely used compare to sigmoid.

Table 5: Kernel types and Mean Squared Error for SVM

Kernel Type	Mean Squared Error
linear	0.295
polynomial	0.491
sigmoid	0.241
radial basis function	0.241

Before creation of Random Forrest classifier, it is required to pass a value for the number of trees. Below table shows mean squared error return for different number of trees. According to the result we can see the minimum squared error return for the number of trees equals to 60. If it is less than 50 the error is increasing and from 60 to 100 all the error values are greater than 0.250.

Table 6: Number of trees and Mean Squared Error for RF

Number of trees	Mean Squared Error
5	0.304
10	0.277
20	0.277
30	0.268
50	0.259
60	0.250
70	0.268
80	0.259
90	0.268
100	0.268

kNN need algorithm need to pass number of neighbours as a parameter to the algorithm. Below Table 7 illustrated the mean squared error received for each neighbour's size. Here we can see the minimum error has given for neighbour size 10. Thus, in this study we have used number of neighbours as 10.

Table 7: Number of neighbours and Mean Squared Eroor for kNN

Number of neighbours	Mean Squared Error
5	0.304
10	0.286
15	0.295
20	0.295

5.2 Performance Evaluation

We have discussed performance evaluation methods in the Chapter 4 in detail. In summary, the performance evaluation matrix consists with precision, recall, f1-score and accuracy which measured by k-fold cross validation. Following table outlines the performance measures calculated for each algorithm.

Table 8: Evaluation matrix for the classification algorithms

Evaluation metric	Opportunity outcome	SVM	RF	Decision Tree	kNN	Logistic Regression
Accuracy		0.759	0.750	0.750	0.714	0.786
Precision	Won	0.730	0.770	0.800	0.704	0.785
	Lost	0.816	0.725	0.702	0.732	0.787
	Weighted Average	0.769	0.750	0.755	0.717	0.786
Recall	Won	0.885	0.770	0.784	0.820	0.836
	Lost	0.608	0.725	0.721	0.588	0.725
	Weighted Average	0.759	0.750	0.750	0.714	0.786
F1 Score	Won	0.800	0.770	0.759	0.758	0.810
	Lost	0.697	0.725	0.741	0.652	0.755
	Weighted Average	0.753	0.750	0.750	0.710	0.785

For a given instance original dataset have been divided into two parts 25% test data and 75% training data. Then, the data has been passed to different algorithm and calculated the above-mentioned performance matrix. These values have been extracted using sckit-learn classification report. It gives a summary of overall weighted averages of accuracy, precision, recall and f1 score.

The table 8 illustrate the prediction capability of each model. We can observe each model has different prediction accuracies with respect to the class. It is important to analyse which model has the better performance over predicting a specific class.

Table 9: Highest and Lowest performing models

Accuracy		Precision		Recall		F1 Score	
Highest	Lowest	Highest	Lowest	Highest	Lowest	Highest	Lowest
Logistic Regression	kNN	Logistic Regression	kNN	Logistic Regression	kNN	Logistic Regression	kNN

Table 10: Highest and Lowest performance based on the outcome

Outcome	Precision		Recall		F1 Score	
	Lowest	Highest	Lowest	Highest	Lowest	Highest
Won	kNN	Decision Tree	RF	SVM	kNN	Logistic Regression
Lost	Decision Tree	SVM	kNN	RF and Logistic Regression	kNN	Logistic Regression

Based on the above table 8 it is evident logistic regression has better performance where kNN has the lowest based on their accuracy and weighted average of precision, recall and F1-score. But if we closely analyse the values in the Table 8 we can see there is not much of a difference in other models as well.

The table 10 summaries the performance gained for each category. As we discussed in chapter 4 Precision, Recall and F1 Score all based on the confusion matrix. The model should be selected based on what kind of false we can tolerate between false positive and false negative. In our case false negative means, predicting winnable opportunity as a losing one and false positive means predicting losing opportunity as a winning one. Recall measure the performance based on the true positive values. Which means how correct is our prediction between won opportunities. Thus, minimizing false negative leads to maximize the recall value. Precision is needed to be consider when it can't tolerate the false positive value. That is when losing opportunity predict as a winning one. Here it is not that much of a big deal compare to the false negative scenario. The F1-Score can be considered as alternative to the accuracy. Which balances between accuracy and recall. Since the SVM has the highest recall value compare to the others which can be taken as the most suitable classification model for the given problem.

5.3 Cumulative Accuracy Profile (CAP) Curve Analysis

The cap curve analysis carried out to observe how quickly a prediction model can determine all its data points of a target class using minimum number of probes [52]. Here we are trying to find out how quickly the given models identify all 'Won' opportunities from the given dataset. Here for the implementation we are using python with sckit-learn APIs.

5.4 Random model

Random model illustrates a completely uninformative model, which has linear growth for won observations proportion to total observations.

5.5 Perfect model

Perfect model illustrates a model, which predict all the won instances of the dataset with number of probes is equalling to number of instances.

5.6 Cap curve Analysis

Cap curve is using area under the curve to determine the Accuracy Rate (AR) of a model. AR is given by area under the prediction model until the random model(aP) dividing by area under the perfect model until the random model (aR). The prediction model said to be really robust if the rate is closer to the 1.

$$AR = aR \,/\, aP$$

There is another way to analyse CAP curve by reading the plot. First it requires to draw a vertical line from the 50% of the x axis until it meets the corresponding model plot. From that point to y axis we draw a horizontal line. Finally, calculate the percentage of correct 'Won' cases identified from the total number of 'Won' cases. If the percentage value is less than 50% means the proposed model is a poor model, where 50% – 100% means the model is good model. However, if it is within 90% - 100%, model should be tested for overfitting [53].

Table 11: Cap curve analysis results

Model	AR	Pot analysis (Won %)
SVM	0.56	70.49
Random Forest	0.56	72.13
Decision Tree	0.73	73.77
Linear Regression	0.61	70.49
kNN	0.60	73.78

Table 11 illustrate the result extracted from the CAP curve analysis. According to the AR values we can see SVM and Rando Forest models have the lowest accuracy rate 0.56 where Decision Tree has the highest AR 0.73. Linear Regression and kNN has 0.61 and 0.60 AR respectively. If you see the figure 24 the decision tree classifier goes quite close to the perfect model where others don't have such consistent behaviour. Since all of these methods has more than 0.5 of AR, we can say all are good models. However, decision tree is comparatively stronger since it has the highest AR.

According to plot analysis result it is evident all model has been performed well in successful predicting criteria. After observing 50% of the given test data it has predicted more than 70% of the total successful cases. Then all these models are good models.

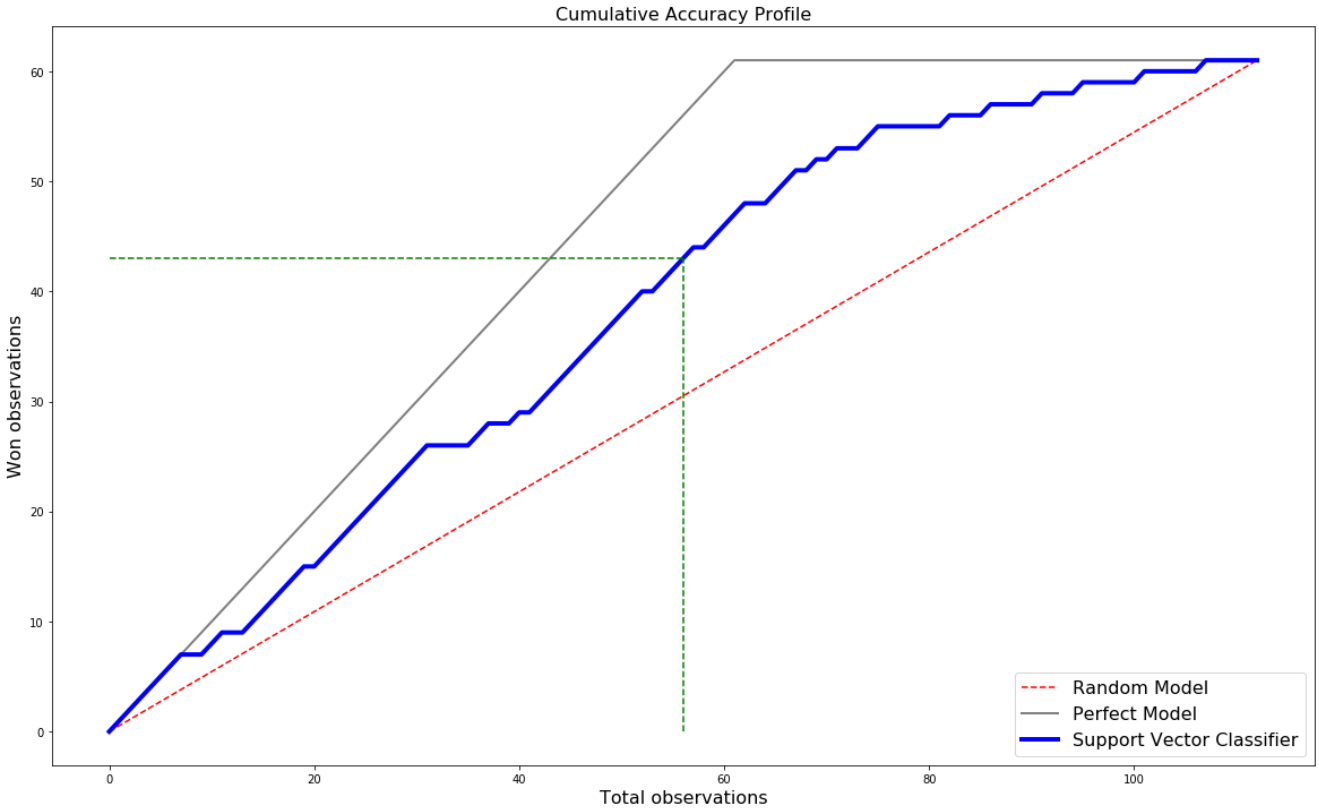

Figure 27: CAP curve for SVM model

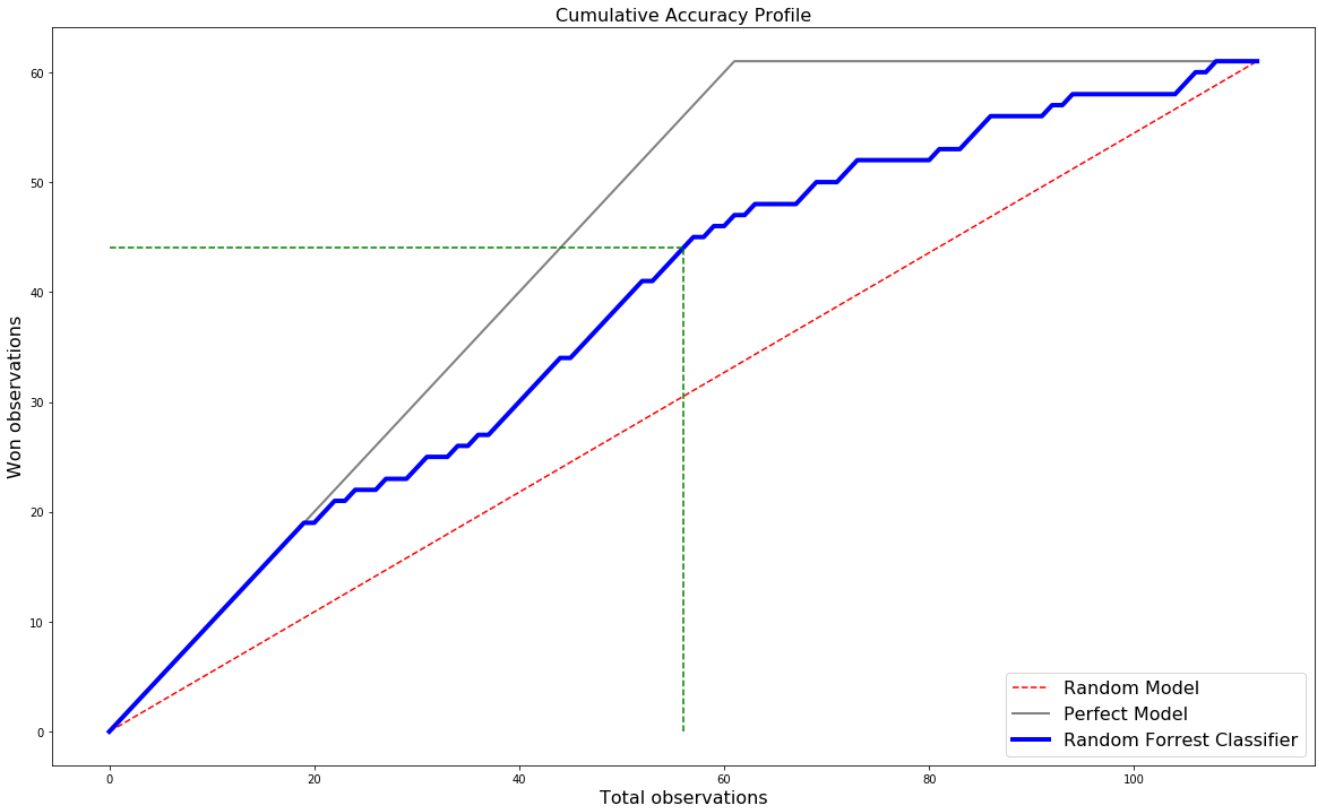

Figure 28: CAP curve of Random Forest model

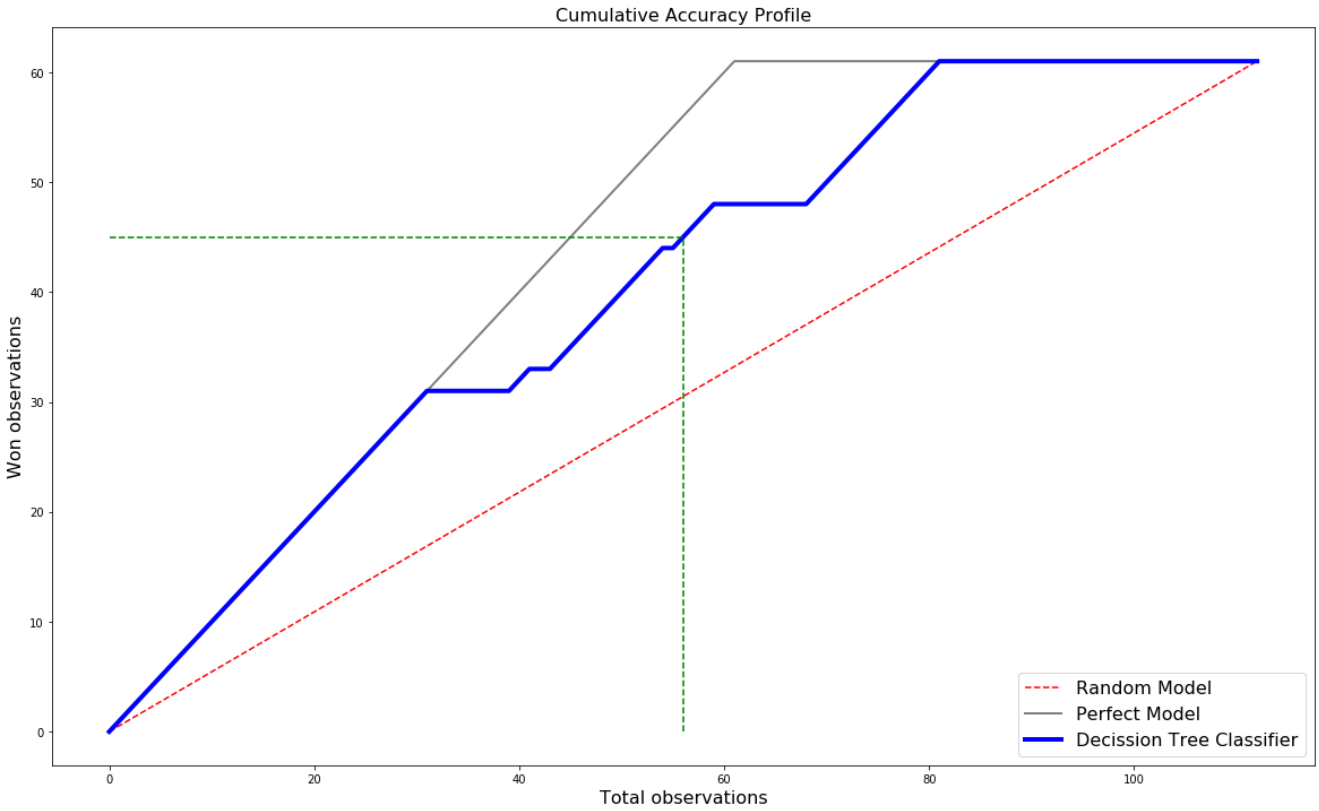

Figure 29: CAP curve of Dicession tree classifer

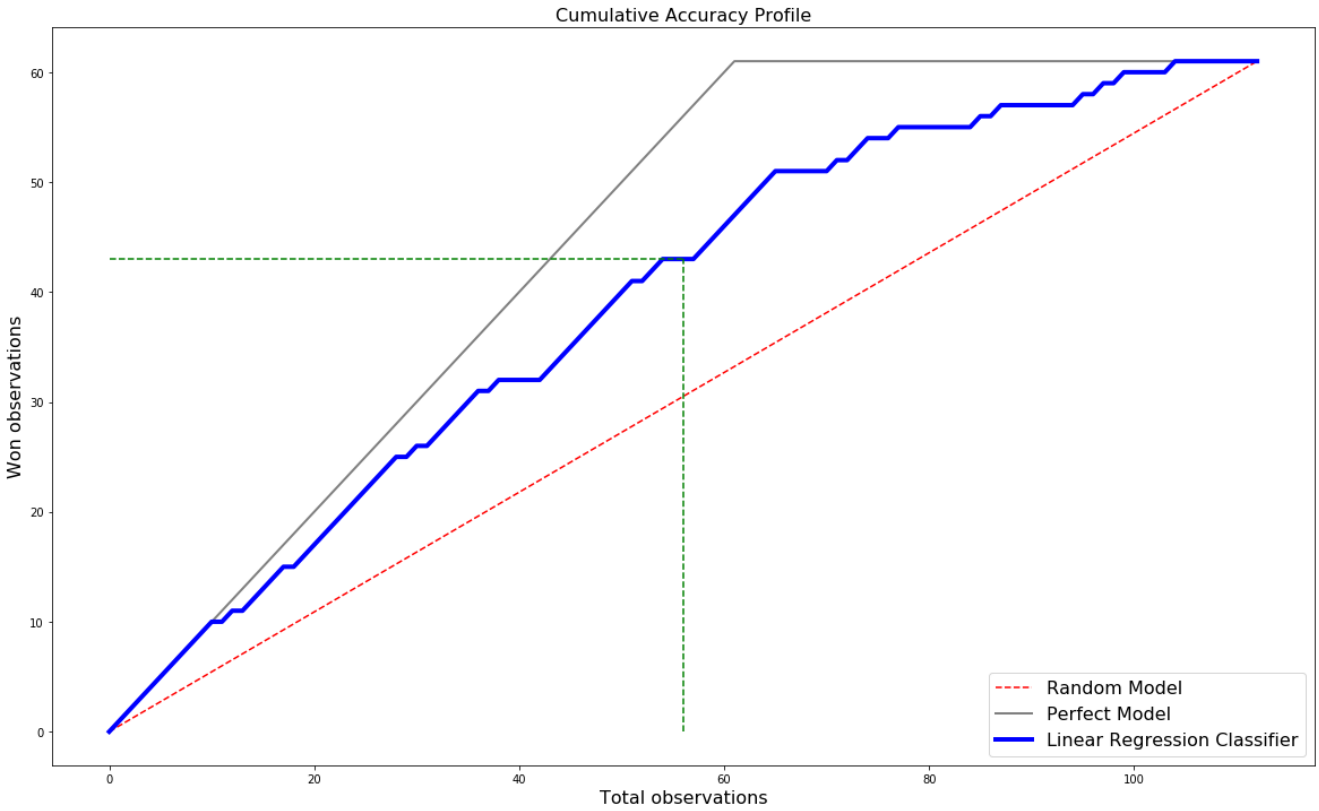

Figure 30: CAP curve of Linear Regression classifier

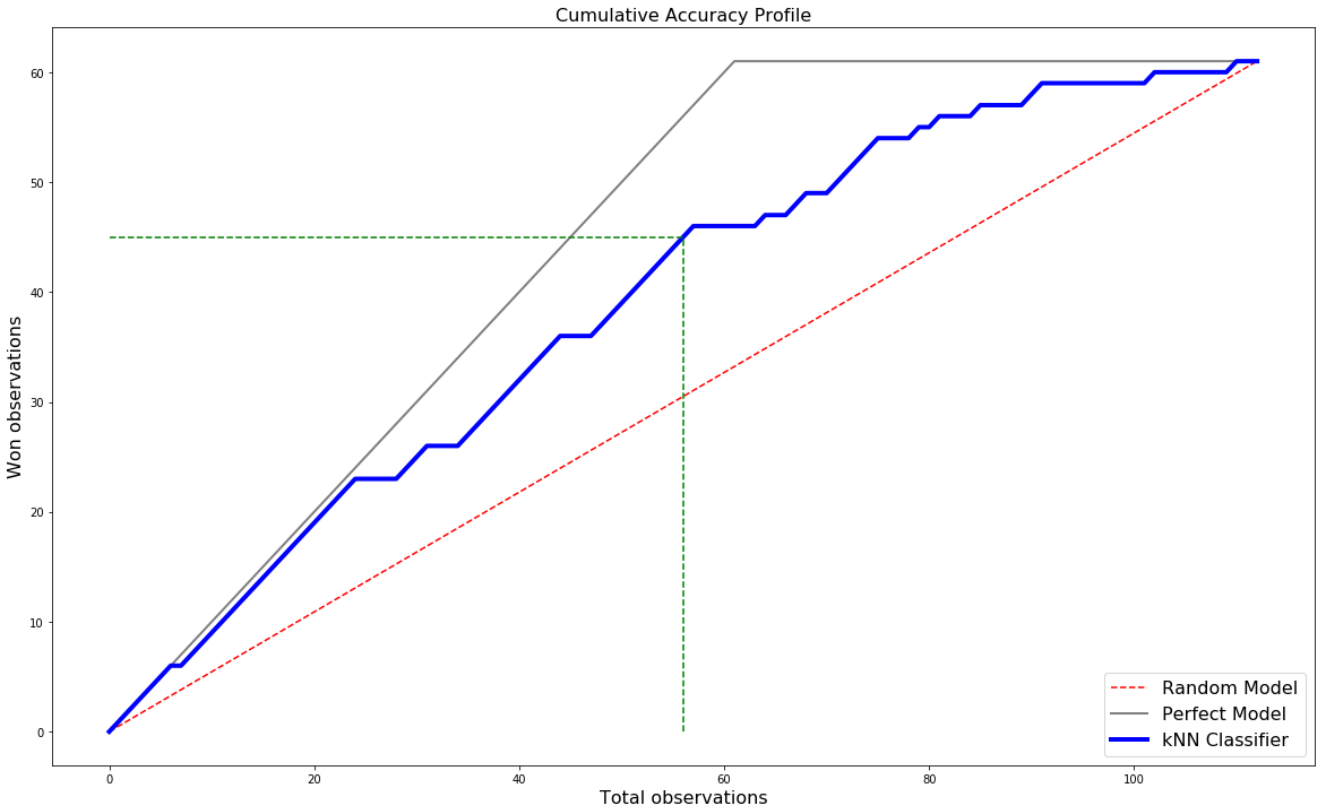

Figure 31: CAP curve of kNN Classifier

Chapter 6

6.1 Conclusion

Finding the most prospective client is the most challenging task for a sales person. Selling to zebras provides a way find out most potential client from their sales pipeline. Still these methodology requires considerable background checks knowledge about past sales and sales people's experience. Thus, sales teams allocate considerable amount of time to dig deep into a sales deal.

In this research we proposed a machine learning models approach to predict the outcome of a given deal. As shown in table 8, we can see that most of our machine learning models were able to predict successful result around 70% of an accuracy. The current selling to zebras selling process has been based only on zebra score. Sales people put their effort to find necessary information on each prospect in order to calculate the zebra score for each opportunity in the sales pipeline. This zebra score sometimes inaccurate since the score has been assigned by a human. In practical scenario, it is very hard to go and assign a score to every opportunity in the sales pipeline. Inexperience of sales person can be ended up being underperforming while trying to find information on every opportunity he has assigned to. The main objective of this study is to accelerate the selling to zebras selling method by predicting outcome of a given opportunity. The proposed model was able to predict the outcome of a given opportunity, around 70% of a success rate. Then the sales people can select opportunities based on this prediction. So that, they can start applying the selling to zebras method to opportunities, which were predicted as 'Won' by the machine learning model. This will accelerate the selling process since it's saving the initial and intermediate research time on a sales opportunity. Furthermore, this insight can be used by sales managers to allocate their resources accordingly.

By observing the result of this experiment, we can come to a conclusion that there is a possibility to predict the outcome of a given opportunity based on the historical sales information of an organization. Therefore, it has proven that the hypothesis of this research is correct to a certain extent by considering the evaluation result.

6.2 Future work

In this study, the model predicts the outcome of a given sales opportunity based on the information which they provided. The output of this model is a set of opportunities which have higher chance of winning. Then the sales teams can work on those profiles by giving highest priority. However the proposed method hasn't give any insight on which opportunity they should work on first. As a future study, we can scale up this model to predict "next most potential opportunity" from their sales pipeline.

Nowadays there are lot of web based services available to find competitors of a given company. It could be an interesting topic to research further on the techniques which predicts most potential similar customers to an existing "won" deal using machine learning and web mining.

References

[1] Koser, J. and Koser, C. (2009). *Selling to zebras*. Austin, TX: Greenleaf Book Group Press.

[2] M. J. Berry and G. S. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Hoboken: John Wiley & Sons, Incorporated., 2011.

[3] A. McAfee and E. Brynjolfsson, "Big Data: The Management Revolution. (cover story)," Harv. Bus. Rev., 2012

[4] F. Söhnchen and S. Albers, "Pipeline management for the acquisition of industrial projects," Ind. Mark. Manag., 2010.

[5] T. M. Smith, S. Gopalakrishna, and R. Chatterjee, "A Three-Stage Model of Integrated Marketing Communications at the Marketing-Sales Interface," J. Mark. Res., 2006.

[7] M. Lambert and J. Basilio, "Sales Forecasting: Machine Learning Solution to B2B Sales Opportunity Win-Propensity Computation Supervisor:"

[6] M. Bohanec, M. Kljajić Borštnar, and M. Robnik-Šikonja, "Explaining machine learning models in sales predictions," Expert Syst. Appl., vol. 71, pp. 416–428, 2017.

[11] S. Russell and P. Norvig, "A modern, agent-oriented approach to introductory artificial intelligence," ACM SIGART Bull., 2007.

[12] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of Machine Learning. 2013.

[13] Kernel Based Algorithms for Mining Huge Data Sets. 2006.

[14] L. P. Kaelbling, M. L. Littman, A. W. Moore, Reinforcement learning: A survey, Journal of artificial intelligence research (1996) 237–285.

[15] C. M. Bishop, Pattern recognition and machine learning, springer, 2006.

[16] F. C. Pampel, Logistic regression: A primer, Vol. 132, Sage, 2000.

[17] C.-Y. J. Peng, K. L. Lee, G. M. Ingersoll, An introduction to logistic regression analysis and reporting, The Journal of Educational Research 96 (1) (2002) 3–14.

[18] S. Dreiseitl, L. Ohno-Machado, Logistic regression and artificial neural network clas- sification models: a methodology review, Journal of biomedical informatics 35 (5) (2002) 352–359.

[19] Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica (Ljubljana).

[20] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained K-means Clustering with Background Knowledge," Int. Conf. Mach. Learn., pp. 577–584, 2001.

[21] C. Ding and X. He, “K-means clustering via principal component analysis,” Proc. twentyfirst Int.Conf. Mach. Learn., vol. Cl, no. 2000, p. 29, 2004

[22] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. View at Publisher View at Google Scholar · View at Scopus

[23] J. C. Culberson, B. P. Feuston, V. Svetnik, C. Tong, A. Liaw, and R. P. Sheridan, “Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling,” *J. Chem. Inf. Comput. Sci.*, 2003.

[24] Bishop, C. M. (*2*006). Machine Learning and Pattern Recognition. Springer, New York.

[25] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Factor to ROC, In- formedness, Markedness & Correlation,” Tech. Rep. SIE-07-001, School of Infor- matics and Engineering, Flinders University, Adelaide, Australia, 2007.

[26] K. Potdar, T. S., and C. D., “A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers,” *Int. J. Comput. Appl.*, 2017.

[27] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation,” *J. Mach. Learn. Technol.*, 2011.

[28] J. Moeyersoms and D. Martens, “Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector,” *Decis. Support Syst.*, 2015.

[29] Investopedia. (2019). Why Operating Margins Matter. [online] Available at: https://www.investopedia.com/terms/o/operatingmargin.asp [Accessed 23 May 2019].

[30] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." In IJCAI, vol. 14, no. 2, pp. 1137-1145., 1995

[31] Rish, I.,“An empirical study of the naive Bayes classifier”, in IJCAI 2001 workshop on empirical methods in artificial intelligence ,vol. 3, No. 22, pp. 41-46,2001

[32] Langley, P., & Sage, S., “Induction of selective Bayesian classifiers”. In Proceedings of the Tenth international conference on Uncertainty in artificial intelligence (pp. 399- 406). Morgan Kaufmann Publishers Inc.

[33] Dougherty, J., Kohavi, R., & Sahami, M., “Supervised and unsupervised dis- cretization of continuous features”. In ICML (pp. 194-202),1995

[34] Yu, L., & Liu, H., “Efficient feature selection via analysis of relevance and redundancy”. The Journal of Machine Learning Research, 5, 1205-1224.,2004

[35] S. Natek and M. Zwilling, “Data Mining for Small Student Data Set – Knowledge Data Mining for Small Data Set As Part of Higher Education,” *Slov. - Act. Citizsh. by Knowl. Manag. Innov.*, 2013.

[36] Zhang Guozheng, Chen Yun, and Fu Chuan, “A study on the relation between enterprise competitive advantage and CRM based on data mining,” 2007.

[37] Ling Zheng, Feng Li, Hui Gui, "The research of ontology-assisted data mining technology", *Information Management and Engineering (ICIME) 2010 The 2nd IEEE International Conference on*, pp. 285-288, 2010.

[38] S. M. S. Hosseini, A. Maleki, and M. R. Gholamian, “Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty,” *Expert Syst. Appl.*, 2010.

[39] K. Khan, U. Karthikeyan, Y. Li, J. Yan, and K. Muniyappa, “Single-molecule DNA analysis reveals that yeast Hop1 protein promotes DNA folding and synapsis: Implications for condensation of meiotic chromosomes,” *ACS Nano*, 2012.

[40] Y. Wei, Y. J. Zhang, and Y. Cai, “Growth or longevity: The TOR’s decision on lifespan regulation,” *Biogerontology*. 2013.

[41] B. Jiang, “Geospatial analysis requires a different way of thinking: the problem of spatial heterogeneity,” *GeoJournal*, 2015.

[42] T. F. Bahari and M. S. Elayidom, “An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour,” *Procedia Comput. Sci.*, 2015.

[43] J. Merkert, M. Mueller, and M. Hubl, “A Survey of the Application of Machine Learning in Decision Support Systems,” *Twenty-Third Eur. Conf. Inf. Syst.*, 2015.

[44] G. Meyer *et al.*, “A machine learning approach to improving dynamic decision making,” *Inf. Syst. Res.*, 2014.

[45] R. Florez-Lopez and J. M. Ramon-Jeronimo, “Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal,” *Expert Syst. Appl.*, 2015.

[46] K. C. Green, A. Graefe, and J. S. Armstrong, “Forecasting Principles,” in *International Encyclopedia of Statistical Science*, 2011.

[47] J. D’Haen, D. Van Den Poel, and D. Thorleuchter, “Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique,” *Expert Syst. Appl.*, 2013.

[48] M. Bohanec, M. Kljajić Borštnar, and M. Robnik-Šikonja, “Modeling attributes for forecasting {B2B} opportunities acquisition,” *Proc. 34th Int. Conf. Organ. Sci. Dev.*, 2015.

[49] M. Luckert, M. Schaefer-Kehnert, W. Löwe, M. Ericsson, and A. Wingkvist, “A classifier to determine whether a document is professionally or machine translated,” in *Lecture Notes in Business*

Information Processing, 2016.[50] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica (Ljubljana)*. 2007.

[51] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha, and V. Honrao, "Predicting Students' Performance Using ID3 and C4.5 Classification Algorithms," *Int. J. Data Min. Knowl. Manag. Process*, 2013.

[52] H. Toh and K. Horimoto, "Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling," *Bioinformatics*, 2002.

[53] W. K. Härdle, D. Prastyo, and C. Hafner, "Support Vector Machines with Evolutionary Feature Selection for Default Prediction," 2017.