INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

VAS Service Prediction for Telco Customers

A Project Proposal by

Mrs. Sanduni Chamodani

Supervised by

Mr. Alroy Mascrenghe

Submitted in partial fulfilment of the requirements for the MSc in Advanced Software Engineering degree at the University of Westminster.

September 2022

DECLARATION

I hereby declare that the work presented in this dissertation and all associated subcomponents and materials are my own work, and none of them has never been submitted or presented as content for any other degree or other educational program. Extracted facts from credible external sources have been properly cited and given due credit.

Name of Student: Raterala Mudiyanselage Sanduni Kanishka Chamodani

Registration No: w1802124 / 20200090

Sandi

Signature: ………………. Date: 2022.09.01

ABSTRACT

A mobile telecommunications company's value-added services provide clients a range of services. Annual revenue from value-added services is significant for telecommunications businesses. The mobile telecommunications market in Sri Lanka is about to reach a point where mobile users will begin to understand that their phones can do more than just make phone calls. Even if users do not use all of the services, mobile network operators' VAS offers are drawing more subscribers since they have realized that mobile VAS has "value." Mobile network companies utilize VAS to boost the number of connections on their networks.

Value Added Services are offered by a value-added service provider externally or internally by the mobile network operator. The network operators make a significant amount of money from these services. It is vital to identify customers from the current customer base who are qualified for each VAS and recommend these services to them in order to reap the greatest benefits. The proper clients will be more likely to use the service and boost revenue for the business if it is recommended to them. These VAS will increase both the consumer base and customer satisfaction. The contented customers of a business are its most important asset, to sum up. On the other hand, a customer will be dissatisfied with the operator if he is suggested a service that he does not desire. Because the consumer will find the messages that make those suggestions burdensome and will become irate with the network operator. Therefore, it is critical to identify the right customers for each VAS.

This research tries to research, design and develop accurate VAS prediction model for telecommunication customers using machine learning. The research used a dataset from a prominent Mobile Service Provider in Sri Lanka and extracted 9 important features to build the model. Different machine learning algorithms, including classical algorithms and ensemble methods, are used to build many models. Stacking, Random Forest, XG Boosting, Bagged CART were used alongside with Logistic Regression, K-nearest neighbor (KNN) and Naïve bayes (NB) algorithms for the predictive modeling and analysis. Predictive model which was built using bagging Classification and Regression Trees algorithm (CART) showed the best results among other models with accuracy of **82.97%**.

Key Words: VAS, Machine learning, Ensemble, Classification

ACKNOWLEDMENT

I would like to take this opportunity to gratefully acknowledge the support and the contribution of all people who have been helping throughout the project. First, I would like to express my special gratitude to my supervisor, Dr Alroy Mascrenghe, for the ongoing advice and assistance provided when reviewing my work and dedicating important time. This constant evaluation helped me to conduct the research project in the right direction. Also, I would also like to thank Mr. Guhanathan Poravi, module coordinator for the research module, who gave me continuous support to compile a comprehensive thesis by providing all the related and additional materials and reviewing the thesis work throughout the research duration. Next, I would like to express my gratitude to all the interviewees and evaluators who sacrifice their precious time to help this project succeed. Additionally, I would want to express my gratitude to the Informatics Institute of Technology's administration, faculty, and staff for their assistance. Thank you to my coworkers for their assistance. Finally, I want to thank my family for all of the help they have given me with my research work.

Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 CHAPTER OVERVIEW

A thorough project outline is provided in this chapter along with an introduction to the research. This contains the problem domain, problem definition, and problem statement for the project. Project background, problem domain, problem definition and problem statement are included in here. Moreover, the related literature is analyzed to find the research gap. Then, the purpose, goal, objective, and scope of the research are established. In conclusion, the chapter thoroughly explains the background, problem, and solution.

1.2 BACKGROUND

In recent years, Mobile Telecommunication Industry gained massive profit from voice, SMS and data. In Sri Lanka, there are several such mobile operators such as Mobitel, Dialog, Hutch and Airtel. Since a large number of operators entered the market last decade, intense competition arose. In the beginning, they provide a variety of voice and data bundles to entice users. Then, instead of providing voice and data, these operators attempted to provide other services. Value Added Services (VAS) were introduced into the field at that time. Those services play an essential role in today's mobile telecommunications operator services, where customers can pay for services. VAS are digital services that are added to mobile networks in addition to voice services, such as ringtones, games, message functions, icons, and so on, and that produce a significant profit for the mobile firm each year. To reap the most benefits, it is necessary to identify clients who are eligible for each VAS from the existing customer base. Therefore, in this research, the author tries to create an efficient and accurate model to select eligible customers for VAS.

In the initial chapters, this document defines problem domain, problem definition, problem statement, research motivation and existing work. In existing work, limitations and improvements are examined. Following that research gap, research contribution, research challenge, research questions, research aim, and objectives are thoroughly examined. Following that, the Methodology section explains the approaches that will be used in this study. Finally, the schedule of the project and the deliverable plans are defined.

1.3 PROBLEM DOMAIN

1.3.1 MOBILE TELECOMMUNICATION INDUSTRY IN SRI LANKA

Until 1989, the telecommunication industry has always been a state-owned industry in Sri Lanka. Then the government decided to change the state monopoly to private sector. The first mobile operator in Sri Lanka is ‘Celltel’ and it introduced mobile telecommunication to Sri Lanka. At that time, a ‘Brick sized’ pocket radio type mobile phone only used by upper class society. The market has become more competitive with the expansion of mobile telecommunications sector, especially with the entry of new mobile operators. Today there are five mobile operators namely Dialog, Airtel, Mobitel, Etisalat and Hutch (“The Mobile Telecommunication Industry In Sri Lanka Management Essay,” n.d.). Sri Lanka's mobile telecommunication industry is one of the country's fastest growing industries and it contributes to economic growth by investments, employment and innovation with introducing latest technology. The mobile telecommunications sector in Sri Lanka is one of the most rapidly expanding industries in the country, due to introduction of new technology and rising public demand. In terms of technology, Sri Lanka's telecommunications business has always been at the forefront of the region, having been the first to introduce most technologies such as GSM, CDMA, Wimax, 3G, HSPA, and so on. Because of improved communication, all segments of the population have improved their lifestyles and quality of life.

1.3.2 VALUE ADDED SERVICES (VAS)

In Sri Lanka, the mobile telecommunications industry is approaching a point where mobile subscribers are realizing that their phones can do more than just make phone calls. They have recognized that mobile VAS has a "value." Even if customers do not use all the services, mobile network operators' VAS offerings are attracting an increasing number of subscribers. VAS is being used by mobile network providers to increase the number of connections within their networks.

'Value-added' services are non-core services that go beyond standard service offerings like voice calls and fax transmissions in the telecommunications business. Mobile value-added services (VAS) are services that go beyond normal communications services and add value to the core service offering. SMS (short message service), MMS (multimedia messaging service), and data access are usually considered core features for general mobile phone use. Those Value added services are Internet-based SMS, MMS, ringtones, games, message functions, icons and so on.(Wickramaratne, 2004)

1.4 PROBLEM DEFINITION

Annually, Value Added Services make a significant profit for the mobile company. Those services are provided either internally by the mobile network operator or externally by a value-added service provider. These services bring in a lot of money for the network operators. To gain the most benefits, it is necessary to identify customers who are eligible for each VAS from the existing customer base and recommend these services to them. If the service is recommended to the right customers, they are more likely to use it, increasing the company's profit. Client satisfaction will also improve as a result of these VAS, as will the customer base. Finally, a company's most valuable resource is its delighted consumers. On the other side, if a consumer is recommended a service that he does not want, the customer will be displeased with the operator. Because the messages that make those suggestions will be a burden for the consumer, and he will be dissatisfied with the network operator. As a result, it is critical to identify the right customers for each VAS.

1.4.1 PROBLEM STATEMENT

Value Added Services do not reach to the preferred customers and customers who don’t prefer those services are annoyed by suggesting those services.

1.5 RESEARCH MOTIVATION

It is beneficial to suggest a service to a client who prefers it, and it is also beneficial to prevent consumers who do not prefer such services from suggesting those services. Customers have a lot of concerns about promotional and service messages and calls, based on my experience as a Mobitel (Pvt)Ltd employee. Some of those customers are annoyed by the unwanted texts and phone calls, and because of that they are disappointed with the mobile operator. Therefore, it's critical to identify the most suited clients and only recommend the service to them.

1.6 EXISTING WORK

Citation	Brief Description	Limitations	Improvement
(Vahidi Farashah et al., 2021)	Combines the X-Means algorithm, the ensemble learning system, and the N-List structure to analyze the customer portfolio of a mobile telecommunication company and provide value added services	Model requires long computation time.	Achieved around 97.6% accuracy.
(Szymkowiak et al., 2018)	An Apriori algorithm is used for customer basket analysis	This research suggests a complex model	Model has an optimal accuracy in classification
(Srivastava et al., 2018)	In this research, a portfolio optimization of customer basket is used	Model has a low accuracy	High speed
(Seyedan and Mafakheri, 2020)	Presented some classification algorithms and their applications	In this project a complex model is suggested	Model has high classification accuracy
(Kurniawan et al., 2017)	Used associative and data mining techniques	The results have high Mean Absolute Error (MAE) and high Root Mean Squared Error (RMSE).	Model has fast execution time

Table 1 Related Works

1.7 RESEARCH GAP

In the mobile telecommunications industry, there is only one recent research paper for customer basket analysis in VAS services. In addition, no research has been conducted in this field in Sri Lanka. Therefore, these gaps will be addressed in this study. In addition, questionnaires are used to build data sets in the majority of customer basket analysis studies. As a result, the data sets' accuracy is questioned. Author intends to use a legitimate dataset from a prominent mobile operator in Sri Lanka for this study. As a result, the accuracy of the results will improve.

1.8 RESEARCH CONTRIBUTION

1.8.1 TECHNICAL CONTRIBUTION

This system will be implemented using reinforcement learning methods. Reinforcement learning has not been used for customer basket analysis for VAS in telecommunication industry. Also, the author hopes to try many machine learning algorithms and ensemble learning methods and build the most accurate and efficient model.

1.8.2 DOMAIN CONTRIBUTION

There is a lack of research focused on customer basket analysis for VAS in telecommunication industry. Also, there is not any such research done in domain of Sri Lanka. So, the author hopes to create a most accurate and efficient model to classify customers in Sri Lanka domain.

1.9 RESEARCH CHALLENGE

Prediction is done considering the previous behavior of customers. Human behavior can be changed from one individual to another and throughout time. Therefore, even though the prediction is correct for one person, another may get a wrong result.

1.10 RESEARCH QUESTIONS

RQ1: What are the newer advancements and techniques in machine learning that can be used to get more accurate results for creating models?

RQ2: How Machine Learning can be automated so that user doesn’t have to create models specific for each dataset?

RQ3: How can we use big data in this research, so we can improve the performance and accuracy?

1.11 RESEARCH AIM

The aim of this research project is to design, develop and test an accurate, efficient model to select customers for a VAS in mobile telecommunication industry.

1.12 RESEARCH OBJECTIVE

Research Objectives	Explanation	Learning Outcome
Problem Identification	• Analyze reasons for customer complains towards Mobile operators • Discuss and confirm the findings with the domain expertise. • Identify the research problem	LO1
Literature Review	• Analyze the research done in basket analysis for VAS in telco system • Analyze the research done in customer basket analysis • Identify the research gaps in the existing research done in basket analysis for VAS in telco system • Identify different techniques and algorithms to develop the model	LO1, LO2, LO4, LO6
Data Gathering and Analysis	• Identify the requirements that are needed for the project	LO2, LO3, LO4, LO5, LO6

	• Choose a specific local platform in the mobile telecommunication industry • Obtain legal permission from the company in order to collect the data • Analyze how these data should be pre-processed before the training	
Research Design	• Find out most efficient algorithms by considering previous research findings from the literature review • Identify the problem-solving approach • Design the architecture of the prototype	LO2, LO3, LO4, LO6
Implementation	• Implement the model • Train the model • Build the prototype	LO2, LO3, LO4, LO6
Testing and Evaluation	• Evaluate the model • Test the prototype	LO2, LO3, LO4, LO6

Table 2 Research Objectives

1.13 PROJECT SCOPE

1.13.1 IN-SCOPE

The scope that is covered in the project is as follows.

- Initial data gathering: Get a legitimate data set and preprocess them to get more accurate data. In this stage, noise removal, feature selection, data normalization and handling class imbalance is done.
- Building a model: Get 70% of data from data set and train data using several algorithms and create a model. Test the model using remaining 30% of data from the dataset.

- Model evaluation: Evaluate various models created in the above processes and picking the best model.

1.13.2 OUT-SCOPE

The parts that will not be covered in the project are as follows

- Automating a data retrieval from the mobile platform is not handled under this project implementation

1.14 PROTOTYPE FEATURE DIAGRAM

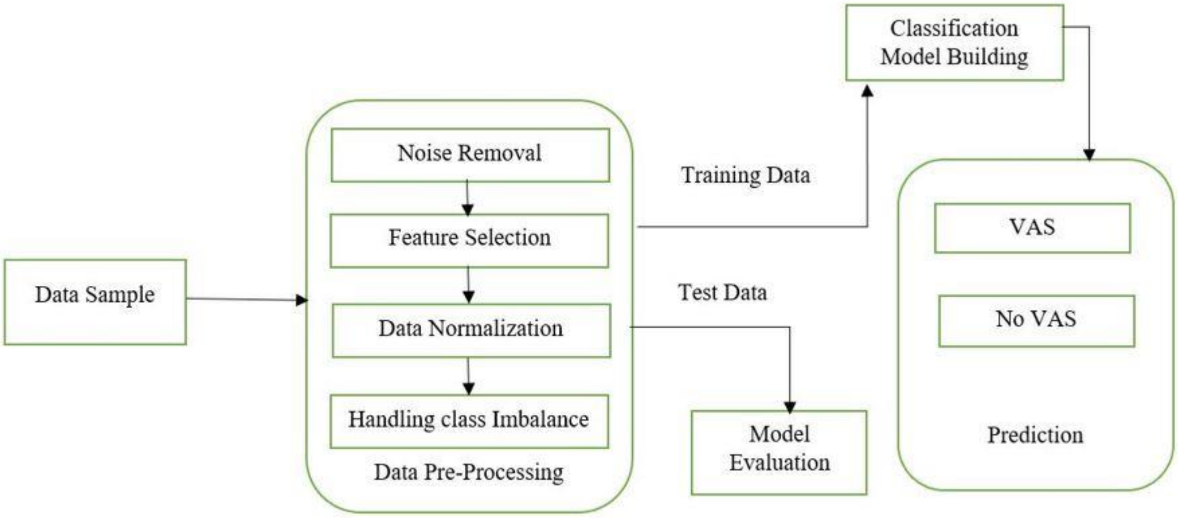


Figure 1 Prototype Feature Diagram

1.15 CHAPTER SUMMARY

This chapter covered the complete introduction to the research project. Project background, problem domain, problem definition and problem statement are discussed here. Further, this chapter describes how the research gap is identified by analyzing past research works. Moreover, the research contribution, question, aim, objective and scope are described. In summary, this chapter explains the background, problem and solution in detail.

CHAPTER 2: LITERATURE REVIEW

2.1 CHAPTER OVERVIEW

Machine learning prediction models have been solved thousands of research problems throughout past years, among those studies, the area of market basket analysis for VAS services has gotten limited attention from the researchers even though close parallel approaches such as customer basket analysis for supermarkets have commonly studied. This chapter will examine the available customer basket analysis prediction research while reviewing the details of algorithms used, results achieved and limitations

2.2 PROBLEM DOMAIN

2.2.1 TELECOMMUNICATION INDUSTRY

All telecommunications/telephone companies and internet service providers make up the telecommunications sectors, which play a critical role in the expansion of mobile communications and the information society.

Traditional phone conversations are still the industry's most profitable source of revenue, but because to advancements in network technology, telecom is becoming less about speech and more about text (messaging, email) and graphics (e.g., video streaming). High-speed internet connectivity is widely available for computer-based data applications including broadband information services and interactive entertainment. The major broadband communication technology is digital subscriber line (DSL). (Value-added) services supplied through mobile networks have seen the most rapid increase. (“Telecommunications industry,” 2022)

Telecommunication is the transmission of information over significant distances to communicate. Visual signs, such as beacons, smoke signals, semaphore telegraphs, signal flags, and optical heliographs, or sound messages, such as coded drumbeats, lung-blown horns, and loud whistles, were used in the past via telecommunications. Telecommunications in modern times entails the use of electrical technologies such the telegraph, telephone, radio, and microwave communications. Fiber optics and their related electronics, orbiting satellites, and the Internet are all used in communication channels. Modern telecommunications industry players manufacture communication equipment and provide voice, data, and broadband services across wireline and wired infrastructure, which includes cables, networks, servers, computers, and satellites.

The telephone industry is interrelated to the computer sector. Cable companies are replacing telephone companies in numerous ways. Low-cost alternatives came as a result of technological breakthroughs, making equipment more accessible to people and enterprises all over the world.

In today's world, the telecommunications system relies on a centralized network to send data from one area to another. The move from physical lines to wireless took some time, but it happened ultimately. (“Telecommunications Industry,” 2022)

This industry is extremely sensitive to even minor changes in legislative, technological, and economic considerations, and it faces its unique set of issues as a result. With wireless and broadband providing the necessary drive for their expansion, industry players have developed different solutions to overcome these obstacles and continue to connect individuals and companies.

The telephone, long-distance, cable/video, cellular, and telecommunications equipment businesses are all becoming increasingly interconnected in today's world. Globalization trends, competitive impacts of new technologies, and the changing regulatory environment are all major issues for this industry.

Telecom is the world's fifth largest and fastest-growing industry. Telecommunications play an essential role in the global economy; in 2008, the global telecommunications industry generated $3.85 trillion in sales. The worldwide telecoms industry's service revenue was predicted to be $1.7 trillion in 2008, and it is expected to reach $2.7 trillion by 2013. The telecommunications industry is divided into two categories: equipment and services.

Companies in the equipment sector provide products that are utilized by both customers and other companies in the same industry. Customers utilize these products to obtain telecommunications services, while other companies use them to develop, maintain, and deliver telecommunications infrastructure and services. Satellite and broadcast network equipment, wireless and wireline equipment, and computer networking equipment are all part of the equipment sector. Wired and wireless services, as well as internet and other broadband services, are included in the Services Sector.

2.2.2 TELECOMMUNICATION INDUSTRY IN SRI LANKA

Telecommunications, computer, and information services are important components of the services sector that grew steadily in 2019 in Sri Lanka. By 2025, Sri Lanka's National Export Strategy seeks to generate $5 billion in revenue, 200,000 direct jobs, and 1000 Information Technology and Business Process Management (IT-BPM) start-ups through the IT-BPM business. In 2020, total credits from the ICT sector were estimated to be over $1 billion. The country began the process of implementing fifth generation (5G) technology in 2018. Software companies in the United States have had success selling solutions to private consumers as well as some government bodies. Sri Lanka is constructing a nationwide fiber optic network that will be connected to multiple international cables.

The market is gearing up for the transition from 4G to 5G mobile services. In 2019, Dialog Axiata and Mobitel performed pre-commercial 5G experiments, with Dialog repurposing 20% of its LTE antennae for 5G compatibility.

The telecommunications sector in Sri Lanka is one of the most dynamic in the country, contributing considerably to investment, employment, productivity, innovation, and overall economic growth, both directly and indirectly. Five mobile operators serve a population of 22 million people in Sri Lanka's telecommunications market. Sri Lanka's telecommunications sector attracts a lot of foreign direct investment. Telecommunications use has increased, with total fixed line and mobile phone density growing to 142 per 100 people, owing to an increase in mobile users. The sector is dominated by mobile phone companies. Telecommunication services are now available throughout the country thanks to ongoing infrastructure construction. Three fixed-line operators, five mobile phone operators, and 11 Internet service providers are all in fierce competition. The U.S. exported approximately $2.4 million of telecommunications equipment to Sri Lanka in 2020.

2.2.3 VALUE ADDED SERVICES (VAS)

Value-added services are typically promoted as premium features and add-ons to core functionality. Telecommunications companies use them to increase demand for core services, although they can often operate on their own. They are usually designed to create operational and/or administrative synergy among the product's variety of services, rather than simply diversify the capabilities of the product bundle. Value-added services are thought to benefit both customers and service providers because they can give increased data and analytics for business usage in addition to adding product capability for end users.

On a conceptual level, value-added services in the telecommunications sector provide value to the regular service offering, encouraging customers to use their phones more and allowing the operator to increase their average income per user. SMS, MMS, and data access have traditionally been considered value-added services for mobile phones, but in recent years, SMS, MMS, and data access have increasingly become core services, and VAS has begun to omit such services. Mobile VAS services can be categorized into consumer behavior VAS, network VAS and enterprise VAS. Major value-added services are live streaming, location-based services, missed call alerts, mobile advertising, online gaming, ring tones, ring back tones, mobile TV, M-Commerce based services and WAP content downloads.

2.3 CONCEPT MAP

A concept map is used to define and represent the whole topic of the literature review. It demonstrates how the study topics converge from a general perspective to specific subdomains. This was used to organize the material to be discussed in the literature review.

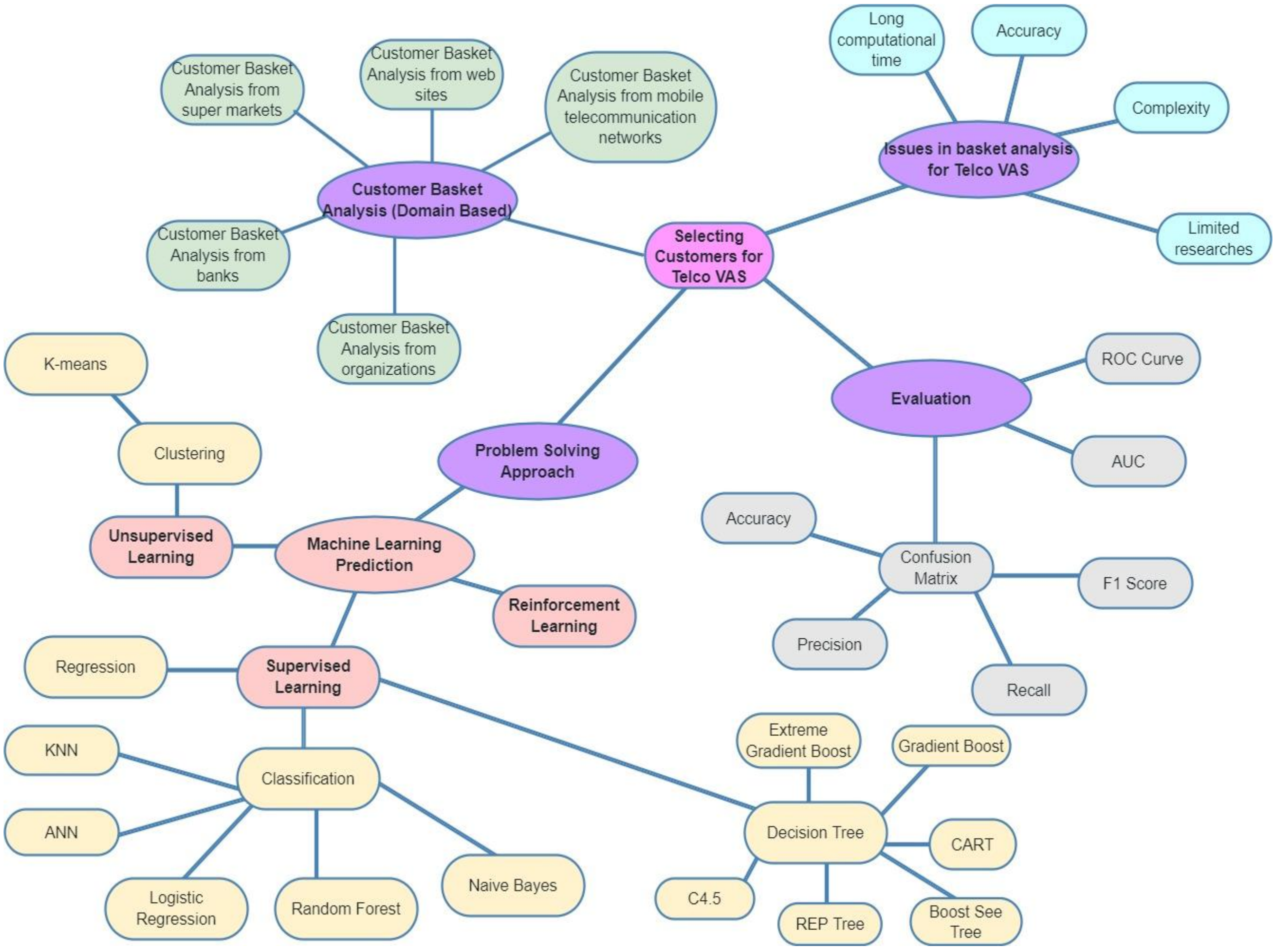


Figure 2 Concept Map

2.4 EXISTING WORK

2.4.1 CUSTOMER BASKET ANALYSIS TO IDENTIFY CUSTOMER BEHAVIORS

Customer basket analysis has been used in a number of studies to identify customer behavior. This section will refer to researches that study customer basket analysis and define concepts and techniques.

In research (Seyedan and Mafakheri, 2020) published a classification of these algorithms and their applications. In this paper, big data analytics (BDA) applications in supply chain demand forecasting are investigated to propose a classification of these applications, find gaps, and provide ideas for further research. Time-series forecasting, clustering, K-nearest-neighbors, neural networks, regression analysis, support vector machines, and support vector regression are the categories in which these methods and their applications in supply chain management are classified. Even though the presented classification has good accuracy, the model is complex.

In research (Szymkowiak et al., 2018) and his colleagues proposed an Apriori algorithm for customer basket analysis. On the large statistical population, the Apriori associative algorithm has an infinite constraint. They were able to obtain the needed accuracy by using data and items from a supermarket in their research. As a result, one of the most important advantages of this model is that it has a fast basket analysis speed and medium accuracy, while one of the research's major shortcomings is its lack of comprehensiveness and high adaptability.

In research (Srivastava et al., 2018) used an effective portfolio optimization approach of customer shopping termed utility mining. They proposed utility mining as a better data mining model. They were able to conduct the consumer basket analysis process fast and accurately using the technique offered, but they did not have the possibility for further progress.

In research (Kurniawan et al., 2017) presented a transaction-based client basket analysis methodology. They used associative and data mining techniques such as neural networks and Apriori in their research. One of the most significant advantages of their work was the speed with which basket analysis could be completed. However, one of the model's significant drawbacks was its lack of precision for online portfolio analysis, as well as its lack of comprehensiveness and its inability to perform well on huge data sets.

2.4.2 CUSTOMER BASKET ANALYSIS FOR TELCO VAS CUSTOMERS

In research (Vahidi Farashah et al., 2021) proposed an analytical model for telco Vas customers' basket clustering using ensemble learning approach. This is the first model that created for analyzing the customers for VAS in telecommunication industry. In that paper, the X-Means algorithm, the ensemble learning system and the N-List structure are combined to analyse the customer portfolio of a mobile telecommunication company and provide value-added services. Optimal number of clusters and clustering of customers in a company are determined using the X-Means algorithm. New elder customers are assigned to categories using the ensemble learning algorithm. The ensemble learning algorithm is also used to assign categories to new elder customers and the N-List structure is used for customer basket analysis.

In this paper, the N-List algorithm-based technique is used to evaluate the customer basket and the suggested ensemble learning system is used to improve the accuracy of customer basket analysis. The proposed N-List method assures that the comprehensiveness of the service is preserved while increasing the service execution speed. This study's suggested ensemble learning system combines three machine learning algorithms: deep neural networks, C4.5 decision trees, and the SVM-Lib algorithm. (The support vector machine algorithm's library core). At each stage, the proposed ensemble learning system uses maximum votes to transmit the optimal response to the output.

The main contributions of this paper are as follows.

- K-Means and X-Means clustering algorithms are combined to generate an efficient clustering algorithm to determine the initial grouping of data
- The TelecoVAS Customer Basket of a Mobile Communication Company was analyzed using a combination of ensemble learning and the N-List algorithm
- Using the N-List approach to improve the results achieved by combining machine learning methods in an ensemble learning system.

To deliver attractive value-added services to telecommunication clients, the proposed method in this research is based on X-Means clustering methods, N-List structure for extracting frequent patterns, and ensemble learning system.

Main stages of this research are as follows

Data Preprocessing

Data clearing method was used in this paper as data pre-processing method. Data is checked to find out if a row or a column contains null or unused value, if it is so, then the mean of the next and previous values will be calculated and replaced with null values. This data clearing eliminates outliers and produces more consistent data.

Data Normalization

Data normalization is used to increase clustering accuracy. All datasets are mapped into matrices, and matrix rows are normalized.

Customers clustering using XK-means algorithm

Combination of K-Means and X-Means algorithm is used for clustering customers based on behavior information. The X-Means clustering algorithm receives behavior information of customers and then it directs each customer to a cluster based on the behavior information. Optimal K value is found using X-Means algorithm by introducing the all customers of telecommunication company to the X-Means algorithm.

To assign labels to new customers, this study uses the X-Means clustering technique, which is an extended version of K-Means. As a result, the X-Means algorithm's input is the telecommunication company's customers. The output of this algorithm is k. Finally, the K-Means clustering algorithm is applied with the number k. Customers of the telecommunication firm are the input to the K-Means clustering algorithm, and labelling the customers is the result.

The ensemble learning

Most popular classification algorithms in ensemble learning are used such as deep neural network, the C4.5 decision tree with the Information Gain Kernel and the SVM-Lib algorithm for classifying new customers in mobile telecommunication companies.

The C4.5 decision tree is integrated with the Information Gain core and the SVM-Lib algorithm in the ensemble learning system, and the best batch is selected from the batches offered as the eventual result for the new customer specified at each stage.

The training data, which is for 70% of the total data, is put into the algorithms, and a model is created. Experimental data are also put into the models that are created to categorize people based on their behavior. For new customers, a new category is chosen during the ensemble phase. Customers in the target group behave the same way as other customers do. Following the implementation of the process search system and the assignment of a new category to the new customers, the N-List structure was applied to all client baskets in the selected category, and finally, a set of services was supplied for the new customers based on the analysis.

Basket Analysis using N-list algorithm

In this paper, basket of customers interested in receiving value added services are analyzed based on their behavior extraction and customer transaction records. In this study, N-list algorithm is used to analyze the customer cart. Customer transactions are processed using N-List algorithm based on its tree structure and offers customer services based on extracted repetitive rules and transactions. A set of characteristics that are useful in repeating transactions is taken from repetitive transactions and then used in the ensemble learning system.

2.4.3 BENCHMARKING

According to the above existing work, there have been just a few attempts at analysis, specifically related to telecommunication Value Added Services. Even though, the above study uses clustering techniques, deep neural networks, online hybrid similarity as a method for analysis to improve the classification accuracy, it costs long computation time. Also, each of the proposed methods has challenges such as inaccuracy and high error of recommendation. Using different supervised machine learning classifiers, this study will determine the optimum machine learning method for solving this specific research challenge. Also, this research will try to improve the accuracy of models using ensemble techniques by comparing the performance using standard machine learning model evaluation criteria.

2.5 PROBLEM SOLVING APPROACH

2.5.1 DATA PREPROCESSING APPROACHES

The process of converting raw data into a format that may be understood is known as data preparation. It is also an important step in data mining as we cannot work with raw data. Before implementing machine learning or data mining techniques, it has to be checked the quality of data. Data validation and data imputation are important parts of the pre-processing process. The purpose of data validation is to determine whether the data is comprehensive and accurate. The purpose of data imputation is to rectify errors and fill in missing numbers, which can be done manually or automatically.

The study (Batista, G.E., Prati, R.C. and Monard, M.C., 2004) has evaluated ten different methods of under and over-sampling techniques namely Random over-sampling, Synthetic Minority Oversampling Technique (SMOTE), Random Under sampling, Tomek links undersampling, Condensed Nearest Neighbor Rule undersampling (CNN), One-Sided Selection undersampling (OSS), Neighborhood Cleaning Rule undersampling (NCL) and some combinations like Smote + Tomek links, CNN + Tomek links, Smote + ENN. The results have suggested that in general, over-sampling methods provide more accurate results than under-sampling methods. And also the proposed combined sampling methods Smote + Tomek and Smote + ENN have performed better than other methods. Moreover, Random over-sampling has achieved competitive results with the more complex methods even though it is a simple technique that is computationally less expensive.

2.6 EVALUATION APPROACHES

Confusion Matrix

A confusion matrix is a table that is used to define a classification algorithm's performance and it is a very popular measure. The performance of a classification algorithm can be visualized and summarized using a confusion matrix. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa. The confusion matrix consists of four basic characteristics. Those characteristics are used to define the measurement metrics of the classifier. Those are:

1. TP (True Positive): number of positive examples classified accurately
2. TN (True Negative): number of negative examples classified accurately
3. FP (False Positive): number of negative examples classified as positive
4. FN (False Negative): number of actual positive examples classified as negative

Accuracy

Accuracy is the number of correct predictions (True Positives + True Negatives) made by the model over all kinds of predictions made. Accuracy can be considered as the most straightforward metric to evaluate classification models.

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

The value of accuracy is increased by the number of correct predictions made by the model. Accuracy is a good measure when the target variable classes in the dataset are balanced. Therefore, accuracy is insufficient on its own. So, other metrics must be considered such as recall and precision

Recall

Recall or sensitivity is the ratio of True Positives to all the correct predictions (True Positive + False Negative) in the dataset. Recall measures how good the model is at correctly predicting positive classes. Actual positive classes are the focus of recall. It shows how many of the positive classes the model can accurately predict.

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

The value of recall is lowered when there are higher number of False Negatives.

Precision

Precision is the ratio of True Positives (TPs) to all the positives predicted by the model (True Positive + False Positive). Precision measures how good the model is when the prediction is positive. Positive predictions are the objective of precision. It shows how many of the positive predictions came true.

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

Because there is a trade-off between precision and recall, we can't aim to maximize both. Precision decreases recall and vice versa when precision rises. Depending on the task, we can aim for maximum precision or recall.

As an example, email spam detection model can be considered. In there, precision has to be maximized, because model has to be correct when an email is detected as spam. Also, normal email should not be labelled as spam (false positive). On the other hand, we need to maximize recall for a tumor detection task since we want to detect as many positive classes as possible.

F1 Score

F1 score combines precision and recall into a single number. It is the weighted average of precision and recall. Since it considers for both false positives and false negatives, F1 score is a more relevant indicator than accuracy for situations with uneven class distribution. The best value for F1 score is 1 and the worst is 0.

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

ROC Curve

By combining confusion matrices at all threshold values, the ROC curve describes the model's performance at different threshold values. The true positive rate (sensitivity) (TPR) is represented on the ROC curve's x axis, while the false positive rate (FPR) is represented on the ROC curve's y axis (1- specificity).

$$TPR = \frac{(TP)}{(TP + FN)}$$

$$FPR = 1 - \frac{(TN)}{(TN + FP)} = \frac{(FP)}{(TN + FP)}$$

AUC

The area under the ROC curve between (0,0) and (1,1), which can be determined using integral calculus, is referred to as the AUC. AUC basically sums up the model's performance across all thresholds. The highest possible AUC value is 1, indicating that the classifier is perfect. The classifier is better if the AUC is close to 1.

2.6 CHAPTER SUMMARY

This chapter focused on a depth review on the problem domain, existing works in customer basket analysis and evaluation methods. The knowledge gained from this chapter along with author's analysis will be utilized throughout the project.

CHAPTER 3: METHODOLOGY

3.1 CHAPTER OVERVIEW

This chapter's aim is to go over how to choose a research methodology, a development methodology, and a project management methodology. The concept, approach, strategy, choice, horizon, and data gathering methods of research methodology are described in detail. The life cycle model, design methodology, and evaluation methodology are then highlighted as part of the discussion of the development approach. Finally, the project management approach is detailed along with the required deliverables, deadline, and resources. In conclusion, the chapter determines how to carry out the research project using each strategy.

3.2 RESEARCH METHODOLOGY

Research Philosophy	**Positivism** was selected among other philosophies such as realism, interpretivism, and pragmatism. The understood problem and the proposed solution are based on the data collected and are not biased by researcher’s ideologies. Also, the data set is highly structured and have large samples and data can be measured and they are quantitative. So, the author has selected positivism over interpretivism.
Research Approach	**Deductive** research approach was selected among other approaches such as deductive, inductive and abductive. This is because the research based on a hypothesis and aims to test and prove the hypothesis it has defined. Also, this is quantitative research.
Research Strategy	**Experiment research strategy and survey research strategy** is chosen for this research among the candidates of case studies, experiment, surveys, interviews and action research.

Research Choice	The choice determines whether this research involves one or more approaches. **Mixed methods** were used because both quantitative and qualitative methodologies were used.
Time zone	**Cross-sectional** was chosen as the time horizon among other options such as longitudinal. This is due to the fact that all of the datasets for this study were collected at the same time. The data set should be defined at the start of the study. There is no need to collect data for an extended period of time.

Table 3 Research Methodology

3.3 DEVELOPMENT METHODOLOGY

3.3.1 LIFE CYCLE MODEL

Agile life cycle method is selected among other life cycle models such as Waterfall, V-shaped, Evolutionary Prototyping, Spiral Method, Iterative and incremental method. The Agile SDLC model combines iterative and incremental process models with a focus on process adaptation and customer satisfaction through the delivery of a working software product rapidly. Agile methods break a project into small, incremental steps and those are provided in iterations. Changing requirements of the project can be done by this iterative and incremental process models. Therefore this life cycle model is selected.

3.3.2 DESIGN MEHODOLOGY

There are several design methodologies that can be use in projects. They are level-oriented design, data flow-oriented design, data structure-oriented design and object-oriented design. Among them **data flow-oriented** design is selected because this system needs to be designed considering data flow.

3.3.3 EVALUATION METHODOLOGY

The research uses the most popular typical machine learning model evaluation approach of confusion metric because the research output is a machine learning prediction model. Furthermore, the research will summarize the output model's performance using common methodologies such as the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC).

3.4 PROJECT MANAGEMENT METHODOLOGY

Agile Prince 2 is selected as the project management methodology among other methodologies such as Agile, Adaptive, DSDM, CPM, CCPM, Feature Driven, Six Sigma, RAD, Scrum, Waterfall and so on. Agile Prince 2 is selected because it allows to focus on management, recursive planning and flexible delivery while adapting to risks.

3.4.1 RESOURCE REQUIREMENT

The resources required to complete the project are identified based on the objectives, expected solution, and deliverables. The following are the software, hardware, and data resource requirements.

Software Requirement

- **Operating System (Windows 10 / MacOS / Ubuntu) -** The Operating System is used to execute all of the software that is installed on top of it. The Ubuntu Operating System will be utilized for this study because it is developer friendly and simple to use.
- **R / Python -** The language that will be used to develop the proposed system. Python is a general-purpose language that can help the project progress; however R is a research-friendly language that is ideally suited for this project.
- **R Studio -** The development environment required to develop the proposed solution
- **Zotero -** The research assistant tool used to manage and backup research papers and artifacts.
- **MS Office -** Tool to create reports and documentations.
- **Google Drive -** To backup files related to the project

Hardware Requirements

- Core i7 processor - To be able to perform high resource intensive tasks of ML.
- 16 GB RAM - To be able to manage the huge volumes of datasets that are used in ML researches.
- Disk space of 40 GB or above – To store the application code and testing files and the data sets needed for processing.

Skills Requirement

- The knowledge of machine learning algorithms needs to be required.
- The knowledge on Reinforcement learning required.

Data Requirement

- Machine learning datasets – Customer data set regarding their past behavior is required for this project

3.4.2 SCHEDULE IN GANTT CHART

Please find Gantt Chart in Appendix A.

3.4.3 DELIVARABLES AND DATES

Deliverable	**Date**
Project Proposal The proposal of the research project	11th Nov 2021
Literature Review Document Review of existing research work and solutions	25th Nov 2021

Software Requirement Specification Document (SRS) Document which specifies the software system which is to be developed	20th Dec 2021
First Version of Prototype First version of prototype of the design and implementation of this project	06th Jan 2022
Interim Progress Report Document on the current progress of the research	10th Feb 2022
Testing and Evaluation Testing and evaluation of designed prototype	24th Mar 2022
Draft Report Report on the current progress of the research	21st Apr 2022
The Thesis The final report documenting the project and research process and decisions	09th May 2022
The Final Prototype The Completed prototype of the design and implementation of this project	09th May 2022
Research Paper Research paper that will be created for publishing this research	22nd May 2022

Table 4 Deliverables and Dates

3.4.4 RISK MANAGEMENT

In the following table, Probability of occurrence and Magnitude of Loss is from 1 to 5 scale.

Risk Item	Severity	Frequency	Mitigation Plan
Changing requirements of the project - Requirements can be changed while doing the project in each iterations	3	5	Cohesion and coupling have to be used.
Lack of knowledge in relevant areas – Lack of knowledge in some concepts of machine learning such as reinforcement learning	5	1	Managing time for self-study, experimenting, getting feedback from experts
Health related issues – The author may get sick	5	3	Submit Mitigating circumstances and extend the deadline
Wrong Estimations – The project deadlines may be delayed due to inaccurate estimates	5	3	Allocate a buffer time in planning
Limited Access to tools and technologies - It is possible that the particular tool or technology required to complete an activity does not exist or is insufficient for the task at hand.	2	3	Alternative tools of technologies have to be found.

Table 5 Risk Management

3.5 CHAPTER SUMMARY

The selection of research, development, and project management approaches were covered in this chapter. The research methodology is also described in terms of its philosophy, approach, strategy, choice, and horizon, as well as its data collection techniques. Following that, the life cycle model, design methodology, and evaluation methodology are highlighted in order to discuss the development approach. The project management technique, including the planned timetable, deliverables, resource requirements, and risks with mitigations, is explained in the last section.

CHAPTER 4: REQUIREMENT GATHERING

4.1 CHAPTER OVERVIEW

Focus of this chapter is on requirement gathering for the project. First, the stakeholders were identified and their roles with the system are explained in the rich picture. The Saunder's Research Onion is then used to illustrate stakeholders, and all stakeholder viewpoints are presented. A review of requirement elicitation techniques and an analysis of the requirements gathered come next. The system is described in a context diagram and a use case diagram. Finally, functional and non-functional requirements of the system are also discussed

4.2 RICH PICTURE

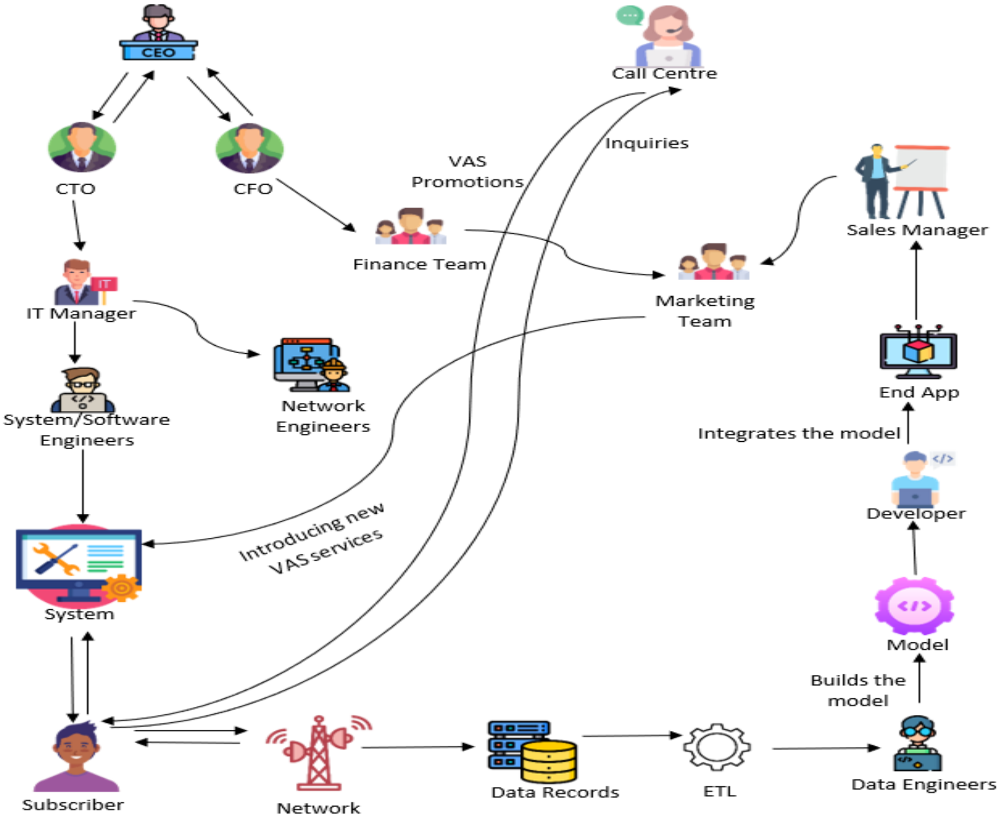


Figure 3 Rich Picture

The goal of the rich picture diagram was to show how various system stakeholders interacted with it. It helps to see relationships and connections and understand the complexity of entire situation. The above rich picture shows the stakeholders in the system and high-level connections between stakeholders, processes and systems.

4.3 STAKEHOLDER ANALYSIS

All stakeholders in the system are identified and represented in the onion model diagram. The role and the view of each stakeholder is identified.

4.3.1 ONION MODEL

Please find the Stakeholder Onion Model in Appendix B.

This figure shows the stakeholders and their roles associated with the project. The pressure points are as follows.

1. The developer should complete the project to satisfy the requirements.
2. The developer should complete the project within the timeline.
3. The developer should make sure the system is secure.
4. The developer should ensure about the quality of the system.
5. The developer should ensure the system satisfies the guidelines of ML experts.

4.3.2 STAKEHOLDER VIEWPOINTS

Stakeholder	**Role**	**Benefits**
Data Engineers, Business Analysts, DevOps Engineers	Operational Beneficiary	Maintains the product
Strategic Managers	Functional Beneficiary	Consumes the product in order to collaborate with senior management on decisions

Product Owner	Functional Beneficiary	Owns the product and is in charge of ensuring its stability
Project Manager	Functional Beneficiary	Has overall responsibility for the planning, design and execution the project
Developer	Financial Beneficiary	Develops the product
Supervisor	Advisory	Provides the guidance which requires to complete the project successfully
ML Expert	Expert	Evaluates the system and provides expert feedbacks
Management	Expert	Offer expert opinion on the domain to enhance the product
Hacker	Negative Stakeholder	Tries to collect the information from the system and tries to cause a problematic behaviour in the system
Competitors	Negative Stakeholder	Competitors (other telecommunication providers) can acquire knowledge to build and improve their system.

Table 6 Stakeholder Viewpoint

4.4 SELECTION OF REQUIREMENT ELICITATION METHODOLOGIES

Researching and collecting about a system's requirements from users, customers, and other stakeholders is known as requirements elicitation. Additional findings from the requirement gathering stage aided the author in understanding the limitations and expectations of the customers. In this section, multiple such options are explored such as literature review, interviews, brainstorming, observation and questionnaires.

1. **Literature Review**

   One of the primary methods of obtaining requirements is through a literature review. This is done using the technology, domain, and existing work. A thorough literature review makes it simple to identify gaps and issues in the existing systems, and these gaps are highly beneficial for engineering requirements. As a result, a thorough literature review of the research domain, existing systems, and potential approaches and technologies was carried out. The drawback of this method to gather requirements is time costing.

2. **Interviews**

   Interviews can be considered as one of most common method of requirement elicitation. There are 2 types of interviews such as open-ended and structured. In open-ended interviews there is no pre-defined questions. Interviewer can ask context free questions to understand the problem. On other hand, in structured interviews there is predefined questions. Data Science Tech Lead, Senior Tech Lead – Software Engineering, Software Architect were among the interviewees. They should have in-depth knowledge of the system, procedure, priority, and other functional and non-functional requirements. Interviews enable us to focus more carefully on one another, which supports the ask follow-up questions and engage in in-depth discussions with the researcher. Therefore interviews are a suitable technique for this project.

3. **Observations**

   One of the most popular elicitation methods, observation of the existing solutions in this domain is crucial to determining the efficiency of the features currently provided and what the feature gaps are, which may then lead to new suggestions and needs for this research. The disadvantage of Observations is it is a time costly method.

4. **Brainstorming**

   In a brainstorming session, a group of people collect a variety of ideas in order to identify answers to specific challenges. Brainstorming sessions are intended to generate lots of new ideas hence providing a platform to share views. As a disadvantage, problems may arise as a result of several duplicate concepts and arguing notions.

4.5 DISCUSSION OF RESULTS

1. Literature Review

There are only limited number of research were done for Telco VAS customers' basket analysis to suggest a suitable service for the customers. Bagging method, stacking method haven't used in recent researches for predicting VAS services. Recent researches have not conducted based on subscribers in Sri Lanka domain. Performances are varied in previous researched when using different algorithms.

2. Interviews

Multiple domain experts are selected from the telecommunication industry to perform interviews. The focused telecommunication platform is Mobitel (Pvt) Ltd, which is a leading mobile service provider in Sri Lanka. The selected experts are Mr. Prabath Jayarathne (Software Architect, Mobitel (Pvt) Ltd), Mr. Veranga Pallawela (Senior Tech Lead, Mobitel (Pvt) Ltd), Mr. Nuwan Senevirathne (Tech Lead - Big Data & data Science, Mobitel (Pvt) Ltd), Mr. Rajitha Fernando (Data Engineer- Big Data & data Science, Mobitel (Pvt) Ltd). Following is an explanation of the expertise's insight.

Current Subscriber behaviour - Existing subscribers are more likely to leave their current service provider since the modern telecommunications industry is dynamic and offers customers enticing and competitive customer packages. Customers can be kept in service by providing desired VAS service according to their requirements. Moreover, this causes to the gain of profit.

Also, on the other hand, customers can get annoyed by suggestions/SMS of VAS services that they do not require. One of the major reasons for customer churning is unwanted messages or calls that they received by the service provider. This can be avoided by finding the services that they need without communicating them.

Data Acquisition - experts described the CDM, CDR, and ETL processes for recording subscriber data. Discuss about training models. Those discussions help select suitable algorithms. It was suggested by experts that the GUI be created user-friendly so that non-technical people could utilize the application.

Training Models - Discussions about prototype requirements, integrating methods, and training models for classification are held with the technical experts.

GUI - There have been discussions on making the GUI user-friendly so that non-technical people can use the application.

3. Observations

There are only few researches were carried out for basket analysis for Telco customers based on VAS services. When considering researches within last five years, there are no researches have conducted in this area in Sri Lanka domain. Additionally, in research studies conducted in research domain context, stacking and bagging are not used to create predictive models.

4. Brainstorming

Brainstorming sessions led to a number of significant project-related choices. The participants of brainstorming sessions are supervisor, fellow students, co-workers etc. The results of brainstorming sessions are identifying research gaps, algorithm selection, system design ideas, tools etc.

4.6 SUMMARY OF FINDINGS

Findings	**Lit.**	**Inter.**	**Obs.**	**Brain.**
Algorithms that can use to build and develop the model	✓		✓	
There is less research are conducted in this context	✓			

There is no research conducted in this context in Sri Lanka for last 5 years	✓	✓		
Anomalies in dataset	✓		✓	
System design ideas		✓		✓
Prototype should have a graphical user interface.		✓		
User interface should be simple, so all the general users can be familiar in quick time		✓		

Table 7 Summary of findings

4.7 CONTEXT DIAGRAM

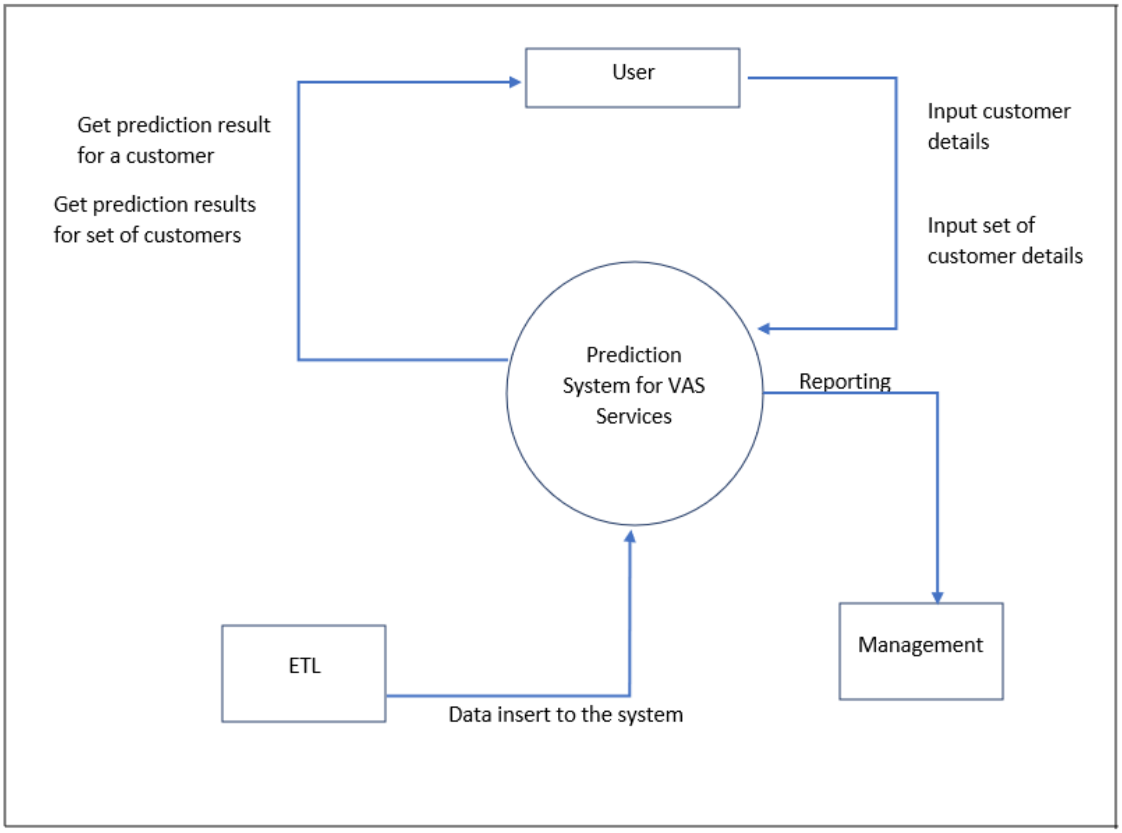

User
Get prediction result for a customer
Get prediction results for set of customers
Input customer details
Input set of customer details
Prediction System for VAS Services
Reporting
Management
ETL
Data insert to the system

Figure 4 Context Diagram

4.8 USE CASE DIAGRAM

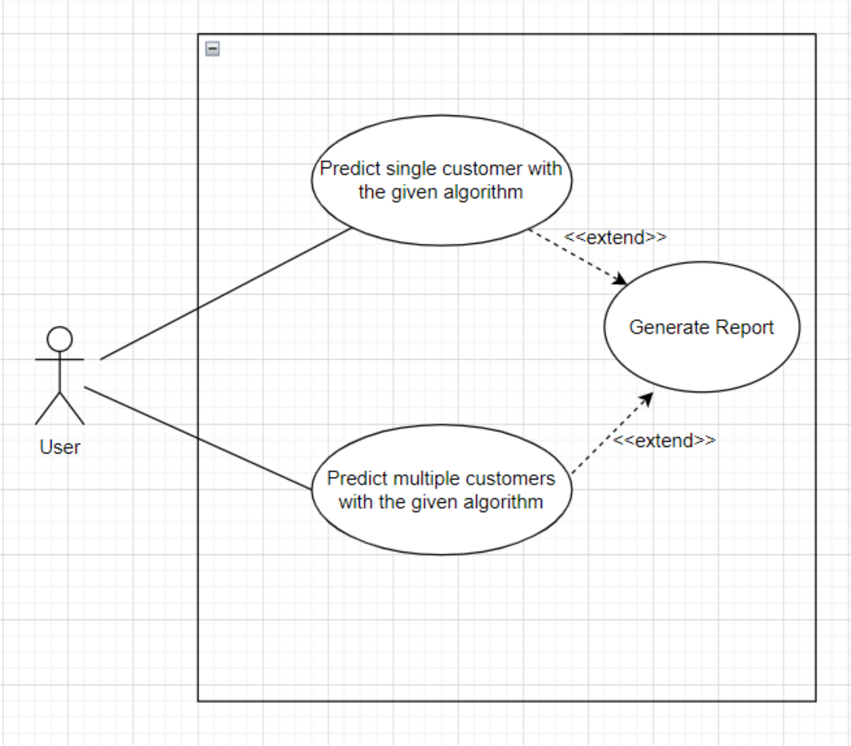

Predict single customer with the given algorithm
<<extend>>
Generate Report
User
<<extend>>
Predict multiple customers with the given algorithm

Figure 5 Use Case Diagram

4.8.1 USE CASE DESCRIPTION

Use case for predict single customer

Use case name	**Predict single customer with the given algorithm**
Description	User needs to predict VAS for a customer
Actors	Users who use the system
Success end condition	User successfully enter the customer details and get prediction result
Pre-conditions	The application must open correctly
Post conditions	The user must load the results.
Main Success Scenario	1. User insert customer details 2. User selects the algorithm 3. User clicks on the predict button

	4. System does the prediction and load the results
Variation	The user entered incorrect data. System shows a popup saying incorrect data
Exceptional flows	Any issue of this use case will affect to entire system

Table 8 Use Case Description 1

Use case for predict multiple customers

Use case name	**Predict multiple customers with the given algorithm**
Description	User needs to predict VAS for multiple customers
Actors	Users who use the system
Success end condition	User successfully upload the customer details in excel sheet and get prediction result
Pre-conditions	The application must open correctly
Post conditions	The user must load the results.
Main Success Scenario	1. User upload customer details as excel 2. User selects the algorithm 3. User clicks on the predict button 4. System does the prediction and load the results
Variation	The user entered incorrect data. System shows a popup saying incorrect data
Exceptional flows	Any issue of this use case will affect to entire system

Table 9 Use Case Description 2

4.9 REQUIREMENT SPECIFICATION

Priority levels of system requirements were defined using the MoSCoW technique, based on their importance.

Priority Level	Description
Must have (M)	These are core functional requirements of the prototype and is compulsory to be implemented.
Should have (S)	These requirements are not essential but will add value to the problem.
Could have (C)	These are optional and never considered essential to the scope of the project.
Will not have (W)	These requirements will not be available in the system and out of scope of the project

Table 10 MoSCoW priority levels

4.9.1 FUNCTIONAL REQUIREMENTS

FR#	**Requirement**	**Priority**
FR1	Dataset has to be pre-processed to remove dataset anomalies	M
FR2	Feature selection has to be performed to remove unnecessary features of the dataset	M
FR3	Data has to be split into two sets as train and test data	M
FR4	Implement several models using different algorithms	M
FR5	Measure the performances of the models	M
FR6	Select the best performing model	M
FR7	Users can access the trained model	M
FR8	User can enter customer details and predict VAS with the given algorithm	M

FR8	User can enter set of customer details and predict VAS with the given algorithm	M

Table 11 Functional requirements

4.9.2 NON-FUNCTIONAL REQUIREMENTS

NFR#	**Requirement**	**Priority**
NFR1	Predictions has to be accurate	M
NFR2	User friendly experience with GUI	S
NFR3	System has high security	M
NFR4	System has a user manual to guide users	C
NFR5	The GUI MUST run on the list of supported web browsers	M

Table 12 Non-functional requirements

4.10 CHAPTER SUMMARY

A rich picture is provided in the initial stage of the chapter to give a high-level idea about the proposed system. Stakeholder analysis has been provided with an onion diagram and descriptions. Observing existing systems, interviews, brainstorming and literature review were used as the requirement elicitation methods for the requirement gathering phase. Then Discussion of results and the summary of findings were discussed. Finally Functional and non-functional requirements were discussed.

CHAPTER 5: SOCIAL, LEGAL, ETHICAL AND PROFESSIONAL ISSUES

5.1 CHAPTER OVERVIEW

The purpose of this chapter is to define the mitigation for the social, legal, ethical, or professional difficulties that may arise during the project.

5.2 SLEP ISSUES AND MITIGATION

Social	Legal
• All interviewees were made aware at the commencement that the information they provided would be used for research and that what they said might be included in the project report. Interviewees' information was kept private, not even in the supporting documentation, if they only agreed to offer anonymous feedback. • Throughout the system's implementation, no data was gathered or sent to servers. The program runs locally on the development computer and won't need network access or open internet connectivity.	• With verbal consent to conduct the research, the dataset was legally obtained from a telecommunications service provider. • The obtained dataset was properly used to create the model, and it has not been released publicly. • Participants in the survey had their personal information and privacy properly protected.

• There are no violations of an ethical, political, or religious type in this research. Ethical	Professional
• The University of Westminster's academic conduct policies are followed when citing and referencing publications that are connected to the research. • The confidentially of the interviewers, evaluators, and industry professionals is preserved, and they will be told that the dissertation will include their comments. Their names weren't mentioned in the thesis, though.	• In the project's design and implementation phases, best practices for software engineering will be adhered to. • The project's outcomes were recorded using actual data that had not been altered. • The data, documentation, and source code used in this study are frequently backed up and secured as needed.

Table 13 SLEP Issues and Mitigation

5.3 CHAPTER SUMMARY

The main purpose of this chapter was to address social, legal, ethical and professional issues regarding the project and how those were resolved.

CHAPTER 6: SYSTEM ARCHITECTURE & DESIGN

6.1 CHAPTER OVERVIEW

Design aspects of the project from core of the system to the user interface are explored within this chapter. This describes the design decisions and diagrams including high level architecture, components, use case sequence, logical flows and class diagrams. The chapter also consists the UI wireframes designed to interact with the user.

6.2 DESIGN GOALS

Below are the design goals of the system

Design Goal	**Description**
Accuracy	In order to obtain accurate predictions, the final predictive models should produce good accuracy.
Scalability	Large datasets and newer, more sophisticated algorithms must be handled by the system without requiring extensive adjustments.
Performance	Performance is critically important that there should not be no errors, lags at either the logical end or the front end. Also, it is important to not take long for the output to be generated. Therefore, the system was designed not to end up with an unnecessarily a complex system.
Reusability	It is important for the components in the system to be built in a flexible way. So those components can be reused in the extensions and other developments.
Maintainability	Changes should be supported, and quick adaptability and modifiability should be achievable.

Table 14 Design Goals

6.3 SYSTEM ARCHITECTURE DESIGN

In the diagram below, the system architecture is illustrated. Here, the data, logic, and presentation layers all adhere to the three-tier architecture. This design follows to the modularity approach, which guarantees design objectives including scalability, maintainability, and reusability. While the other two layers are application oriented, the logic tier is the most concerned on the research.

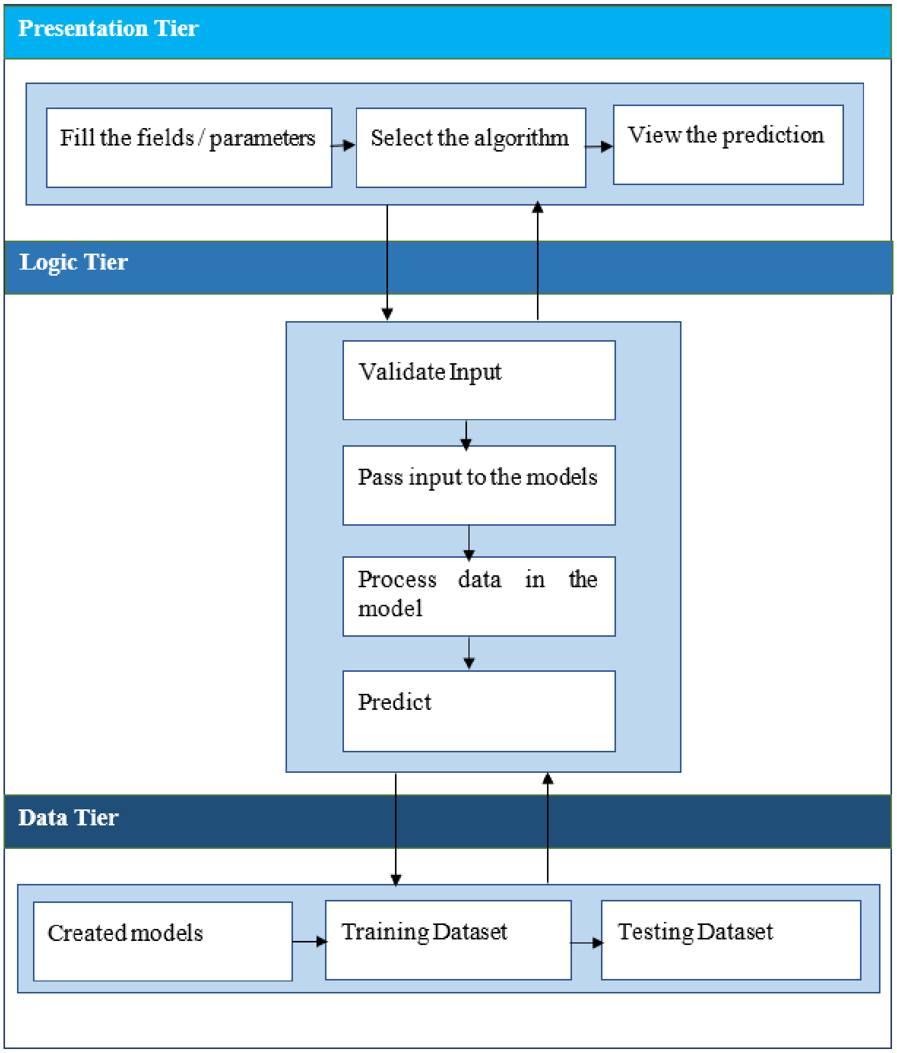


Figure 6 Three-tier system architecture

Data Tier: Data tier holds all the data required for the application and the Client Tier. Training models are created using the data set. Accuracy of the models depend on this dataset. Therefore, data tier is really important for the system success.

Logic Tier: Prediction is done in Logic tier using the data set. Predictive model and the prediction is based on the user inputs. User has a choice between trained models. Accuracies of trained models are displayed in UI, so that the user can select the desired algorithm.

Presentation Tier: Provide view to input the data and in order to predict the VAS service for a particular customer.

6.4 SYSTEM DESIGN

6.4.1 CHOICE OF DESIGN PARADIGM

CRISP-DM was used as a design paradigm to design this project. Cross-industry process for data mining is known as CRISP-DM. A structured way for planning a data mining project is provided by the CRISP-DM methodology. It is a reliable and well-proven methodology. There are 6 phases in this methodology such as business understanding, data understanding, data preparation, modelling, evaluation and deployment.

6.4.2 DATA FLOW DIAGRAM

The components that have been identified from the 3-tier architecture to be developed the data flow between each of them are used for data flow diagram. The link between system components and data flow is shown in below diagram.

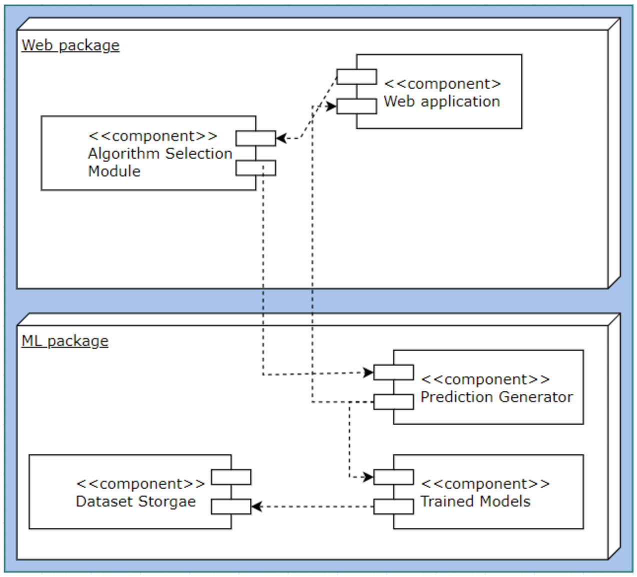


Figure 7 Data Flow Diagram

6.4.3 SEQUENCE DIAGRAM

Below sequence diagrams show the sequence of execution for predicting VAS service. In here user fills the feature data and the algorithm through the web interface. Those data are passed to the trained model and suitable VAS predicted. Generated VAS returned to the user through web interface.

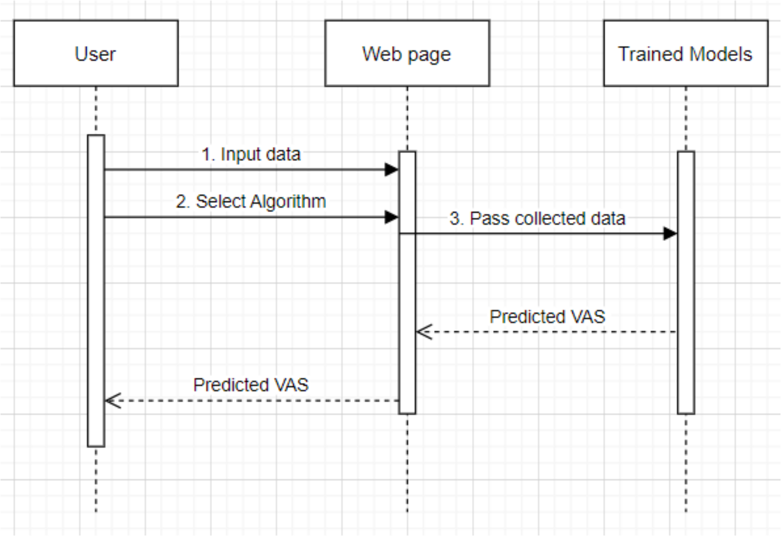


Figure 8 Sequence Diagram for single user

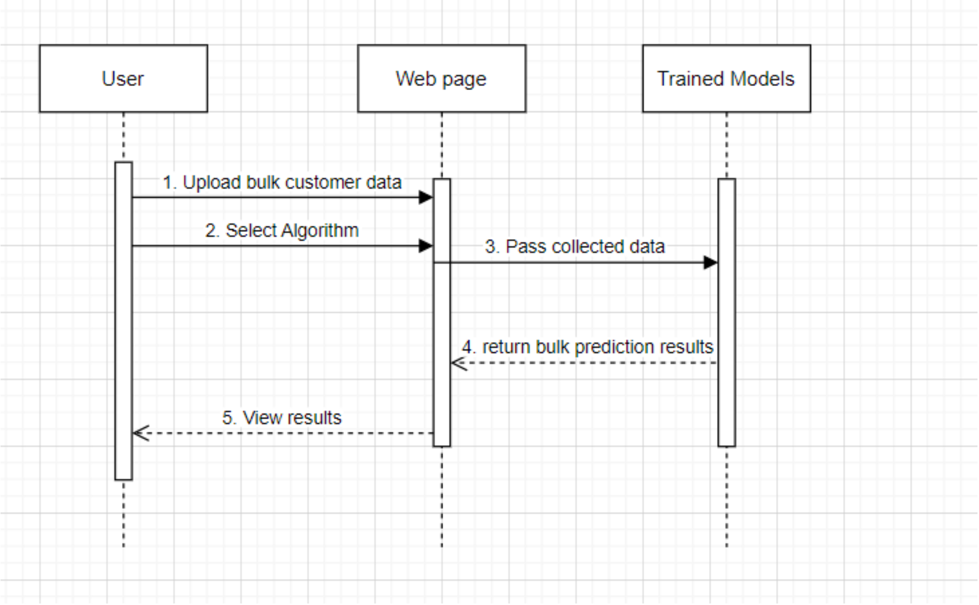


Figure 9 Sequence Diagram for multiple users

6.4.4 SYSTEM PROCESS DIAGRAM

The System Process Flowchart explains on the flow of the algorithms when the system predicts a VAS.

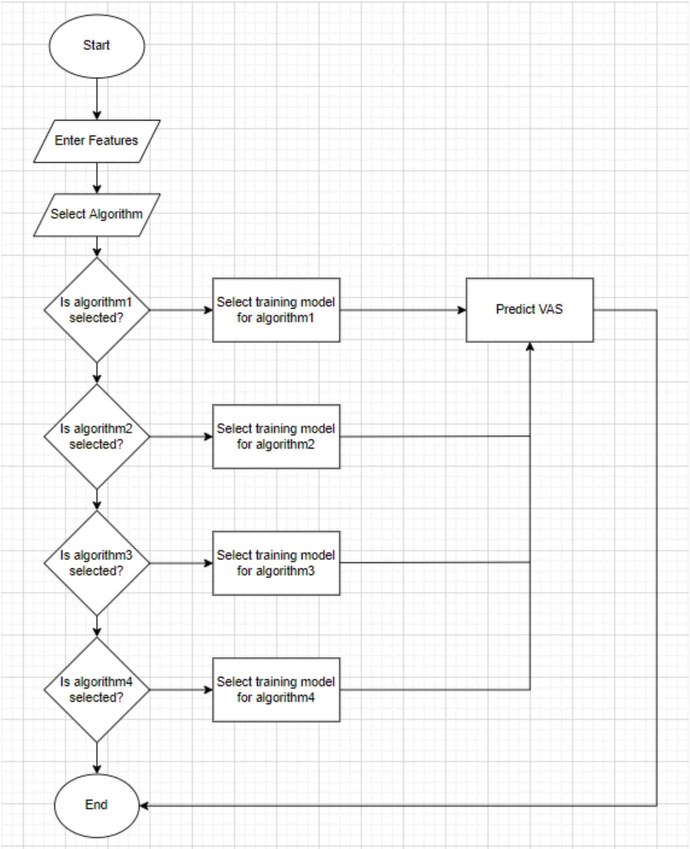

Start
Enter Features
Select Algorithm
Is algorithm1 selected?
Select training model for algorithm1
Predict VAS
Is algorithm2 selected?
Select training model for algorithm2
Is algorithm3 selected?
Select training model for algorithm3
Is algorithm4 selected?
Select training model for algorithm4
End

Figure 10 Decision structure flowchart

6.4.5 UI WIREFRAMES

In UI, there are interfaces which interact with the user in order to perform the prediction. UI consists of two tabs for single customer prediction and multiple customer prediction. Single customer prediction view allows users to input single customer details and get VAS prediction. Multiple customer view allows the user to upload multiple customer details as an excel file and get the prediction result.

6.5 CHAPTER SUMMARY

The system's architecture and design are described in this chapter. Discussions included the general system architecture, design goals, design paradigm, component diagrams, and sequence diagrams. The system user interfaces are then presented, followed by an illustration of the user flow.

CHAPTER 7: IMPLEMENTATION

7.1 CHAPTER OVERVIEW

The implementation of the suggested system is covered in this chapter. The main result of the combination of the literature review, the requirements that were gathered, and the design is this. It explains how the design choices were converted into executable code and discusses significant code segments from the system's essential parts as well as the research contribution.

7.2 TECHNOLOGY SELECTION

7.2.1 TECHNOLOGY STACK

Below are the technologies selected to be used in various aspects of the project.

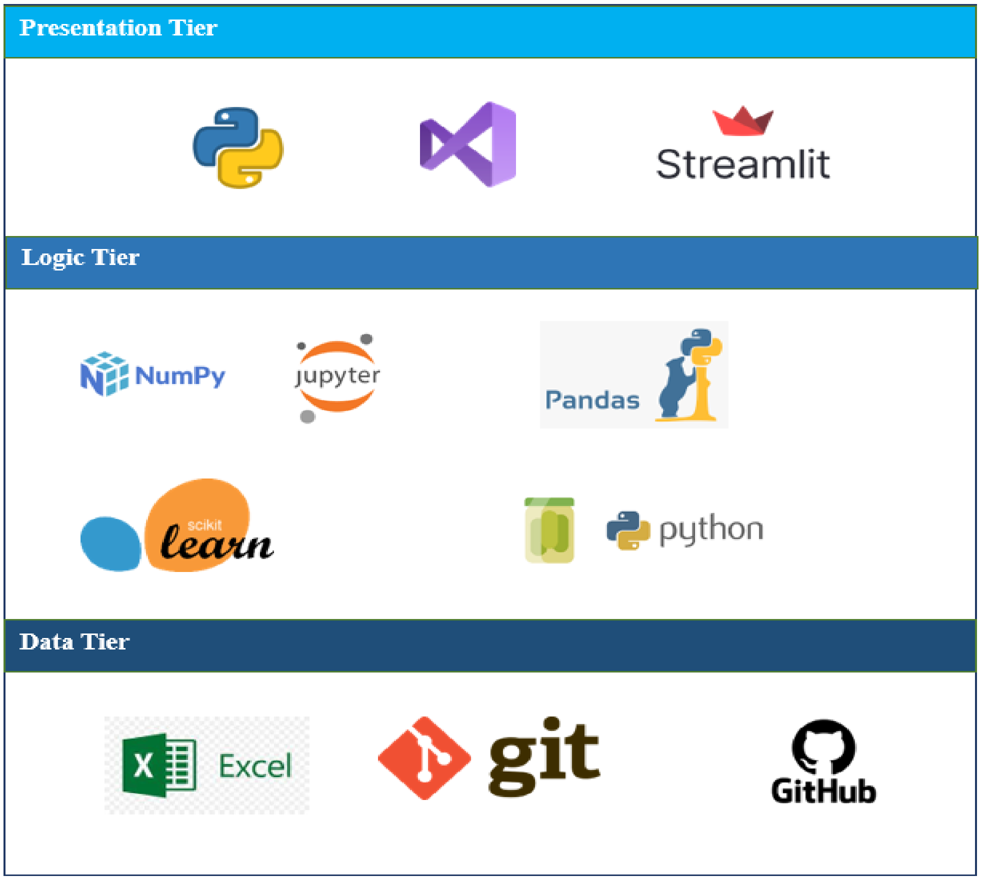


Figure 11 Technology Stack

7.2.2 DATA SELECTION

The main requirement of any data science project is data. Insufficient data will result in the project not being able to be developed successfully. Initially identified data requirements are,

1. Dataset must be acquired from a reliable source.
2. Dataset should be related to Sri Lankan domain
3. Dataset should be a real-world dataset

The author was able to get a comprehensive data set with 20000 records and 11 features from prominent telecommunication service provider in Sri Lanka. A process known as ETL (extraction, transform, and load) is being applied to the raw subscriber data. Obtained dataset also a transformed product from ETL process.

Dataset features are shown in below table

Feature	Description
package	Value added service
product_desc	Description of VAS
deducted	Deducted amount in Rupees for the sim
dateval	Date of data processed
gender	Male or Female
age_cat	Age of customer
con_age	Connection age of the sim
language_id	Language
total_voice_usage_min	Total voice usage
total_data_usage_mb	Total data usage
device_type	Device type
total_revenue	Total revenue
vas_revenue	Revenue of VAS services
customer_id	Id of the cutomer

Table 15 Feature description of the dataset

7.2.3 SELECTION OF THE DEVELOPMENT FRAMEWORK

Streamlit is selected as the graphical user interface development framework. It offers developers useful features and tools that make the framework incredibly adaptable and accessible. Since the scikit-learn framework is built on top of widely used libraries like numpy, matpotlib, and Scipy and makes data mining and analytical tasks simple and effective, therefore it is chosen for creating predictive models.

7.2.4 PROGRAMMING LANGUAGE

Python is used as the programming language for creating graphical user interface and predictive models. This choice is influenced by the fact that Python can be used to implement both the Core and the Backend API using a single programming language for quick development, while also benefiting from a wide range of frameworks and libraries, including PyTorch, AllenNLP, and FastAPI

7.2.5 LIBRARIES

Python has two main libraries for data modelling named SciKit and Tenserflow. Among them SciKit was selected for this project since Tenserflow mostly used in deep learning and neural networks while SciKit is generally used in general machine learning.

7.2.6 IDE

Since Jupyter Notebook IDE is a web-based interactive development environment and offers the freedom to build and arrange workflows in data science projects, it is chosen to be used with Python modeling. Since VSCode IDE is a cross-platform source code editor with a strong web development community and productivity-boosting features like syntax highlighting and auto-indentation, it is chosen for GUI development.

7.2.7 SUMMARY OF TECHNOLOGY SELECTION

Summary of technology selection is as follows.

Component	Tools /Technologies
Programming Language	Python
ML Library	sklearn
IDE	Streamlit

UI Framework	Jupyter Notebook, VSCode

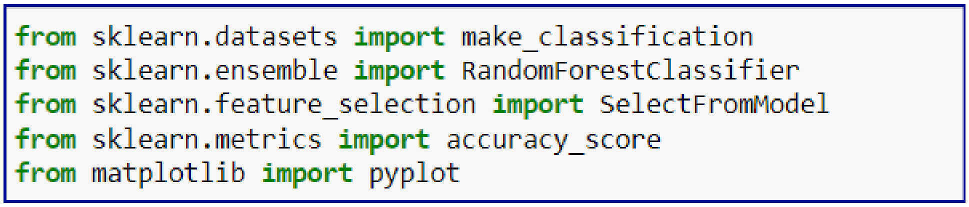
Table 16 Summary of Technology selection

7.3 IMPLEMENTATION OF CORE FUNCTIONALITIES

7.3.1 DATASET SELECTION

The original dataset consists of 13 features. Initially 3 features were eliminated such as customer_id, product_desc (Product description), deducted_amt (deducted amount). Customer_id does not provide any meaning to the dataset considering practical aspects. Moreover, product_desc and deducted_amt are fully dependant on the package. Therefore, those features are selected based on the package. Product description and deducted amount is pre-defined for the package. So, there is no connection between those features and other features for choosing a VAS service for a customer.

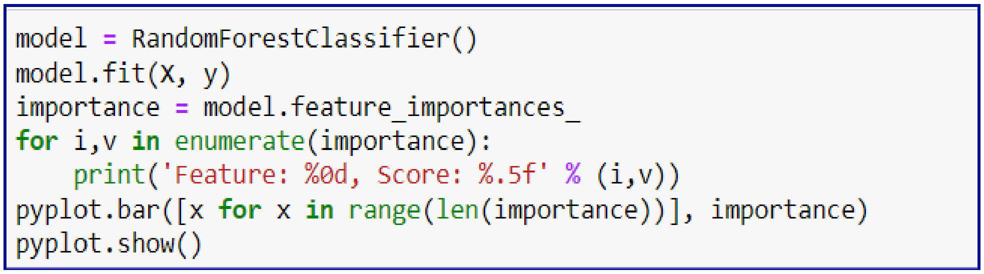
Feature importance technique was performed in order to identify less important features. The term "feature importance" relates to methods for scoring each input feature for a certain model; the scores merely indicate the "importance" of each feature. A higher score indicates that the characteristic will have more of an impact on the model being used to forecast a particular variable. In here, the feature importance technique was performed with the support of Random Forest classifier. Below are the libraries that are used to perform the feature importance technique.



```
from sklearn.datasets import make_classification
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import SelectFromModel
from sklearn.metrics import accuracy_score
from matplotlib import pyplot
```

Figure 12 Libraries used for feature engineering

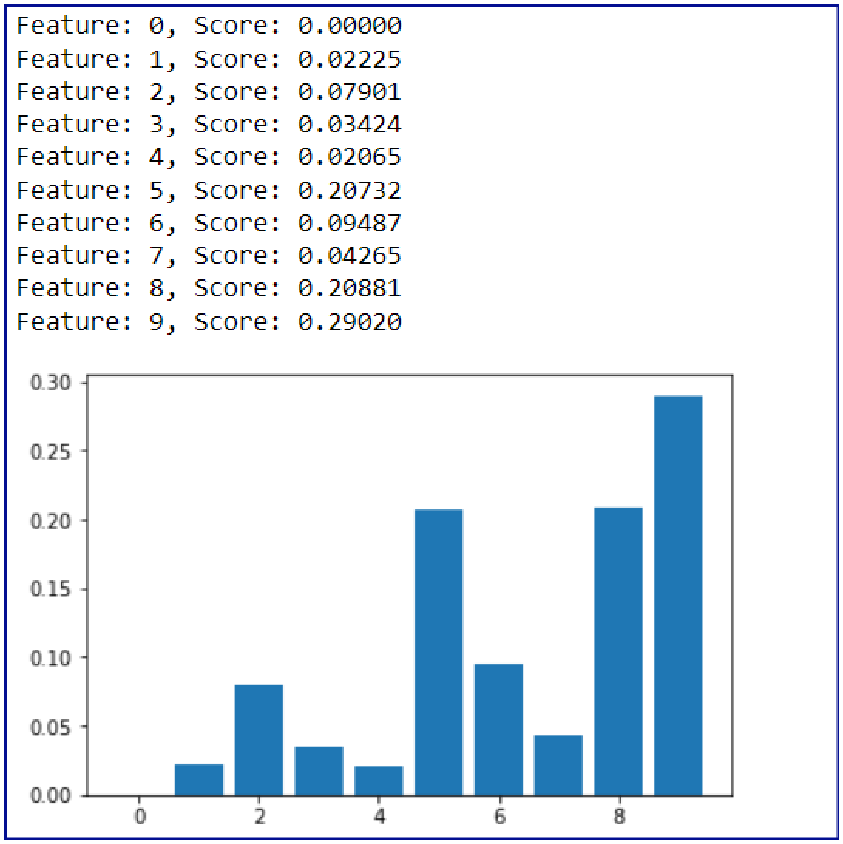
In order to generate importance below steps are followed.



```
model = RandomForestClassifier()
model.fit(X, y)
importance = model.feature_importances_
for i,v in enumerate(importance):
    print('Feature: %0d, Score: %.5f' % (i,v))
pyplot.bar([x for x in range(len(importance))], importance)
pyplot.show()
```

Figure 13 Feature engineering core code snippet

The output is as follows.



```
Feature: 0, Score: 0.00000
Feature: 1, Score: 0.02225
Feature: 2, Score: 0.07901
Feature: 3, Score: 0.03424
Feature: 4, Score: 0.02065
Feature: 5, Score: 0.20732
Feature: 6, Score: 0.09487
Feature: 7, Score: 0.04265
Feature: 8, Score: 0.20881
Feature: 9, Score: 0.29020
```

Figure 14 Feature importance diagram with score

According to the feature analysis, ‘dateval’ is eliminated from the dataset.

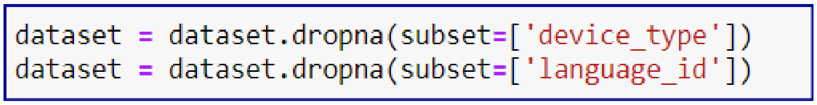
7.3.2 DATA PREPROCESSING

Preparing raw data to be appropriate for a machine learning model is known as data preprocessing. In order to build a machine learning model, it is one of crucial step. Real-world data typically includes noise, missing values, and may be in an unusable format, making it impossible to build machine learning models on it directly. Data preprocessing is necessary to clean the data and prepare it for a machine learning model, which also improves the model's accuracy and effectiveness.

Remove null Values

Data handling involves several crucial procedures, one of which is removing null values from the dataset. Any machine learning algorithm's performance and accuracy are negatively impacted by these null values. Therefore, before using any machine learning technique on the dataset, it is crucial to remove null values.

In order to remove null values below steps are followed



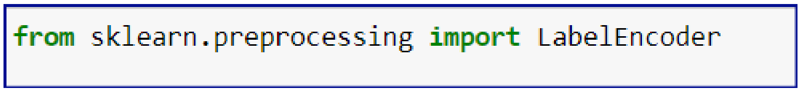
```
dataset = dataset.dropna(subset=['device_type'])
dataset = dataset.dropna(subset=['language_id'])
```

Figure 15 Remove null values

Label Encoding

Label encoding is the process of transforming labels into a numeric form so that they can be read by machines. The operation of those labels can then be better determined by machine learning techniques. It is a major supervised learning pre-processing step for the structured dataset.

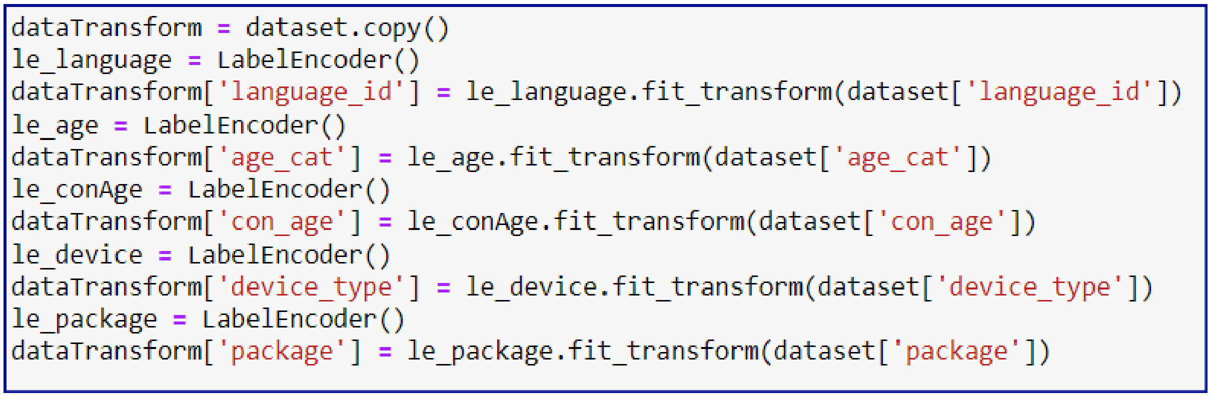
Below is the library that used to perform label encoding.



```
from sklearn.preprocessing import LabelEncoder
```

Figure 16 Libraries used for label encoding

Label encoding is done as follows.

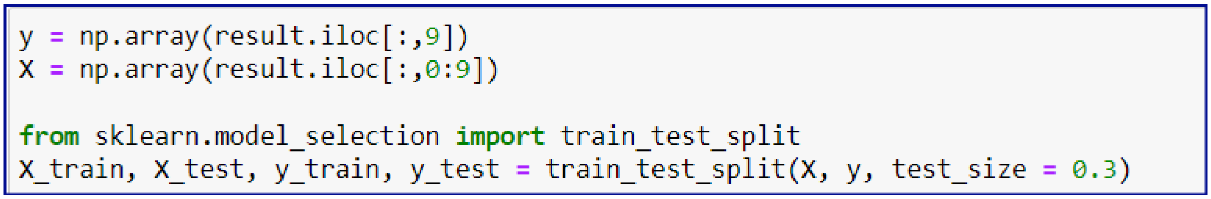


```
dataTransform = dataset.copy()
le_language = LabelEncoder()
dataTransform['language_id'] = le_language.fit_transform(dataset['language_id'])
le_age = LabelEncoder()
dataTransform['age_cat'] = le_age.fit_transform(dataset['age_cat'])
le_conAge = LabelEncoder()
dataTransform['con_age'] = le_conAge.fit_transform(dataset['con_age'])
le_device = LabelEncoder()
dataTransform['device_type'] = le_device.fit_transform(dataset['device_type'])
le_package = LabelEncoder()
dataTransform['package'] = le_package.fit_transform(dataset['package'])
```

Figure 17 Code snippet for label encoding

7.3.3 TRAINING MODELS

Datasets have been divided into 70:30 train: test ratios in order to execute predictive modelling using the chosen approaches. The splitting is seen in the following code segment.



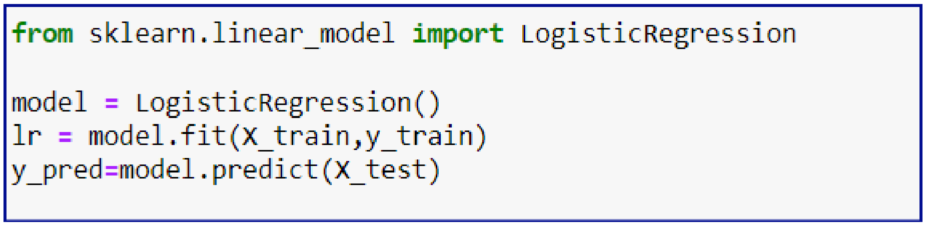
```
y = np.array(result.iloc[:,9])
X = np.array(result.iloc[:,0:9])

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3)
```

Figure 18 Code snippet for splitting data

Training models that are used as follows.

Logistic Regression

Logistic Regression was selected to build one of model. “LogisticRegression” in “sklearn” library is used for the training model.

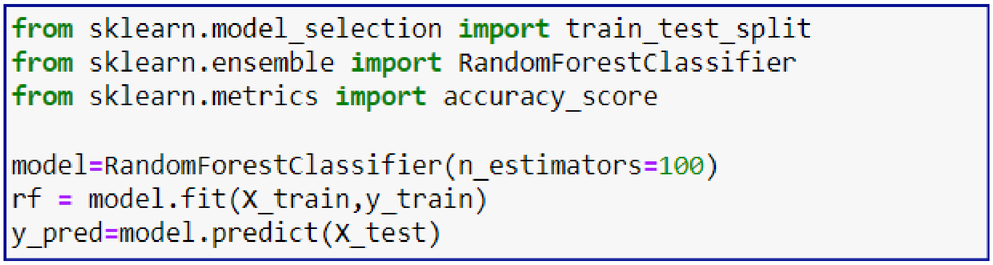


```
from sklearn.linear_model import LogisticRegression

model = LogisticRegression()
lr = model.fit(X_train,y_train)
y_pred=model.predict(X_test)
```

Figure 19 Code snippet for logistic regression

Random Forest

Random Forest algorithm was selected as the 2nd algorithm to build the model. “RandomForestClassifier” in “sklearn” library is used for the training model.

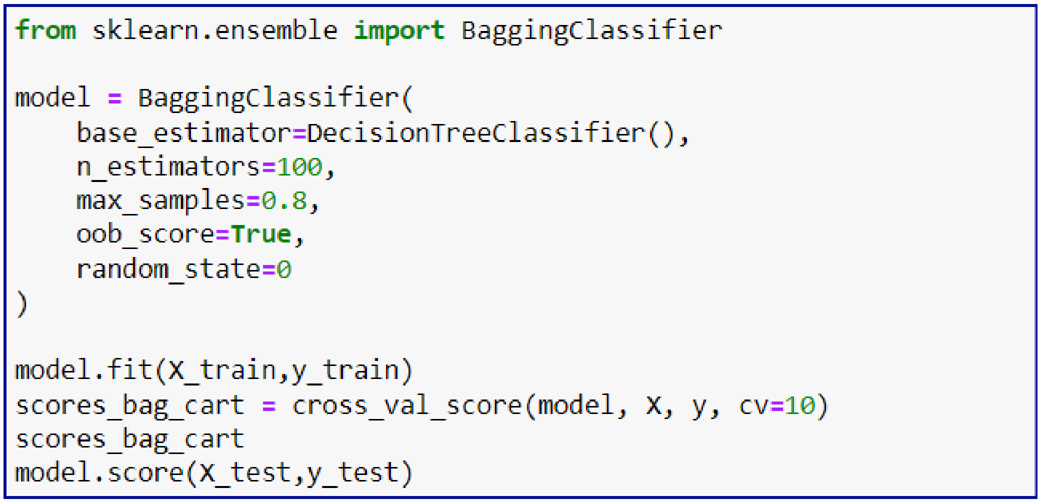


```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

model=RandomForestClassifier(n_estimators=100)
rf = model.fit(X_train,y_train)
y_pred=model.predict(X_test)
```

Figure 20 Code snippet for random forest

Bagged CART

Classification and Regression Trees algorithm was selected to build 3rd model. “DecisionTreeClassifier” in “sklearn” library is used for the training model.

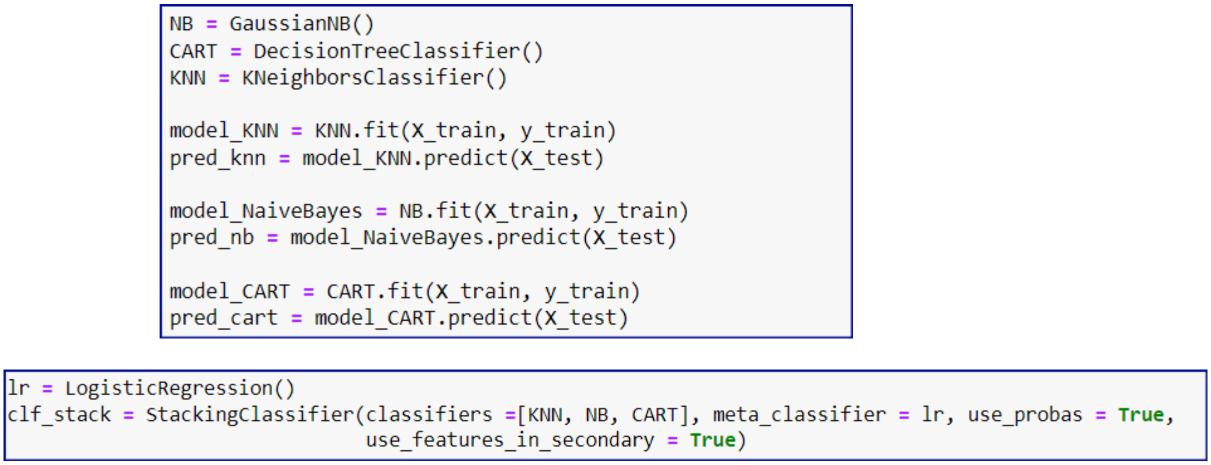


```
from sklearn.ensemble import BaggingClassifier

model = BaggingClassifier(
    base_estimator=DecisionTreeClassifier(),
    n_estimators=100,
    max_samples=0.8,
    oob_score=True,
    random_state=0
)

model.fit(X_train,y_train)
scores_bag_cart = cross_val_score(model, X, y, cv=10)
scores_bag_cart
model.score(X_test,y_test)
```

Figure 21 Code snippet for bagged cart

Stacking

Stacking ensemble technique was used to build 4th model. K-nearest neighbors algorithm, Naive Bayes algorithm, and CART algorithm were used as weak learners while Logistic regression algorithm was used as the meta learner. “sklearn” and “mlxtend” packages were used in here.



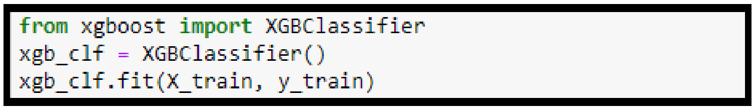
```
NB = GaussianNB()
CART = DecisionTreeClassifier()
KNN = KNeighborsClassifier()

model_KNN = KNN.fit(X_train, y_train)
pred_knn = model_KNN.predict(X_test)

model_NaiveBayes = NB.fit(X_train, y_train)
pred_nb = model_NaiveBayes.predict(X_test)

model_CART = CART.fit(X_train, y_train)
pred_cart = model_CART.predict(X_test)
```

```
lr = LogisticRegression()
clf_stack = StackingClassifier(classifiers =[KNN, NB, CART], meta_classifier = lr, use_probas = True,
                               use_features_in_secondary = True)
```

Figure 22 Code snippet for stacking

Boosting

By sequentially feeding datasets to the same algorithm while assigning weights to the data, the boosting ensemble technique creates different datasets. The aggregated final result will be provided by the sequence's final model. Here, a decision tree and XG boosting are used. Snippets of code showing the implementation are below.

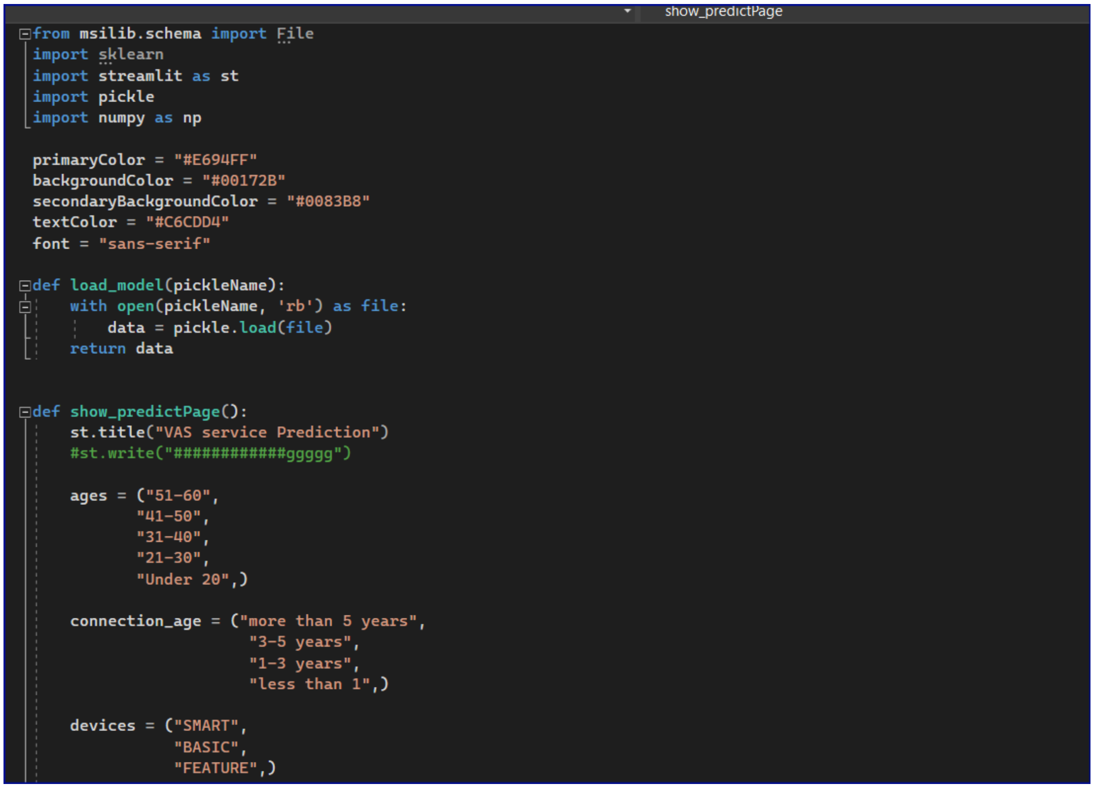


```
from xgboost import XGBClassifier
xgb_clf = XGBClassifier()
xgb_clf.fit(X_train, y_train)
```

Figure 23 Code snippet for boosting

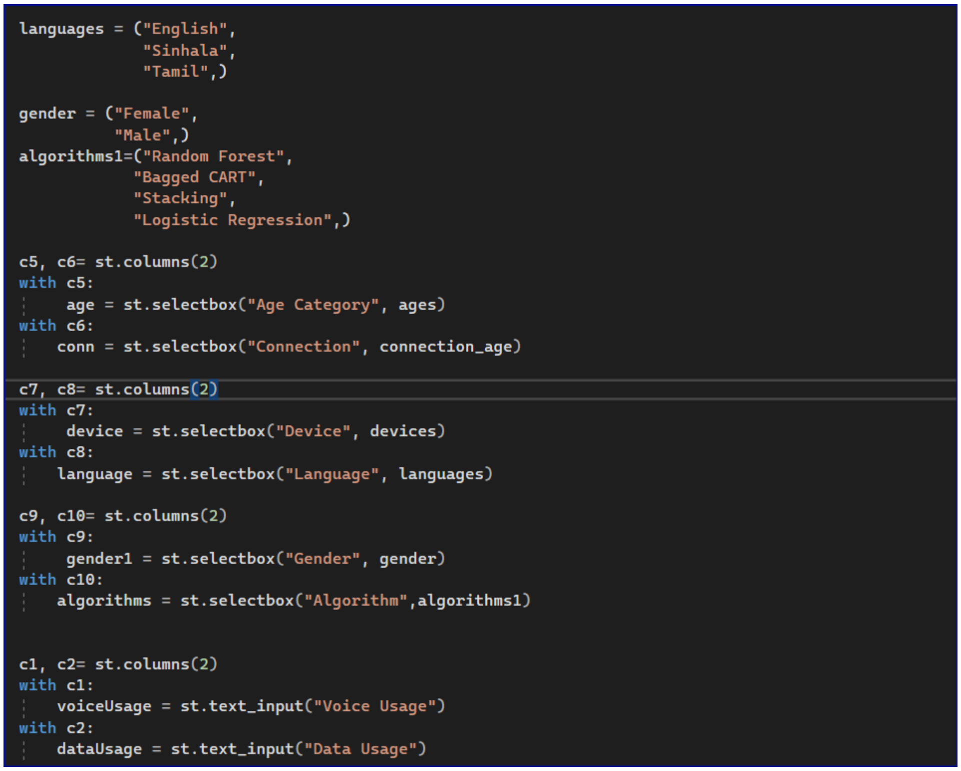
7.4 INTEGRATING MODELS WITH GUI

Basic HTML technology and the Streamlit framework are used to create the graphical user interface. The UI implementation of the prototype is demonstrated in the following code samples.



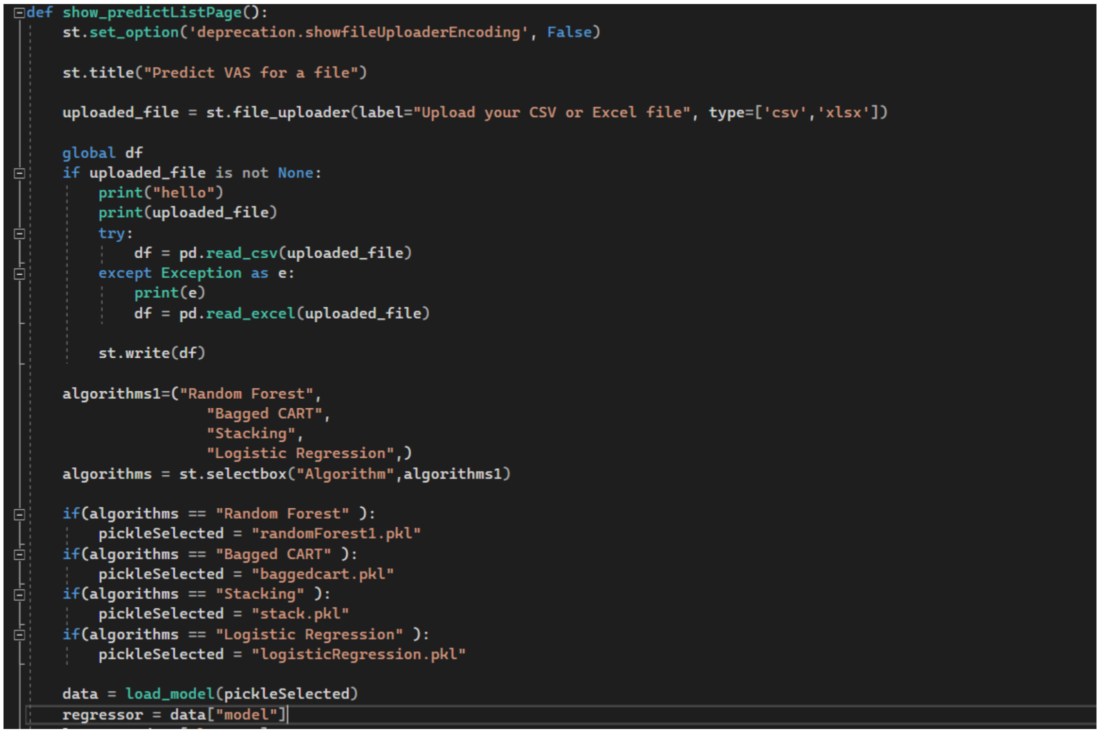
```
from msilib.schema import File
import sklearn
import streamlit as st
import pickle
import numpy as np

primaryColor = "#E694FF"
backgroundColor = "#00172B"
secondaryBackgroundColor = "#0083B8"
textColor = "#C6CDD4"
font = "sans-serif"

def load_model(pickleName):
    with open(pickleName, 'rb') as file:
        data = pickle.load(file)
    return data


def show_predictPage():
    st.title("VAS service Prediction")
    #st.write("############ggggg")

    ages = ("51-60",
           "41-50",
           "31-40",
           "21-30",
           "Under 20",)

    connection_age = ("more than 5 years",
                       "3-5 years",
                       "1-3 years",
                       "less than 1",)

    devices = ("SMART",
               "BASIC",
               "FEATURE",)
```

Figure 24 Predict page for Single customer



```
languages = ("English",
             "Sinhala",
             "Tamil",)

gender = ("Female",
          "Male",)
algorithms1=("Random Forest",
            "Bagged CART",
            "Stacking",
            "Logistic Regression",)

c5, c6= st.columns(2)
with c5:
     age = st.selectbox("Age Category", ages)
with c6:
    conn = st.selectbox("Connection", connection_age)

c7, c8= st.columns(2)
with c7:
     device = st.selectbox("Device", devices)
with c8:
    language = st.selectbox("Language", languages)

c9, c10= st.columns(2)
with c9:
     gender1 = st.selectbox("Gender", gender)
with c10:
    algorithms = st.selectbox("Algorithm",algorithms1)


c1, c2= st.columns(2)
with c1:
    voiceUsage = st.text_input("Voice Usage")
with c2:
    dataUsage = st.text_input("Data Usage")
```

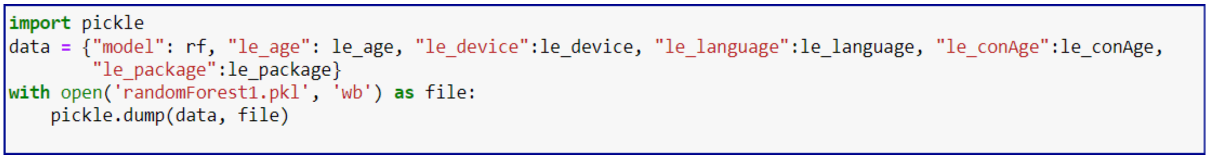
Figure 25 Predict page for Single customer 2



```
def show_predictListPage():
    st.set_option('deprecation.showfileUploaderEncoding', False)

    st.title("Predict VAS for a file")

    uploaded_file = st.file_uploader(label="Upload your CSV or Excel file", type=['csv','xlsx'])

    global df
    if uploaded_file is not None:
        print("hello")
        print(uploaded_file)
        try:
            df = pd.read_csv(uploaded_file)
        except Exception as e:
            print(e)
            df = pd.read_excel(uploaded_file)

        st.write(df)

    algorithms1=("Random Forest",
                 "Bagged CART",
                 "Stacking",
                 "Logistic Regression",)
    algorithms = st.selectbox("Algorithm",algorithms1)

    if(algorithms == "Random Forest" ):
        pickleSelected = "randomForest1.pkl"
    if(algorithms == "Bagged CART" ):
        pickleSelected = "baggedcart.pkl"
    if(algorithms == "Stacking" ):
        pickleSelected = "stack.pkl"
    if(algorithms == "Logistic Regression" ):
        pickleSelected = "logisticRegression.pkl"

    data = load_model(pickleSelected)
    regressor = data["model"]
```

Figure 26 Predict page for multiple customers

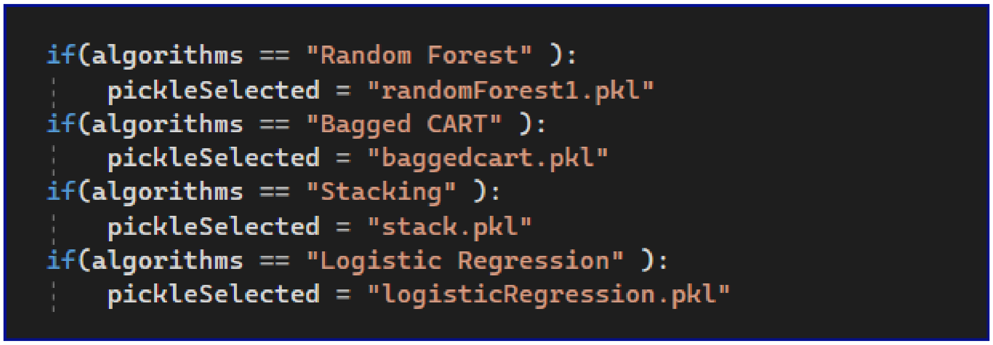
The predictive models created with Python were used for the UI integration. Using the "Pickle" library, trained models were exported. The trained models were exported and saved using the following code snippet.



```
import pickle
data = {"model": rf, "le_age": le_age, "le_device":le_device, "le_language":le_language, "le_conAge":le_conAge,
        "le_package":le_package}
with open('randomForest1.pkl', 'wb') as file:
    pickle.dump(data, file)
```

Figure 27 Code snippet for creating pickle

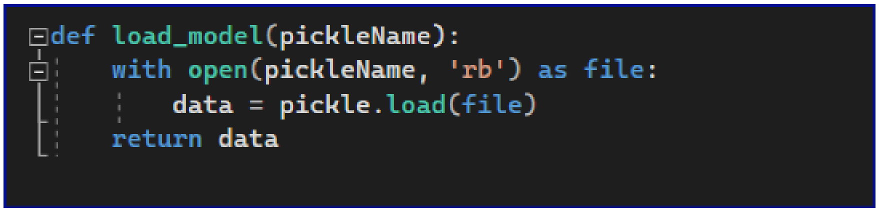
Below code snippet shows how the pickle is selected according to the algorithm that the customer selected from the UI



```
if(algorithms == "Random Forest" ):
    pickleSelected = "randomForest1.pkl"
if(algorithms == "Bagged CART" ):
    pickleSelected = "baggedcart.pkl"
if(algorithms == "Stacking" ):
    pickleSelected = "stack.pkl"
if(algorithms == "Logistic Regression" ):
    pickleSelected = "logisticRegression.pkl"
```

Figure 28 Code snippet for selecting pickle

Below code snippet shows how the pickle is imported to the python project. In there, le_age, le_device, le_language, le_conAge, le_package are label encoders.

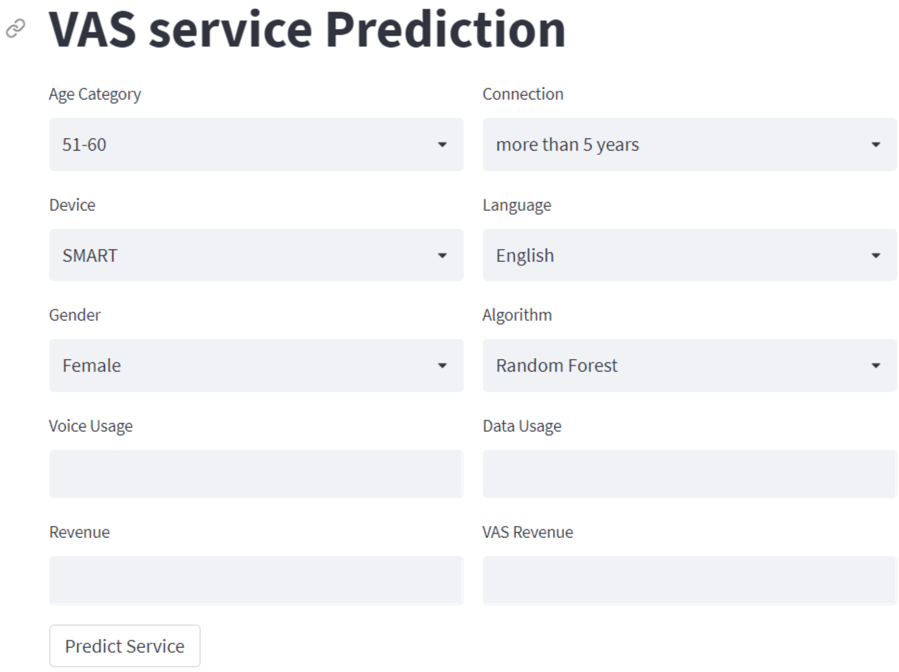


```
def load_model(pickleName):
    with open(pickleName, 'rb') as file:
        data = pickle.load(file)
    return data
```

Figure 29 Code snippet for loading pickle

GUI of the final prototype is shown in below figure.

Single Customer Prediction GUI



VAS service Prediction

Age Category: 51-60

Connection: more than 5 years

Device: SMART

Language: English

Gender: Female

Algorithm: Random Forest

Voice Usage:

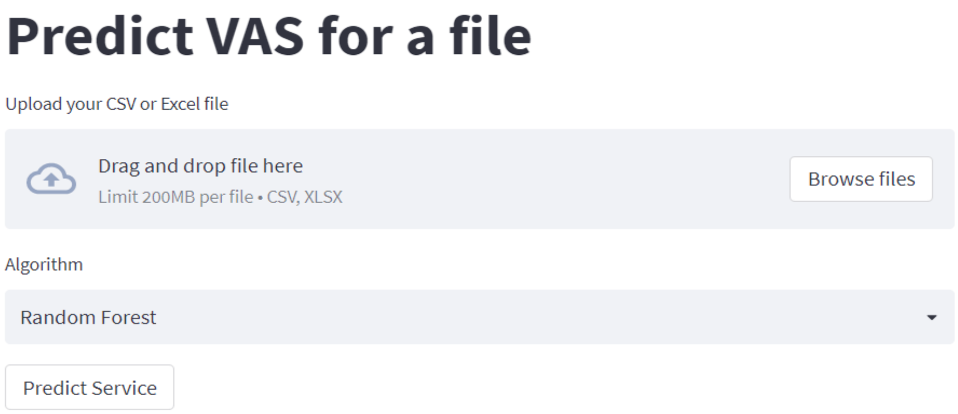
Data Usage:

Revenue:

VAS Revenue:

Figure 30 Single customer prediction GUI

Multiple Customer Prediction GUI



Predict VAS for a file

Upload your CSV or Excel file

Drag and drop file here

Limit 200MB per file • CSV, XLSX

Algorithm

Random Forest

Figure 31 Multiple customer prediction GUI

7.5 CHAPTER SUMMARY

This chapter covered the technology stack, programming language, libraries, UI Framework, and IDE used in the implementation of this system. Next, the core functionality implementation is described including data analyzing, preprocessing, model training and GUI integration. A few functional snippets of code created for the core of the system were displayed. Finally, the main user interfaces are shown.

CHAPTER 8: TESTING

8.1 CHAPTER OVERVIEW

This chapter provides an overview of the testing techniques used to make sure the system is functioning as planned. There were both functional and non-functional tests run, and the test plan and test criteria were addressed. Along with benchmarking results, unit testing and integration testing have been carried out.

8.2 OBJECTIVES AND GOALS OF TESTING

The main goal of testing the software product is verify that the system is operating in accordance with the anticipated expectations by the gathered requirements. The main objectives of the system's testing process are listed below.

- Identify the system's bugs and flaws.
- The system must be checked before being published to the production server to ensure that it is free of any significant issues.
- Verify that each functional requirement has been met.
- Verify that each non-functional requirement has been met.

8.3 TESTING CRITERIA

Here, testing is done module-by-module, allowing the researcher to test even the smallest components. Model testing is done by benchmarking, functional testing, integration testing, unit testing and non-functional testing.

8.4 MODEL TESTING

To evaluate the performance, the model is tested first. Confusion matrix and Receiver Operating Characteristic curve (ROC) are used for model evaluation. Accuracy, precision, recall and F1 score are considered to evaluate the performance in the model.

"Bagging CART" is selected as the higher performance algorithm among other models which were trained using Random Forest, Stacking and Logistic Regression.

8.4.1 CONFUSION MATRIX

Here is the confusion matrix which is obtained for the model.

Prediction	**Reference**		
	B2W	**CRBT_SRV**	**MQSM**
B2W	44	671	19
CRBT_SRV	108	4682	107
MQSM	16	408	421

Table 17 Confusion Matrix

Here test dataset consists of 6476 records while 'CRBT_SRV' package has highest number of records.

To measure the performance, following metrics are used.

1. True Positive (TP): Refers the number of predictions where the classifier correctly predicts the positive class as positive
2. True Negative (TN): Refers the number of predictions where the classifier correctly predicts the negative class as negative
3. False Positive (FP): Refers the number of predictions where the classifier incorrectly predicts the negative class as positive
4. False Negative (FN): Refers the number of predictions where the classifier incorrectly predicts the positive class as negative

There are no negative or positive classes in this matrix unlike the binary classification. Therefore TP, TN, FP, FN have to be calculated for each individual class.

	B2W	CRBT_SRV	MQSM
TP	44	4682	421
TN	5618	500	5505
FP	690	215	424
FN	124	1079	126

Table 18 Confusion Matrix 2

8.4.1.1 ACCURACY

The accuracy of the model is defined as the ratio of correct classifications (true positives and true negatives) to the total number of cases. The model's accuracy is determined using the equation below.

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

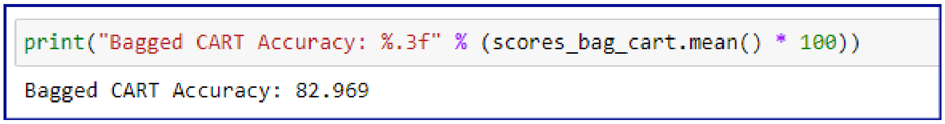
Accuracy of each class can be calculated as follows.

Accuracy for B2W = $\frac{44 + 5618}{44+5618+690+124} = \frac{5662}{6476} * 100\% = 87.43\%$

Accuracy for CRBT_SRV = $\frac{4682+500}{4682+500+215+1079} = \frac{5182}{6476} * 100\% = 80.01\%$

Accuracy for MQSM = $\frac{421+5505}{421+5505+424+126} = \frac{5927}{6476} * 100\% = 92.27\%$

Overall accuracy of the model can be achieved by below code snippet



```
print("Bagged CART Accuracy: %.3f" % (scores_bag_cart.mean() * 100))
```

```
Bagged CART Accuracy: 82.969
```

Figure 32 Code snippet for accuracy

Therefore, overall accuracy 82.97% accuracy is achieved, and it is very significant achievement.

8.4.1.2 PRECISION

The precision of the model is defined as the fraction of predictions as a positive class are actually positive. The model’s precision can be calculated using below equation.

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

Precision of each class can be calculated as follows

Precision for B2W = $\frac{44}{44+690}$ = 0.05

Precision for CRBT_SRV = $\frac{4682}{4682+215}$ = 0.95

Precision for MQSM = $\frac{421}{421+424}$ = 0.50

Macro average precision = $\frac{0.05+0.95+0.50}{3}$ = 0.5

8.4.1.3 RECALL

The recall of the model is defined as the fraction of all positive samples were correctly predicted as positive by the classifier. The model's recall can be calculated using below equation.

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

Recall of each class can be calculated as follows.

Recall for B2W = $\frac{44}{44+124}$ = 0.26

Recall for CRBT_SRV = $\frac{4682}{4682+1079}$ = 0.81

Recall for MQSM = $\frac{421}{421+126}$ = 0.77

Macro average recall = $\frac{0.26+0.81+0.77}{3}$ = 0.61

8.4.1.4 F1-SCORE

The F1-score of the model combines precision and recall into a single measure. It's harmonic mean and of precision and recall. The model's F1-score can be calculated using below equation.

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

F1-Score of each class can be calculated as follows

F1-Score for B2W = $2 * \frac{0.05*0.26}{0.05+0.26} = 0.08$

F1-Score for CRBT_SRV = $2 * \frac{0.95*0.81}{0.95+0.81} = 0.87$

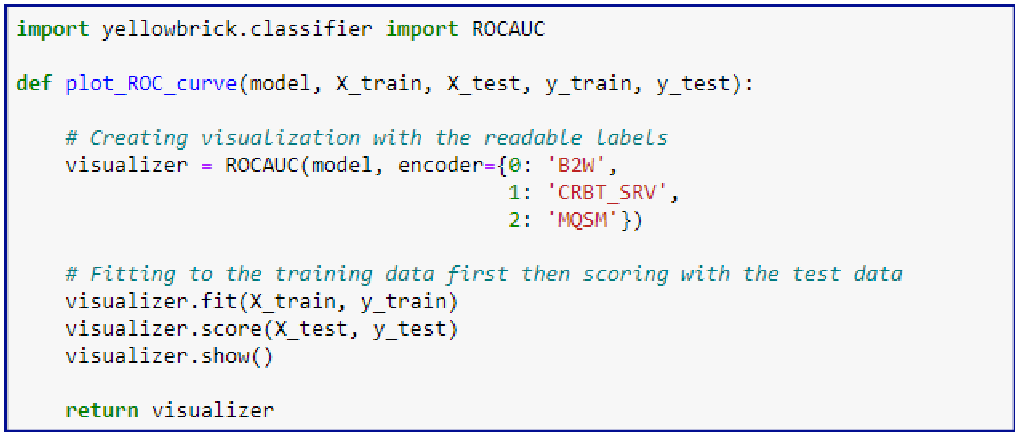
F1-Score for MQSM = $2 * \frac{0.50*0.77}{0.50+0.77} = 0.61$

Macro average F1-score = $\frac{0.08+0.87+0.61}{3} = 0.52$

8.4.2 AUC-ROC

The performance of a model in terms of sensitivity and specificity can be seen using an AUC ROC (Area Under the Curve Receiver Operating Characteristics) plot. Sensitivity is the capacity to recognize accurately entries that belong to the positive class. The capacity to correctly recognize elements that belong in the negative class is referred to as specificity. an AUC ROC plot can help to identify how well the model is able to distinguish between classes.

For multiclass problems, ROC curves can be plotted by one class versus others. The AUC score can be calculated for each class separately. In this model, Yellowbrick library is used to plot AUC ROC plot.



```
import yellowbrick.classifier import ROCAUC

def plot_ROC_curve(model, X_train, X_test, y_train, y_test):

    # Creating visualization with the readable labels
    visualizer = ROCAUC(model, encoder={0: 'B2W',
                                        1: 'CRBT_SRV',
                                        2: 'MQSM'})

    # Fitting to the training data first then scoring with the test data
    visualizer.fit(X_train, y_train)
    visualizer.score(X_test, y_test)
    visualizer.show()

    return visualizer
```

Figure 33 Code snippet for AUC-ROC

Output is as follows

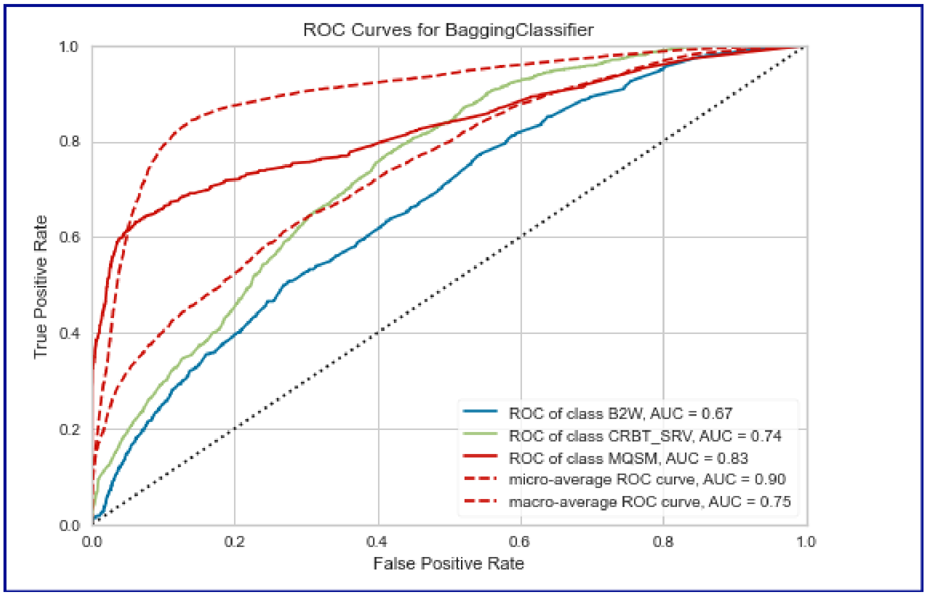


Figure 34 AUC ROC curve

According to the figure, the model can be considered as a good one because it starts at 0 FPR, quickly catches up to about 83% for MQSM, 74% for CRBT_SRV and 67% for B2W of TPR, and then increases gradually toward maximum TPR.

8.5 BENCHMARKING

A benchmarking method is a test that establishes a comparison to give information about the effectiveness or performance of the system that was designed and built. Mainly There are three ways to conduct benchmarking, as follows.

- Competitive Benchmarking: comparing with existing systems
- Functional Benchmarking: comparing with manual traditional methods
- Internal Benchmarking: comparing with predicted and actual results

This project is based on telco customer VAS service prediction model and there are only limited number of such models available (Only found 1 such model). Unfortunately, couldn’t find such free systems to test and compare the results with developed system. Therefore, competitive benchmarking couldn’t be carried out.

As functional and internal benchmarking, dataset was trained using different algorithms and achieved different performances. Performances of the models are shown in below table.

Algorithm	**Package**	**Accuracy**
KNN	KNeighborsClassifier	72.31%
NB	GaussianNB	74.23%
Logistic Regression	LogisticRegression	76.85%
Random Forest	RandomForestClassifier	75.18%
Stacking	KNeighborsClassifier, GaussianNB, DecisionTreeClassifier, LogisticRegression, StackingClassifier	73.70%
Bagged CART	DecisionTreeClassifier	82.97%
XG Boost	XGBClassifier	82.84%

Table 19 Performances

Below are the algorithms that were used to train data.

- K-nearest neighbors algorithm
- Naïve Bayes algorithm
- Logistic regression algorithm
- Random forest algorithm
- Stacking with KNN, NB and CART weak learners
- Classification and regression trees with bagging (Bagged CART)
- XG Boost

Following a comparison with various models and consideration of the results, the "Bagged CART" algorithm is chosen for the model.

8.6 FUNCTIONAL TESTING

Test case	Description	Input data	Expected outcome	Actual outcome	Status
1	The model training with the selected training configuration	Data set and training configurations are provided	The model needs to be trained using the specified training configuration.	The model is trained using the specified training configuration.	Pass
2	The trained model must be able to export and save.		The trained prediction model needs to be usable and operate similarly.	A trained prediction model is usable and operate similarly.	Pass
3	Take inputs and make the prediction	Fill data in ‘predict_page’ and click ‘Predict’ button	Display the prediction	Display the prediction	Pass
4	Able to select the preferred prediction algorithm	Click on ‘algorithms’ dropdown box	Show several algorithms	Show several algorithms	Pass
5	Ability to modify the algorithm without having to re-enter feature data	Change algorithm after predicting the result	Show previously filled data and has the ability to make prediction again using	Show previously filled data and has the ability to make prediction	Pass

			different algorithm	again different algorithm	
6	Measure the performance of selected model	Fill input data and click on ‘Predict’ button	Show accuracy of each model	Show accuracy of each model	Pass

Table 20 Functional Testing

8.7 INTEGRATION TESTING

Test case	**Description**	**Input data**	**Expected outcome**	**Actual outcome**	**Status**
1	Pre-processing data set	Row data set	Clean data set which does not have anomalies	Clean data set which does not have anomalies	Pass
2	Training models	Pre-processed data	Trained model	Trained model	Pass
3	Saving models	Trained model	Can be loaded to the backend of GUI	Can load to the backend of GUI	Pass
4	Loading models	Saved model	Can use for frontend of GUI	Can use for frontend of GUI	Pass
5	Predicting results for different algorithms	Data that taken from the user through GUI	Display predicted result	Display predicted result	Pass
6	Selecting algorithm	Selected algorithm from dropdown box	User can select the preferred algorithm	User can select the preferred algorithm	Pass

Table 21 Integration Testing

8.8 NON FUNCTIONAL TESTING

8.8.1 SYSTEM SPECIFICATION

Specifications that are used to test non functional requirements are as follows.

CPU	11th Gen Intel(R) Core (TM) i7-11800H @ 2.30GHz
Memory	32.0 GB (31.8 GB usable)
OS	Windows 10
GPU	Intel UHD Graphics 620

Table 22 System Specifications

8.8.2 ACCURACY TESTING

The models that were used in the system completed an accuracy test, and the results are presented in the following table.

Algorithm	**Accuracy**
Naïve Bayes	74.23%
K-nearest neighbors	72.31%
Logistic Regression	76.85%
Random Forest	75.18%
Stacking	73.70%
Bagged CART	82.97%
XG Boost	82.84%

Table 23 Accuracy Testing

8.8.3 PERFORMANCE TESTING

Time that are taken for training models are analyzed for each algorithm and results are shown in below table.

Algorithm	**Model train time (S)**
Naïve Bayes	7
K-nearest neighbors	8
Logistic Regression	8

Random Forest	10
Stacking	10
Bagged CART	67
XG Boost	16

Table 24 Performance Testing

8.8.4 USABILITY TESTING

Non-technical individuals performed usability testing on the GUI to assess its usability and user-friendliness. Although the feedback was mostly positive, several improvements were recommended.

8.8.4 COMPATIBILITY TESTING

Edge, Google Chrome and Mozilla Firefox, 3 of the supported browsers, are manually checked for compatibility concerns with the Frontend Web application. It is confirmed that there are no significant problems.

8.9 LIMITATIONS OF TESTING PROCESS

Limitations of the testing process are,

- One machine is used to test the models. To get a clear measurement of the performances, it is preferable to test the models on various machines with various specifications.
- Limit to one dataset. It is better if different datasets can be got from different service providers to measure the performance of the system

8.10 CHAPTER SUMMARY

The chapter's opening discussion covered the goals and objectives of testing as well as the testing standards. Based on the confusion matrix, accuracy, precision, recall, and ROC, model testing was carried out and discussed. System underwent benchmarking using the used algorithms. Then integration testing was used and finally, the limitations of testing process were discussed.

CHAPTER 9: EVALUATION

9.1 CHAPTER OVERVIEW

This chapter explains how the system's many aspects as importance, development and usability were assessed by domain experts and the intended audience. The author’s self-assessment is given after their evaluation. It is also mentioned if the system's original goals were attained.

9.2 EVALUATION METHODOLGY AND APPROACH

A project's evaluation serves as a measurement of its success. Feedback was gathered on the research challenge, the project's design, development, testing, and other key aspects. This system was designed and implement to support Telco companies by suggesting VAS services for their customers. Both technical and non-technical users will be using this system. Due to the system is based on several algorithms and complex logics, the end user may find it challenging to understand its core functionalities. Therefore, system should be straightforward to use.

Both domain and technical specialists from the telecommunications and machine learning fields reviewed the system.

9.3 EVALUATION CRITERIA

To evaluate the implemented system and get accurate feedback on it, a quantitative evaluation approach was used. The table below displays the evaluation criteria.

Evaluation Criteria	**Purpose of Evaluation**
The overall notion of the project	This criterion covers the overall concept of the project
Literature survey	This is to confirm that a thorough review of the literature on the subject and the relevant technologies was done before a gap in the literature was discovered.

System design and implementation	This is to get feedback regarding the system's-maintained standards
Prototype	This criterion entails to Verify the performance of the implemented system as a proof-of-concept for the suggested solution.
Difficulty of the final application	This is to analyze the non-functional requirements for the system.
GUI and user experience of the application	This is to analyze the non-functional requirements for the system.

Table 25 Evaluation Criteria

9.4 SELF EVALUATION

Self-evaluation has been used to discuss an author's assessment of their own work. The author chose to do an analysis based on topics related to the research. Based on the evaluation criteria specified in the previous section, a self-evaluation was also completed.

Evaluation Criteria	Self-evaluation by the author
The overall notion of the project	The concept of the project is crucial because it is quite uncommon to find research on VAS services for telecom subscribers, particularly in the Sri Lankan domain. Additionally, as an employee of prominent telecommunication company in Sri Lanka, author identifies how important to suggest the suitable VAS service for suitable customers to gain maximum benefit.
Literature survey	Initially a thorough literature survey was done on market basket analysis and VAS service prediction works to find gaps that exists in the domain.
System design and implementation	It can be rated as a good system considering the entire system. GUI and architectural improvements can be done as future improvements.

Prototype	The prototype is well executed and has a good accuracy. GUI can be improved by adding more components for customization and adding bigger dataset.
Difficulty of the final application	Application is easy to use both technical and not technical people
GUI and user experience of the application	User experience is good for both technical and non-technical users, because it is straightforward.

Table 26 Self evaluation

9.5 SELECTION OF EVALUATORS

In order to obtain evaluations from individuals with a variety of educational and professional backgrounds, evaluators were chosen based on three categories. The following has three categories.

1. Beginner developers who have some background in and experience with machine learning.
2. Domain experts who will profit from the system.
3. Technology experts who are well-versed and skilled in machine learning.

Group	**Affiliation**	**Reason**
Beginner developers	Evaluator 1 Undergraduate Student, UCSC (University of Colombo school of computing)	He is doing a machine learning project as his final year project and therefore has a good knowledge of machine learning models and so on
Domain Experts	Evaluator 2 Software Architect, Mobitel (Pvt) Ltd.	He has more than 15 years' experience in telecommunication domain and software systems
	Evaluator 3	He has more than 10 years' experience in

	Tech Lead - Big Data & data Science Information Systems Mobitel (Pvt) Ltd.	telecommunication domain, data science and software systems. Moreover, he has a knowledge about customer requirements
Technical experts	Evaluator 4 Senior Software Engineer, Sysco Labs	She is working on machine learning projects in the work environment.
	Evaluator 5 Senior Software Engineer, CodeGen	He is skilled in ML technology. He actively participates in ML developments in his capacity as an employee.

Table 27 Selection of evaluators

9.6 EVALUATION RESULTS

9.6.1 QUALITATIVE RESULTS

9.6.1.1 THE OVERALL NOTION OF THE PROJECT

Evaluator	Feedback
Evaluator 1	Very intriguing idea. Genuine datasets from the local telco domain are not available to everyone. The hardest element, in my opinion, is finding a reliable dataset, which we find to be really challenging.
Evaluator 2	It has more business value, since predicting VAS for suitable customers is more beneficial for the business.
Evaluator 3	Identifying customers for VAS services is really tricky and important for our business.
Evaluator 4	It is an attractive research area to explore in machine learning

Evaluator 5	The project covers a wide range of topics including model selection, prediction, and data preparation in the workflow of a machine learning.

Table 28 Overall notion of the project

Evaluation Summary

The concept received positive feedback from all the evaluators, who also noted the importance of a reliable dataset and the need of keeping high accuracy.

9.6.1.2 LITERATURE SURVEY

Evaluator	**Feedback**
Evaluator 1	There has been thorough literature research on a range of models, and the expertise is high.
Evaluator 2	Before coming up with an appropriate methodology, the author seemed to have completed an appropriate literature review.
Evaluator 3	Good literature review is carried out for the project.
Evaluator 4	Prior to creating a successful methodology, the author appears to have completed an appropriate literature review and thoroughly investigated the field.
Evaluator 5	Good literature review was done before the project.

Table 29 Literature Survey

Evaluation Summary

The literature survey received positive feedback from all the evaluators, who also noted the importance of a carrying out a good literature survey before coming up with the methodology.

9.6.1.3 SYSTEM DESIGN AND IMPLEMENTATION

Evaluator	Feedback
Evaluator 1	System design, architecture and implementation seems impressive.
Evaluator 2	The implementation is great. Design and architecture look good.
Evaluator 3	Good implementation. Even non-technical people can easily understand it. UIs can be enhanced.
Evaluator 4	It is easy to understand and follow architecture. Since you have completed UML designs, the design is also good and standard.
Evaluator 5	This project covers practically every typical machine learning workflow. For instance, preparing the data and delivering the best models for the specific dataset.

Table 30 System design and implementation

Evaluation Summary

The System design and implementation received positive feedback from all the evaluators, who also noted the improvements that can be done for GUIs.

9.6.1.4 PROTOTYPE

Evaluator	Feedback
Evaluator 1	The prototype is meeting the requirements and is operating as intended in the suggested system.
Evaluator 2	The size of the input data could be increased, and accuracy is encouraging.
Evaluator 3	The concept is demonstrated through a prototype when the intended outcomes are obtained.
Evaluator 4	Functionality of the prototype is great. The models' accuracy can be increased in the future.

Evaluator 5	Prototype demonstrates the concept and produces the expected result.

Table 31 Prototype

Evaluation Summary

The prototype received positive feedback from every evaluator, who also said that it demonstrated the concept and performed according to the proposed system.

9.6.2 QUANTITATIVE ANALYSIS

Quantitative results are shown below. The audience of the analysis is the same audience used for qualitative analysis.

9.6.2.1 DIFFICULTY OF THE SYSTEM

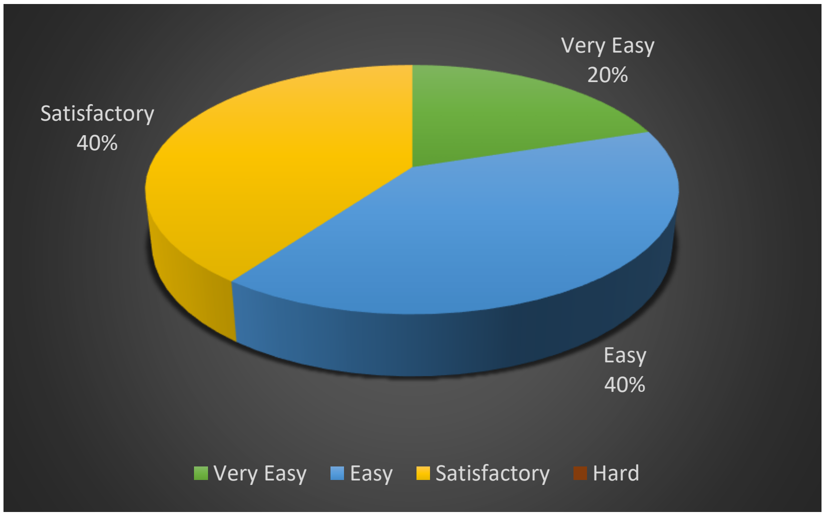


Figure 35 Difficulty of the system responses

The system was evaluated to be satisfactory, easy, or very easy by each evaluation team member. While 20% of those evaluated considered the system to be very easy, the majority of the evaluators chose between easy and satisfactory, with 40% proportion for each.

9.6.2.2 USER INTERFACE AND USER EXPERIENCE OF THE APPLICATION

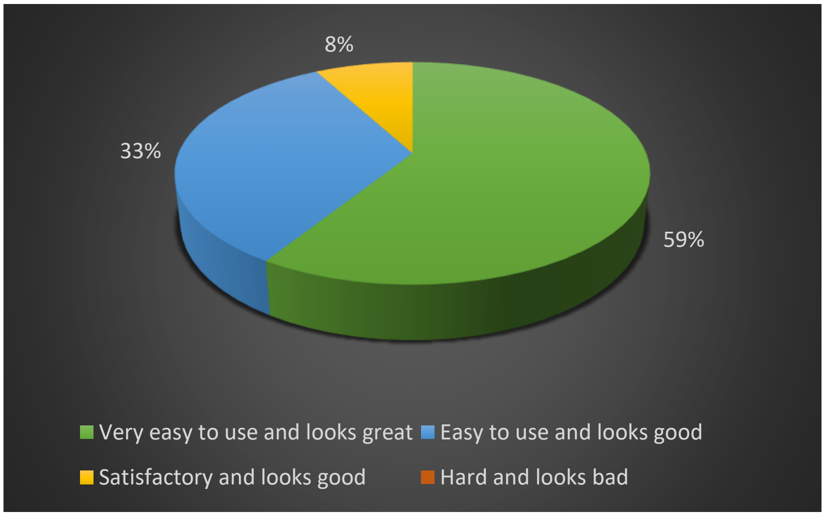


Figure 36 User experience of the system responses

The user experience gets positive feedback from every respondent. Around 59% of those who evaluated the system said the UI was attractive and easy to use. The rest 41% of the evaluators found the system to be at least satisfactorily easy.

9.7 CHAPTER SUMMARY

This chapter includes numerous expert evaluations on a range of categories. The methodology and evaluation approach were the chapter's opening subjects. This chapter includes documentation of the evaluation criteria, the selection of the evaluators, and their suitability to evaluate the project. In terms of concept, literature survey, design, architecture, implementation, and prototype, the outcome of evaluations are presented. Finally, functional and non-functional requirements are evaluated.

CHAPTER 10: CONCLUSION

10.1 CHAPTER OVERVIEW

This chapter contains the project's concluding notes. This chapter mostly covers topics like whether the project's goals and objectives were met, the problems that were encountered, the restrictions placed on the project, how the knowledge gained through the master's degree was applied to the project, and what new abilities were developed. It also provides evidence of the researcher's contributions to the subject.

10.2 ACHIEVEMENTS OF RESEARCH AIMS & OBJECTIVES

10.2.1 AIM OF THE PROJECT

The aim of this project is to design, develop and evaluate a system for predicting VAS services for telco customers based on the given inputs which helps to improve the profit of the telecommunication operator.

The research aims as mentioned above is achieved during the development time of the research project. System was successfully designed, developed and evaluated. Model which was trained using 'Bagging CART' is selected as the most accurate and best performing model.

10.3 UTILIZATION OF KNOWLEDGE FROM COURSE

Module	**Description**
Research Methods and Professional Practices	The principles of conducting research, including what research is, how to write a research proposal, and an introduction to thesis writing, were addressed in this module.
Enterprise Application Development	This subject covers the fundamentals of creating enterprise-level applications, including the underlying theories and the

	essential techniques for software design, such as UML modelling, System development life cycle etc.
Advanced Software Design	The stages for creating an effective business-level application were covered in this module. Through the use of this specific module, students learned about design methodologies, best practices, and how to build a designed application.
Data Mining and Machine Learning	The principles of machine learning were addressed in this module. This module covered both fundamental and more advanced machine learning concepts and methods.

Table 32 Utilization of knowledge from course

10.4 USE OF EXISTING SKILLS

The author has acquired a number of skills throughout the Master's program that she used to complete the research project. The following is a list of some of the skills that were acquired during the course of study:

- Machine Learning - During the study of the Data Mining and Machine Learning module, the author acquired a fundamental understanding of machine learning as well as some advanced techniques.
- Python - In the Data Mining and Machine Learning module, the author gained a foundational understanding of Python.

10.5 USE OF NEW SKILLS

The author has consistently learned throughout the course of the research in order to successfully conduct research and implement proof of concept. The following is a list of some of the talents acquired throughout the research and development phase.

- Python - Although the author has studied the fundamentals of the Python programming language as part of her MSc module, she has self-taught advanced Python approaches for constructing prototypes and predictive models.
- Streamlit - The author self-taught herself the Streamlit framework, which is used to build web apps, and used it to construct the prototype's GUI.

10.6 ACHIEVEMENTS OF LEARNING OUTCOMES

Learning Outcomes	**Description**
L01	Publications on machine learning in the telecommunications sector have been reviewed in order to gain a deeper understanding of the chosen research topic, identify research gaps, and identify the technology needed to solve identified research issues and successfully finish the research project.
L02	Planning and scheduling for the project were completed when creating the project proposal in the first phase. In order to ensure that the project can be finished within the anticipated timescale, prepared plans and schedules were carefully examined. There were times when a few jobs required more time to be provided since the allocated time during planning for those tasks was insufficient to complete them. To gain more time in the timeframe, several tasks have to be accelerated. The work's quality has not been impacted despite the timing problems.
L03	Machine learning methods like "bagging", "stacking" and “ensemble learning” as well as various tools like "python" were used to train the models and evaluate their accuracy in order to make the research project and prototype successful. Streamlit was used for prototype.
L04	This is accomplished and covered in detail in the chapters on system architecture & design and implementation, respectively.
L05	This is accomplished and covered in detail in the chapters on evaluation and social, legal, ethical, and professional issues.

L06	All the components necessary to meet an academic level report for the documentation are included in a comprehensive thesis.

Table 33 Achievements of learning outcomes

10.7 PROBLEMS AND CHALLENGES FACED

Problem	**Solution**
This research required legitimate data set from Sri Lankan telecommunication service provider since this is done based on Sri Lankan subscribers.	Author able to get a significant dataset from one of prominent telecommunication service provider with the support of her supervisor.
There were not many resources available to train the models for performance evaluation.	Based on the data gathered using a single laptop, performance is evaluated.
Obtaining appointment slots to interview experts for gathering requirements is difficult.	Author was able to interview every interviewee, which took longer than expected.
Working on the project was challenging because of the country's current circumstances, which includes frequent power outages.	somehow succeeded in finishing the assignment by putting in extra time and effort.
COVID-19	Emails, social media, and online meetings are used to communicate with the relevant stakeholders.

Table 34 Problems and challenges

10.8 LIMITATIONS OF THE RESEARCH

There were a number of issues and restrictions throughout the entire project. Most of those were quickly resolved, but several required more work. Some of the major problems and restrictions encountered during the project are outlined below.

- The project had a very broad scope. By itself, machine learning is a vast area with many underlying theories and statistical components. The true scope could only be defined once the author began working on implementation.
- In order to get a clear understanding of the models' performances, it is preferable to test them on several different machines with different specifications. But only one laptop was used to test the predictive models.
- It is preferable to have evaluated an alternative dataset with the same features, preferably from a different service provider, in order to obtain a precise measurement of the implemented system's accuracy. However, the testing and evaluation only make use of one dataset.

10.9 FUTURE ENHANCEMENTS

There are numerous potential future enhancements that could be addressed based on the limits and the newer dimension that this project is adding to the domain.

- Encourage more learning styles. The technology only supports supervised and unsupervised learning as of right now. There are numerous additional types, including transfer learning and associative learning.
- The results of batch processing and excel uploading can be used to construct a reporting component.

10.10 CONCLUDING REMARKS

The project and thesis are concluded in this chapter with a discussion of whether the goals and objectives stated at the outset of the project were achieved. How the author's knowledge from prior degree modules and abilities from previous work were applied in this project are covered. Also discussed are the fresh insights gained, the limitations, and the limitations encountered. Finally, future improvements are recommended.

REFERENCES

75, n.d. Sri Lanka - Telecommunications and Information Technology [WWW Document]. URL https://www.trade.gov/country-commercial-guides/sri-lanka-telecommunications-and-information-technology (accessed 4.18.22).

Kurniawan, F., Umayah, B., Hammad, J., Nugroho, S.M.S., Hariadi, M., 2017. Market Basket Analysis to Identify Customer Behaviours by Way of Transaction Data. Knowl. Eng. Data Sci. 1, 20. https://doi.org/10.17977/um018v1i12018p20-25

Seyedan, M., Mafakheri, F., 2020. Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. J. Big Data 7, 53. https://doi.org/10.1186/s40537-020-00329-2

Srivastava, N., Stuti, Gupta, K., Baliyan, N., 2018. Improved Market Basket Analysis with Utility Mining (SSRN Scholarly Paper No. 3170300). Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3170300

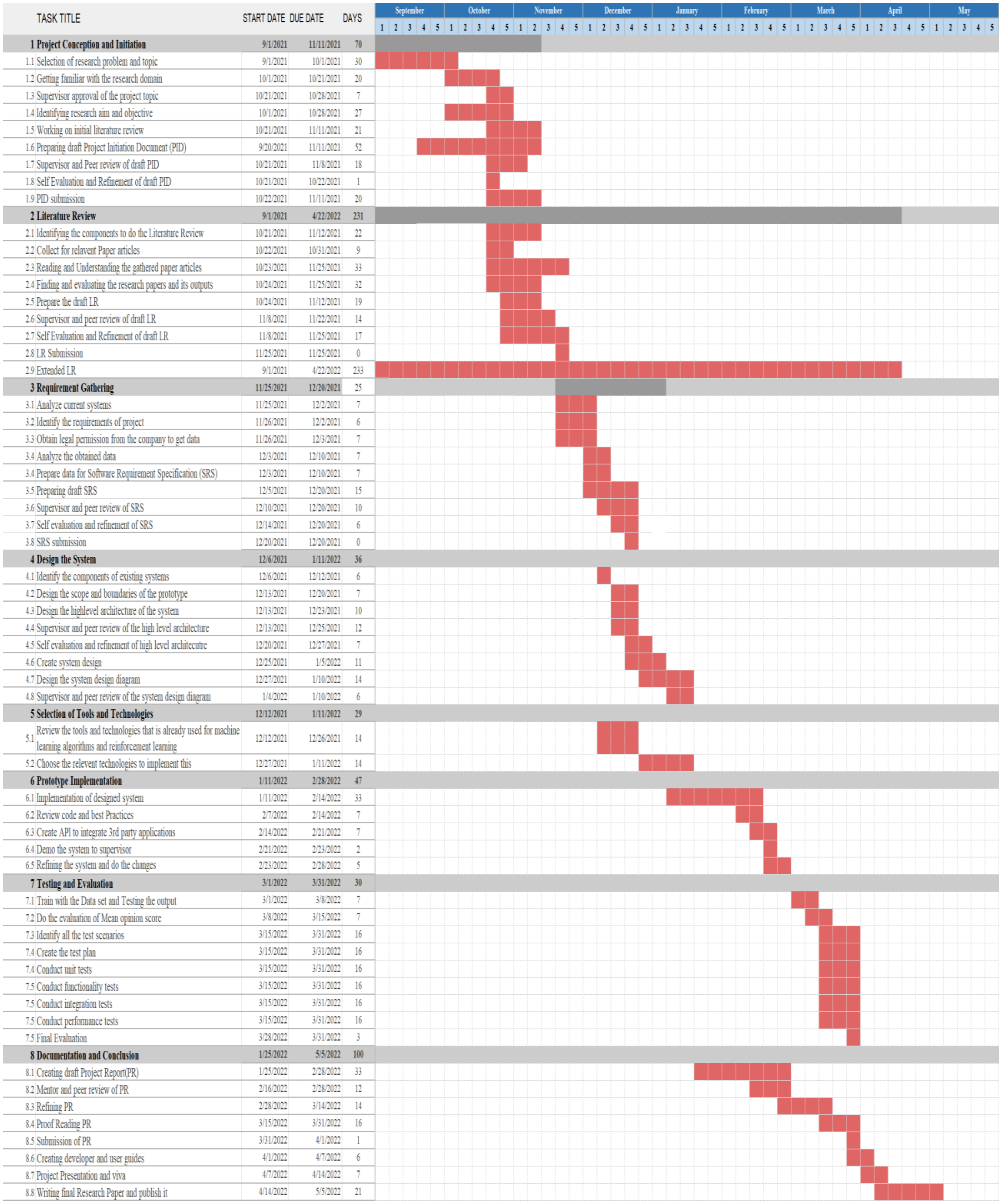
Szymkowiak, M., Klimanek, T., Józefowski, T., 2018. APPLYING MARKET BASKET ANALYSIS TO OFFICIAL STATISTICAL DATA. ECONOMETRICS 22, 39–57. https://doi.org/10.15611/eada.2018.1.03

Telecommunications Industry [WWW Document], n.d. . Corp. Finance Inst. URL https://corporatefinanceinstitute.com/resources/careers/companies/telecommunications-industry/ (accessed 4.18.22).

Vahidi Farashah, M., Etebarian, A., Azmi, R., Ebrahimzadeh Dastjerdi, R., 2021. An analytics model for TelecoVAS customers' basket clustering using ensemble learning approach. J. Big Data 8, 36. https://doi.org/10.1186/s40537-021-00421-1

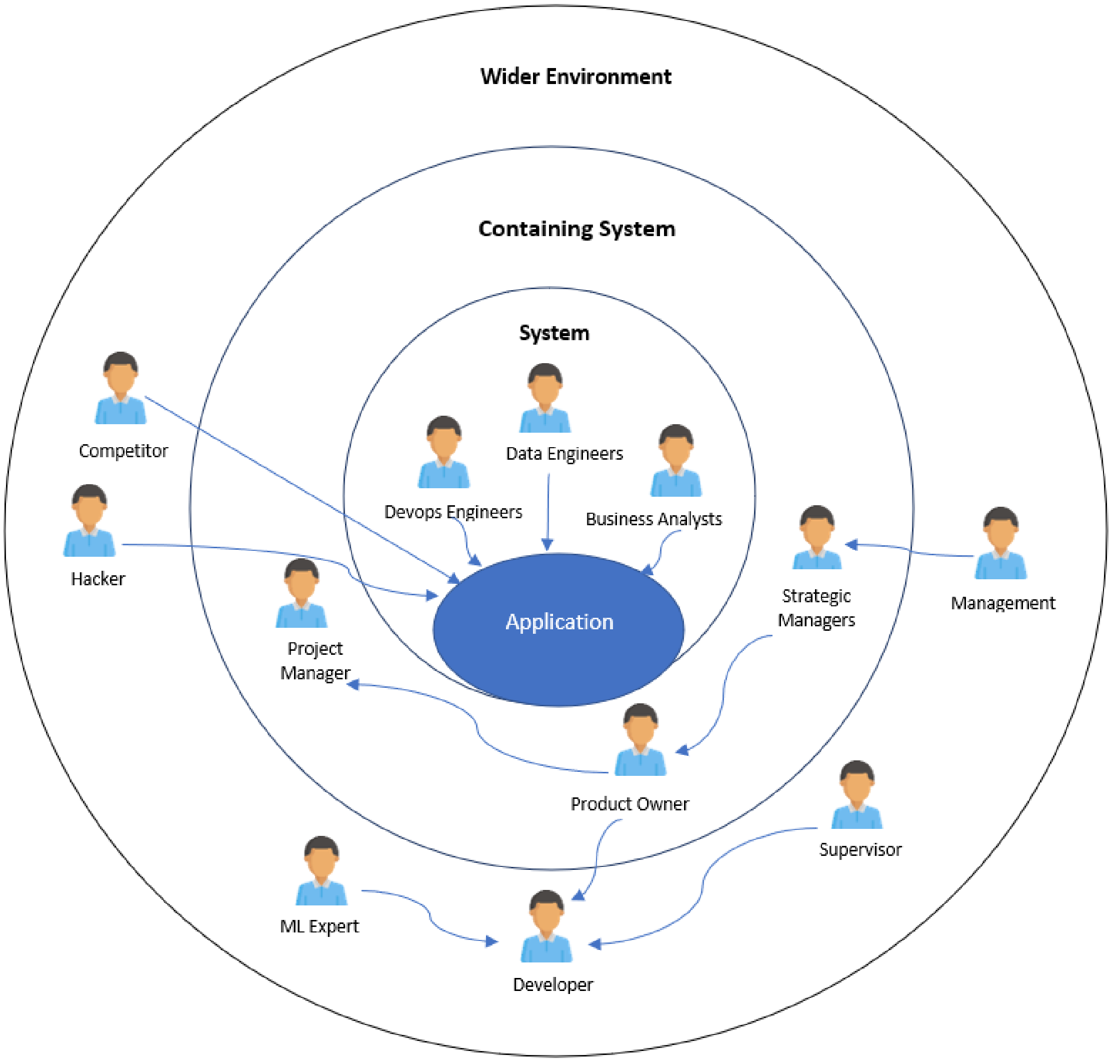
What is a Value-Added Service (VAS)? - Definition from Techopedia [WWW Document], n.d. . Techopedia.com. URL http://www.techopedia.com/definition/25234/value-added-service-vas (accessed 4.18.22).

APPENDIX A : Gantt Chart



	TASK TITLE	START DATE	DUE DATE	DAYS
1	**Project Conception and Initiation**	**9/1/2021**	**11/11/2021**	**70**
1.1	Selection of research problem and topic	9/1/2021	10/1/2021	30
1.2	Getting familiar with the research domain	10/1/2021	10/21/2021	20
1.3	Supervisor approval of the project topic	10/21/2021	10/28/2021	7
1.4	Identifying research aim and objective	10/1/2021	10/28/2021	27
1.5	Working on initial literature review	10/21/2021	11/11/2021	21
1.6	Preparing draft Project Initiation Document (PID)	9/20/2021	11/11/2021	52
1.7	Supervisor and Peer review of draft PID	10/21/2021	11/8/2021	18
1.8	Self Evaluation and Refinement of draft PID	10/21/2021	10/22/2021	1
1.9	PID submission	10/22/2021	11/11/2021	20
2	**Literature Review**	**9/1/2021**	**4/22/2022**	**231**
2.1	Identifying the components to do the Literature Review	10/21/2021	11/12/2021	22
2.2	Collect for relavent Paper articles	10/22/2021	10/31/2021	9
2.3	Reading and Understanding the gathered paper articles	10/23/2021	11/25/2021	33
2.4	Finding and evaluating the research papers and its outputs	10/24/2021	11/25/2021	32
2.5	Prepare the draft LR	10/24/2021	11/12/2021	19
2.6	Supervisor and peer review of draft LR	11/8/2021	11/22/2021	14
2.7	Self Evaluation and Refinement of draft LR	11/8/2021	11/25/2021	17
2.8	LR Submission	11/25/2021	11/25/2021	0
2.9	Extended LR	9/1/2021	4/22/2022	233
3	**Requirement Gathering**	**11/25/2021**	**12/20/2021**	25
3.1	Analyze current systems	11/25/2021	12/2/2021	7
3.2	Identify the requirements of project	11/26/2021	12/2/2021	6
3.3	Obtain legal permission from the company to get data	11/26/2021	12/3/2021	7
3.4	Analyze the obtained data	12/3/2021	12/10/2021	7
3.4	Prepare data for Software Requirement Specification (SRS)	12/3/2021	12/10/2021	7
3.5	Preparing draft SRS	12/5/2021	12/20/2021	15
3.6	Supervisor and peer review of SRS	12/10/2021	12/20/2021	10
3.7	Self evaluation and refinement of SRS	12/14/2021	12/20/2021	6
3.8	SRS submission	12/20/2021	12/20/2021	0
4	**Design the System**	**12/6/2021**	**1/11/2022**	**36**
4.1	Identify the components of existing systems	12/6/2021	12/12/2021	6
4.2	Design the scope and boundaries of the prototype	12/13/2021	12/20/2021	7
4.3	Design the highlevel architecture of the system	12/13/2021	12/23/2021	10
4.4	Supervisor and peer review of the high level architecture	12/13/2021	12/25/2021	12
4.5	Self evaluation and refinement of high level architecutre	12/20/2021	12/27/2021	7
4.6	Create system design	12/25/2021	1/5/2022	11
4.7	Design the system design diagram	12/27/2021	1/10/2022	14
4.8	Supervisor and peer review of the system design diagram	1/4/2022	1/10/2022	6
5	**Selection of Tools and Technologies**	**12/12/2021**	**1/11/2022**	**29**
5.1	Review the tools and technologies that is already used for machine learning algorithms and reinforcement learning	12/12/2021	12/26/2021	14
5.2	Choose the relevent technologies to implement this	12/27/2021	1/11/2022	14
6	**Prototype Implementation**	**1/11/2022**	**2/28/2022**	**47**
6.1	Implementation of designed system	1/11/2022	2/14/2022	33
6.2	Review code and best Practices	2/7/2022	2/14/2022	7
6.3	Create API to integrate 3rd party applications	2/14/2022	2/21/2022	7
6.4	Demo the system to supervisor	2/21/2022	2/23/2022	2
6.5	Refining the system and do the changes	2/23/2022	2/28/2022	5
7	**Testing and Evaluation**	**3/1/2022**	**3/31/2022**	**30**
7.1	Train with the Data set and Testing the output	3/1/2022	3/8/2022	7
7.2	Do the evaluation of Mean opinion score	3/8/2022	3/15/2022	7
7.3	Identify all the test scenarios	3/15/2022	3/31/2022	16
7.4	Create the test plan	3/15/2022	3/31/2022	16
7.4	Conduct unit tests	3/15/2022	3/31/2022	16
7.5	Conduct functionality tests	3/15/2022	3/31/2022	16
7.5	Conduct integration tests	3/15/2022	3/31/2022	16
7.5	Conduct performance tests	3/15/2022	3/31/2022	16
7.5	Final Evaluation	3/28/2022	3/31/2022	3
8	**Documentation and Conclusion**	**1/25/2022**	**5/5/2022**	**100**
8.1	Creating draft Project Report(PR)	1/25/2022	2/28/2022	33
8.2	Mentor and peer review of PR	2/16/2022	2/28/2022	12
8.3	Refining PR	2/28/2022	3/14/2022	14
8.4	Proof Reading PR	3/15/2022	3/31/2022	16
8.5	Submission of PR	3/31/2022	4/1/2022	1
8.6	Creating developer and user guides	4/1/2022	4/7/2022	6
8.7	Project Presentation and viva	4/7/2022	4/14/2022	7
8.8	Writing final Research Paper and publish it	4/14/2022	5/5/2022	21

APPENDIX B : Stakeholder Onion Model


Wider Environment
Containing System
System
Data Engineers
Devops Engineers
Business Analysts
Application
Competitor
Hacker
Project Manager
Strategic Managers
Management
Product Owner
Supervisor
ML Expert
Developer