CUSTOMER CHURN REASONING ANALYSIS MODEL FOR TELECOMMUNICATION INDUSTRY.

Kapugama Geeganage Kasun Thilina

198777H

Faculty of Information Technology

University of Moratuwa

July 2022

CUSTOMER CHURN REASONING ANALYSIS MODEL FOR TELECOMMUNICATION INDUSTRY.

Kapugama Geeganage Kasun Thilina

198777H

Dissertation submitted to the Faculty of Information Technology,

University of Moratuwa, Sri Lanka

for the partial fulfillment of the requirements of the

MSC in Information Technology

July 2022

Declaration

I have used my own work for preparing this report. All the other published and non-published materials which were used in the research study have been acknowledged in the references section on the thesis.

K G Kasun Thilina

Signature of Student:

Date:

Supervised By

Dr. C.P. Wijesiriwardena

Lecturer

Faculty of Information Technology

University of Moratuwa

Signature of Supervisor:

Date:

Acknowledgements

I am so grateful to Dr. Chaman Wijesiriwardana, Senior Lecturer, Faculty of Information Technology, and University of Moratuwa. My research supervisor for guiding me in all aspects and supported in achieving the objectives and milestones in order to complete the project successfully.

My next most gratitude pays to all academic staff from the Faculty of Information and technology, who shared their vast knowledge throughout these two years by providing me with a good environment, which influenced a lot to achieve this goal.

Then, I would thank all non-academic staff of the University of Moratuwa, Sri Lanka for all the efforts and facilities that it has taken to contribute towards this postgraduate program and all my colleagues for their support to complete this research study in many ways. All those whom I have not mentioned name-wise who helped and motivated me.

Furthermore, I extend my earnest thanks to my parents, and siblings, for always being my support system and motivating me always.

Abstract

Customer churn is the most impactful problem in every business and industry. Therefore, every company tries their best to satisfy and maintain existing customers. Today telecommunications companies are facing this problem frequently due to increasing demand of customers every day. It is very difficult to gather new customers and need to allocate a huge cost from company revenues to acquire new customers compared to retaining the existing customer, therefore it is more important to increase their customer retention and work for that. This research is based on churn customer information and the primary objective of this research is to predict the churn reason of a given customer who has predicted to be churn using modern data analytics techniques. It include Logistic Regression, Naive Bayes, Random Forest, Decision Tree, K-Nearest Neighbor, Support Vector Machine and Gradient Boost Classifier. Further, Hybrid Model has been considered using Voting Classifier ML model. The dataset used in this research is obtained through the Data Warehouse of one of the leading telecommunication companies.

Table Contents

Abbreviations

DM	Data Mining
WEKA	Waikato Environment for Knowledge Analysis
CRL	Classification by Rule Learning
DMEL	Data Mining by Evolutionary Learning
EA	Exhaustive Algorithm
GA	Genetic Algorithm
CA	Covering Algorithm
LA	LEM2 Algorithm
DT	Decision Tree
ANN	Artificial Neural Networks
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
IQR	Interquartile Range
SMOTE	Synthetic Minority Oversampling Technique
CSV	Comma Separated Values

List of Figures

List of Tables

Chapter 01

Introduction

1.1 Background and Motivation

In the current competitive world, keeping customers satisfied is the key success factor to compete in the market. As long as companies have been competing for sales, identifying the appropriate target market has been adopted in many industries since a deeper understanding of customers is important to provide a satisfied service. It is commonly accepted that serving all consumers and satisfying them is quite hard because people have different needs and wants and therefore, they cannot be targeted in the same manner.

Furthermore, telecommunication companies play a major role with huge customer demand. Due to some reasoning, today's customer will not hesitate to change their network service provider to another service provider if they are not satisfied with the service of the existing service provider or what they are looking for. The reason for the importance of telecommunication is because of so many tasks of the day today. Also due to the current COVID pandemic of every country, that service usage is increasing with the work from home concept. Therefore, every telecom service provider tries their best to provide their service in a competitive telecom market.

Over and above that, Customer churning is directly dependent on customer satisfaction, and in the telecom service industry, customer churn term used to denote the customer (either prepaid or postpaid) movement from one provider to another. As has been mentioned in previous research, obtaining new clients is more expensive compared to absorbent (hold prevailing clients without moving to another service provider) the existing customers [13, 14]. Therefore, customer satisfaction and attraction are one of the most significant goals in todays' telecommunication organizations since it appears to affect the company's revenue and income [14].

According to the previous research topics analysis, most of the research is based on customer churn prediction and few research are focus on the customer churn reasoning part. It is more important to focus on customer churn reasoning, so then an immediate

solution can be provided to retain customers if churn reason has a solution on the company side and the company can improve its services in the competitive business world. In addition to that, along with the proper management of clients, we could minimize churning susceptibility, maximize company profitability and improve customer satisfaction.

1.2 Aim and Objectives

1.2.1 Aim

The aim of this research study is to identify a data mining model for predicting churn reason of a given customer who has predicted to be churn in telecom industry.

1.2.2 Objectives

This Research aims to achieve following four objectives,

1. To identify the gaps with existing research and proposed.
2. To identify the limitation in accessing confidential customer data and collect customer data.
3. To identify the features using suitable procedures.
4. To predict the churn reason of a given customer who has predicted to be churn.

1.3 Problem Definition

In the telecom Industry they allow customers to change their service provider to another service provider if the customer is not satisfied with the service. So, if a company is unable to find out the actual churn reason before terminating the connection by customer, the company is unable to retain customers and it'll impact company revenue. Based on the described situation, this research focuses on customer churn reasoning through the churn customer data analytics, and it'll be helpful to get immediate action before the customer terminates the connection.

1.4 Problem Statement

How to identify actual churn reason before terminate the connection by customer?

1.5 The Way of Proceeding

This research is based on churn customer information data which is obtained from the data warehouse of one of the leading telecommunication companies. The modern data analytics techniques are used to build a model including Decision Tree, K-Nearest Neighbors, and Logistic Regression. The expected result of this research is to predict the churn reason of a given customer who has predicted to be churn using one of the algorithms which provided the best accusation. The python technology was used as main programming language

1.6 Report Outline

The rest of the thesis is prepared according to the follow, Chapter 2 donate a descriptive literature review on a related topic, Studies done on Customer Churn Analysis, their limitations and practical usage. Chapter 3 explains the Technology Adaptation of the study. The researcher justifies the selection of technologies and approaches among the other available options. Further, chapter 4 presents the Customer Churn Reasoning Analysis. Chapter 5 is devoted to evaluation and design finished by using the researcher even as Chapter 6 gives the implementation of the solution. Further chapter 7 evaluate the proposed solution and finally, Chapter 8 provides the conclusion and further work of the study.

1.7 Summary

As discussed in above, customer reason analysis is more important to company revenue and according to the previous research topics analysis, maximum of the studies is based totally on consumer churn prediction and few research are targeted on the churn reasoning part. So, the next chapter is customer churn analysis literature review, and it provides an overview of current research related to this research.

Chapter 02

Customer Churn Analysis Literature Review

2.1 Introduction

The previous chapter has described the background of the research and a brief explanation about the way of going on. The others' works are focusing through this chapter and about others' approaches to solve similar problems. Customer retention is a key concern factor in many industries and it's particularly acute in the telecommunication industry due to high product diversification and high competition. Thus, customer churn was a major case study for many researchers since identifying the churnable customers and proactively responding to them is important to retain them and survive in the market. As the major objective of the study is to develop a model for customer churn reasoning, this chapter presents the studies on customer churn prediction approaches based on data mining techniques, limitations of these studies and ability to extend for predicting customer churn reasons.

2.2 Rule Based approach of Customer Churn Prediction

The term "Prediction" simply means guessing a future event. This guessing is done by exploring the previous pattern of took place events and identifying the factors and applying them for the destiny event. The prediction in business events has evolved from a time with different approaches. Consider the recent past, rule-based approaches for prediction were quite used in business to plan their business strategies. Rule based approaches are quite domain specific and heavily rely on the experience and professional orientation.

Huang, Y (2011) proposed a novel approach named Classification by Rule Learning. The short term of Classification by Rule Learning is CRL. It is a rule-based classification model for predicting the customer churn in the industry of telecommunication. It can classify churnable consumer segments in addition to capable of produce the prediction chance [16]. It has specified that the proposed approach has shown reasonable time in generating the expected results. Two main procedures were included in CRL: Initially heuristic of hill-climbing and pruning unimportant and

redundant rules were applied for generating rules. Secondly, test case category was predicted according to the rules.

The take a look at was conducted based on UCI repository datasets and Ireland Telecoms which is a telecommunication dataset. Considering the evaluation process, Lift Curve and AUC were used to measure the performance of the proposed model and the performance was compared with the DMEL (Data Mining by Evolutionary Learning) model. The experimental consequences show that CRL is greater efficient than DME.

Adan A (2016) has used rough set theory (RST) for proposing a wise rule-based totally decision-making method which could extract important decision rules relevant to customer churn and non-churn [9]. The churn customers can be classified from non-churn customers by this approach effectively. Then the customers can identify (predict) those customers who will churn or may bossily churn in the near future. Furthermore, they have evaluated the performances of the approach rough set theory (RST) with four rule-generation mechanisms, such as the Exhaustive Algorithm (EA), Genetic Algorithm (GA), Covering Algorithm (CA) and the LEM2 Algorithm (LA). The RST based classification which they visualized as shown in figure 2.1.

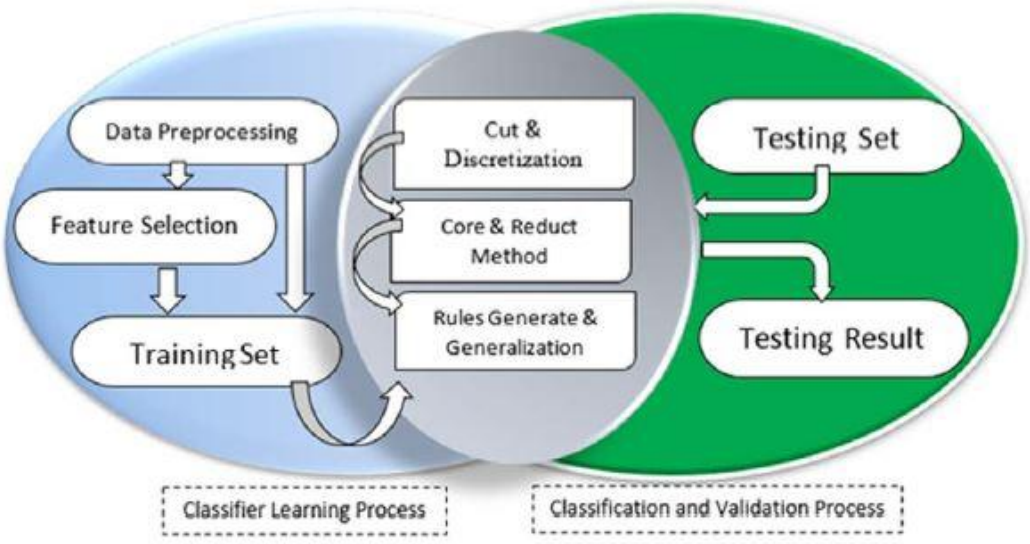


Figure 2.1 Proposed Methodology [9]

According to the evaluations, it indicated that in terms of recall , precision, misclassification rate, lift, coverage, accuracy, and F-measure, other rules generation algorithms were outperformed by the RST classification based on GA. Additionally, some important insights regarding reasons for customer churn which enables retention action development were revealed by the discretization process applied to different attributes.

The existing studies on rule-based approaches of customer churn prediction indicate effective methodologies with use of theoretical knowledge and domain expertise. However, with the development of data intensive nature of the industries and complexities in customer behavior, industries tended to move for data mining based approaches of customer churn prediction. Data mining based techniques were able to identify the hidden patterns of the customers and generate valuable insights.

2.3 Data Mining based customer segmentation

Keramati A (2014) Proposed a novel approach for predict the churnability of the customer by integrating major 4 classifiers, DT (Decision Tree), ANN (Artificial Neural Networks), KNN (K-Nearest Neighbors), and SVM (Support Vector Machine), and came up with a much more accurate hybrid classifier [1]. The study was carried out based on a call center database over a 12-month period of an Iranian mobile company and extracted the features as number of Call Failure, number of Complains, Subscription Length, Charge Amount, Use Frequency, Use Seconds, Frequency of SMS, Distinct Calls Number, Age Group, Service Type and Status [1].

The study has explored the capabilities of these major 4 classifiers and identified that ANN (Artificial Neural Network) significantly outperformed the other three, the study further presented the proposed novel methodology as a more accurate approach for a limited training dataset. It is shown in Figure 2.4, the four classifiers, which we call interior classifiers, are used to build the new one. The proposed approach has shown that the telecom company can gain a substantially higher than 95% accuracy rate for both recall measures and precision ad it is shown in Figure 2.2 and Figure 2.3.

	1	2	3	4	5
Best DT	0.787	0.834	0.814	0.828	0.809
ANN	0.908	0.891	0.869	0.874	0.869
KNN	0.706	0.725	0.727	0.748	0.743
SVM	0.923	0.932	0.912	0.931	0.939
Best hybrid methodology	0.966	0.966	0.98	0.966	0.979

Figure 2.2 Hybrid methodology Recall values [1]

	1	2	3	4	5
Best DT	0.903	0.901	0.936	0.914	0.895
ANN	0.857	0.852	0.845	0.823	0.838
KNN	0.741	0.76	0.762	0.780	0.809
SVM	0.715	0.761	0.813	0.734	0.731
Best hybrid methodology	1	1	1	1	0.928

Figure 2.3 Hybrid methodology Precision values [1]

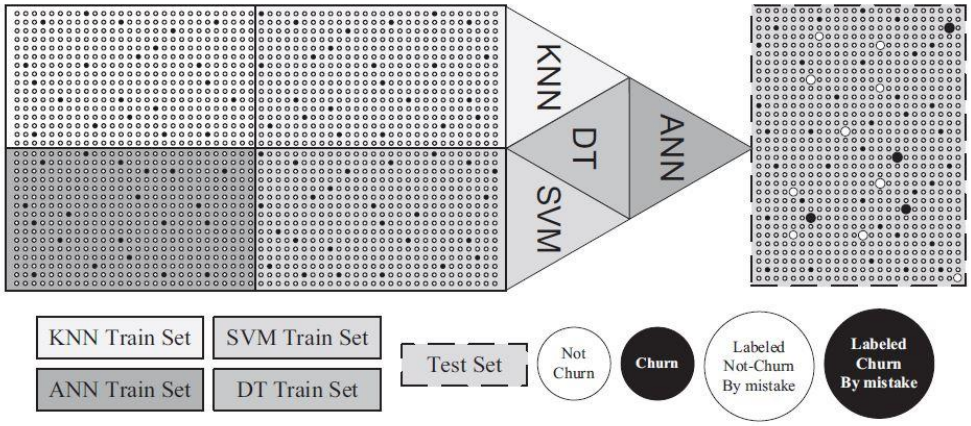


Figure 2.4 Proposed Methodology for Hybrid Classifier [1]

Shin-Yuan H (2006) presented theoretical infrastructure that we use to assess the overall performance of diverse churn prediction fashions constructed through statistics mining techniques [2]. The specialty of this study is it assessed the different churn prediction performance models built via different DM (Data Mining) techniques. The study was carried out based on a wireless telecommunication company in Taiwan, including dataset of about 160,000 subscribers who buy service from the service

provider company, including 14,000 churners who has terminated the contact with service provider company. The features were extracted into four main categories, as customer demography, analysis of bill and payment, analysis of call detail records and analysis of customer service/care [2].

As shown in the Figure 2.5, the proposed approach has used classification techniques under different scenarios such as applying a certain classification technique for all the customers and again the same classification for an identified customer segment [2]. The study mainly adapted two classification models, (DT) Decision Tree and Neural Network. The consequences suggest that both neural and Decision Tree (DT) community techniques can deliver correct churn prediction models and it has shown better results with integrating the churn prediction with customer segments.

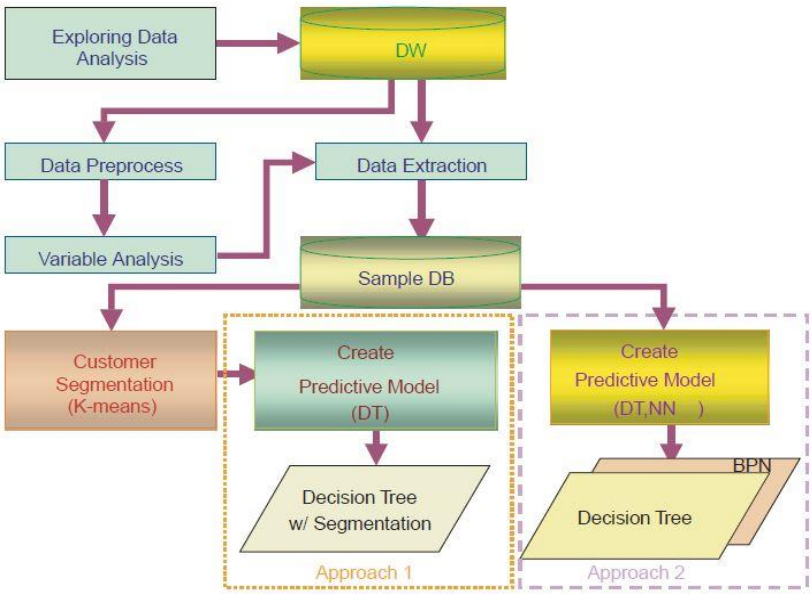


Figure 2.5 Process of Predictive Model Creation [2]

Bingquan H (2012) proposed an approach for detecting the customers who are willing to churn in domestic telecommunication service [7]. Despite others, this study has focused on wide coverage of feature extraction, since it presented a set of features considering precise account information, information of line, bill and payment information, four month call information, demographic profiles, complaint information and service orders. The extracted features were applied for seven classifiers and due to some limitations of certain classifiers, some features were normalized and applied for models. One of major specification of this study was, it has presented new feature set

and combining with existing features and seven classifiers (Logistic Regression, Decision Tree, Naive Bayes, Linear Classifier, ANN, SVN and Evolutionary Data Mining Algorithm) and came up with a new approach of churn prediction [7]. The study has presented a comparison of the seven modeling techniques effectiveness and the comparison of effectiveness between the existing feature sets and the new feature set. It has come up with a conclusion that the most appropriate modeling technique will depend on the objective of the decision makers, and it has presented the scenarios with respect to appropriate modeling technique.

2.4 Customer Churn Prediction Models

D. Kiran and B. Surbhi (2015) have suggested a modern framework for the churn prediction version and implements it the usage of the WEKA information Mining software as shown in the Figure 2.6 [11].

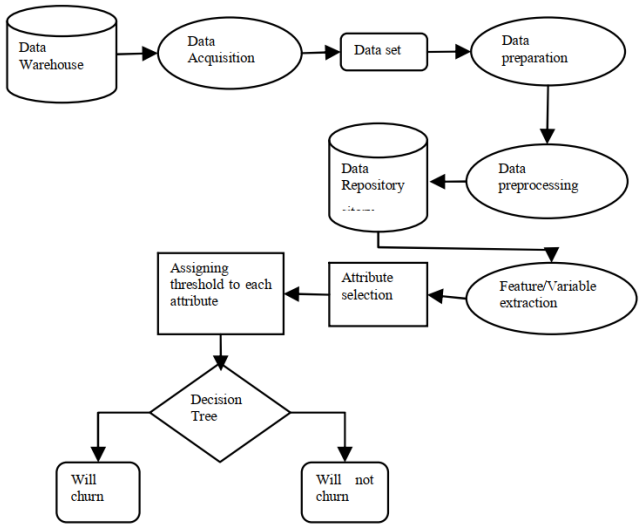


Figure 2.6 Churn Prediction Framework

They have in comparison the efficiency and the overall performance of DT (Decision Tree) and Logistic regression techniques. According to those comparisons, accuracy achieved with DT (Decision Tree) was higher than Logistic Regression data mining technique. These results clearly indicate the efficiency of Decision Tree technique [11]. The findings as shown in the Figure 2.7

Data Mining Techniques	Small Data set		Medium Data set		Large Data set	
	Churned	**Not Churned**	**Churned**	**Not Churned**	**Churned**	**Not Churned**
J-48 Decision Tree	47	3	96	4	606	2
Logistic Regression	43	7	86	14	590	18

Figure 2.7 Analysis of outcomes in data mining techniques [11]

Ammar Saleem Rehman, Saad Ahmed Qureshi, Ali Mustafa Qamar and Aatif Kamal have suggested a new framework for the churn prediction model. The data set contained 106,000 customers in telecommunication traffic data sine 3 months period and incoming, outgoing, voice, SMS (Short Message Service), data, traffic destination (on-net, competition), loyalty, rate plan, and traffic behavior are main features they have used. Further, Data preprocessing has been done with handling class imbalance. They have used number of data mining techniques to classify whether customers are churn or not. Mainly Logistic Regression, ANN (Artificial Neural Networks), K-Means clustering, DT (Decision Trees) including CHAID, Exhaustive CHAID, CART and QUEST. They have considered four variant of decision tree algorithm. Further, the evaluation has been done by using precision, recall and F-measure measurements and Exhaustive CHAID algorithm has been shown best result compared with other algorithm.

DM Algorithm	Recall		F-Measure	
	Active	Churn	Active	Churn
Logistic Regression	0.787	0.450	0.720	0.515
Artificial Neural Networks	0.823	0.325	0.698	0.419
K-Means Clustering	0.503	0.445	0.529	0.416
CHAID	0.792	0.531	0.743	0.583
Exhaustive CHAID	0.773	0.603	0.750	0.628
CART	0.897	0.271	0.740	0.383
QUEST	0.821	0.393	0.727	0.478

Table 2.1 Recall and F-Measure Results

As per the mention in the study, long time period big data set will be considered to test. Further, different counties and different telecommunication company data will be considered in the near future [10].

2.5 Summary

According to the literature review, most research is based on customer churn prediction, and they have used a lot of data mining techniques. Some of them tried to predict through the hybrid models also. In some research, the precision and recall have been used to degree the accuracy and it shows higher consequences. So, Most of the research papers Decision Tree and Logistic Regression DM algorithms are used and which algorithms shows the best results. The next chapter will be explaining about technologies used to solve the problem, which is mentioned in Chapter 1.

Chapter 03

Customer Churn Analysis Technology Adapted

3.1 Introduction

The previous chapter has described others' approaches to solve similar problems. According to the above chapter most researches are based on churn prediction and it can be identify as the gap which is tried to highlight in this study. This chapter gives an impression of the selected technology, which is perfectly adopted for our research to solve the problem, which is mentioned in chapter one, and it presents the usefulness of data mining techniques that distinguishes them from the technologies applied in existing literature.

3.2 Involvement of Data Mining

A huge number of databases and data are collected within various fields, with the rapid growth of Information Technology. There's been a rise to various database and Information technology-based research approaches which were accompanied in order to store and manipulate these data. Data mining is one of these methods which is used for extracting useful facts (also, it could be call as information) and styles from massive dataset. It's also known as understanding discovery technique, understanding mining from information, expertise extraction or analysis of records /pattern [15].

In DM (Data Mining), various artificial intelligence methods and database management statistics are used for extracting patterns from huge data sets. Given the informational advantages, Data Mining is significantly important for transforming data into business intelligence. Moreover, it's been used in a fixed of profiling practices like advertising, surveillance, fraud detection, and scientific discovery.

3.2.1 Data Mining Algorithms and Techniques

The Classification, Clustering, Regression, AI (Artificial Intelligence), Neural Networks, Association Rules, DT (Decision Trees), Genetic Algorithm, and Nearest Neighbor are various algorithms and techniques, which are used for knowledge discovery from databases.

The classification is the maximum usually carried out data mining technique, and it is the technique used in this study. The model is developed using classification techniques to get expected objectives, which are described in above sections. In general, data is one of the major items of every organization and data analysis. This research is based on one of those techniques called classification such as,

1. Logistic Regression

There are so many popular Machine Learning algorithms and Logistic regression is one of them. The category of that algorithms is Supervised Learning. Logistic Regression and Linear Regression have some similarities. It will be changed by the way that those algorithms are used. The Logistic Regression is used to resolve the Classification problems. Furthermore, Linear Regression is used to resolve the problems of Regression

2. Naive Bayes

Naive Bayes algorithm is used for solving classification problems and it comes under the Supervised Learning algorithm. Further, Naive Bayes algorithm is based on Bayes theorem. Mainly, the text classification can be done by using Naive Bayes. The Naive Bayes Classifier is most effective and one of the simple Classification algorithms. This can use to build fast machine learning models. Further. It can make fast predictions.

3. Random Forest

Random Forest comes under Supervised Learning technique and it is a popular ML algorithm. Classification and Regression problems can be solved by using this algorithm. The ensemble learning is concept that algorithm is based on.

4. Decision Tree

Decision Tree comes under Supervised Learning technique. Both classification and Regression problems can be solved by using this DM algorithm. But, most of the time it use to solve classification problems. It is based on tree structure and it contained internal nodes, branches and leaf nodes. The leaf node represents the result. The branches shows the decision rules. The internal nodes shows the dataset features. This algorithm can be understand simply. Because, it follows the same process of real-life

in decision making from human. It can be used to resolve problems which are decision related.

5. K-Nearest Neighbor

K-Nearest Neighbour comes under Supervised Learning technique and that algorithms is one of the simplest Machine Learning algorithms. It is used to solve Classification and Regression problems. But, mostly used for classification.

6. Support Vector Machine

Support Vector Machine or SVM comes under Supervised Learning algorithms and it is one of the most popular algorithms. Support Vector Machine is used to solve Classification problems as well as Regression problems. But mostly it used for Classification problems in ML. The algorithms contained two types, which are linear SVM and non-linear SVM. If the dataset is separable, then can use linear SVM. The data is separable means that, dataset can be separated by single straight line. If dataset cannot be separated by single straight, then it is called as non-linear SVM.

7. Gradient Boost Classifier

Gradient Boost Classifier comes under Supervised Learning algorithms and it is known as boosting algorithms. The algorithms has been built by using combining Decision Trees.

8. Hybrid Model (Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier)

Voting Classifier has been used to develop the Hybrid model and it is a machine learning model. The output is predicted based on their highest probability. It contained two types, which are hard voting and soft voting.

According to the description of this techniques, that DM techniques has been used across this research.

3.3 Model Development Programing Languages and Tools

The model development language and tool technologies are the most important part of the research, and the solution would be a combination of a number of those

technologies.

3.3.1 Programing Languages

These days' researchers use different kinds of programming languages for data science solution development. Python is one of them. Python programming language is easy-to-read nature and powerful analytics packages; therefore, data scientists and engineers are choosing it for data science. It has a number of features such as ensuring data quality, working with multiple data sources, generating visualizations, etc. Because of the above mentioned reasoning, python was the main model development technology.

3.3.2 Tools

The tool is plays most significant role in research to identify suitable model. There are number of python development tools such as Jupyter NoteBook, Pycharm, Google Colab, etc.

The Google Colab is a tool which is provided by Google. It's most important for beginners to learn. Because, the python libraries are exit inbuilt in the tool and processing power is high comparing to the local application execution.

The Jupyter NoteBook is web based python development tool and used for all sorts of data science tasks such as data cleaning and transformation, machine learning, data visualization, EDA (exploratory data analysis), statistical modeling, and deep learning. It can be easy to use and can be run cell by cell to better understand what the code does (clean code). Also, notebook can be converted to a different kind of standard output format such as HTML, PowerPoint, LaTeX, PDF, ReStructuredText, Markdown, Python using web interface simply.

Therefore, Jupyter NoteBook used as a python development tool for this study.

3.4 Data Collection Techniques

There are many data collection methods that exist in data sciences. From that, the dataset of this research is obtained from the Data Warehouse of one of the leading telecommunication companies. The NoSQL is the technology, which is used to fetch data from Data Warehouse. This gives a number of advantages, such as easy to

understand and learn capability to access large amounts of data directly without copying data into other applications, etc.

3.5 Summary

This chapter describes data mining as the technology proposed to analyze customer churn reasoning prediction and mention about eight techniques. Also, hybrid models are also mentioned as the technology to proceed the study. Further, Python programming language has been used in this study and Jupyter NoteBook is used as a tool to implement the solution. The next chapter will be explaining about the Churn Reasoning Analysis Approach.

Chapter 04

Churn Reasoning Analysis Approach

4.1 Introduction

The previous chapter has described technologies used to solve the problem which is mentioned in chapter one. It's important to get consideration bout the churn reasoning analysis approach and it will be discussed in this chapter. The input, output, process, data selection, data preprocessing, data mining, evaluation and features are the subtopics of this chapter.

4.2 Input

The data sources for the model will be collected using the data volumes for churned (already deactivated customers) customers provided by a telecommunication provider by using NoSQL. This data collects as CSV format, and it used as input of the program.

Some sample data is created without considering class variable and that value are input to the trained algorithms to get the predicted output of the class variable.

4.3 Output

Output is obtained for this process as different accuracy of models related to features selected of the dataset. That vales shows through the accuracy, precision, recall and f-1 score values. The values are further discussed through the evaluation subtopic. Also, class variable predication has been done to get output of predication class variable.

4.4 Process

In this process of churn reasoning analysis using data mining classification techniques all standard steps processed, which contains data selection to evaluation, are carried out. Throughout the process, the data set is cleaned, formatted and prepared for mining and interpretation.

4.4.1 Data Selection

It was a way to select data from the telecommunication provider due to the fact that they have customer churned information. So, after the identification of the feature, which is required to proceed the study, the data are requested from one of the telecommunication providers. That data is used to proceed the remaining task.

4.4.2 Data Preprocessing

After collecting the data, data sampling will be performed using randomly selected sets of deactivated customers with the relative information. Next data preprocessing (also called data preparation) phase will be undertaken which includes data cleaning and normalization steps. Data cleaning involves removing incorrect, incomplete, irrelevant, or improperly formatted facts (also, it can be called as information) which includes wrong spelling words caused by human mistakes, missing values, special mathematical symbols, duplicated information, strings ‘‘NULL’’, etc. The normalization step normalizes the values of features into a range (e.g., in between 0 and 1).

After data preprocessing, feature extraction is done to extract a set of features to represent deactivated customers.

4.4.3 Data Mining

Then, classification models will be developed for prediction, using appropriate modeling techniques to find which outfit works well for our model to suit the main objectives. These models will be trained & tested by the using training & testing technique and evaluated by using suitable performance measures. The python and Data Mining technologies are used to proceed the said approach above.

4.4.4 Evaluation

The mean score and Confusion Matrix for Multi-Class Classification main two factors are considered for prediction and evaluation. The mean score values will be calculated for each and every DM technique and check whether which DM technique gives the highest value compared with others. Then after, Confusion Matrix is considered and calculates accuracy, precision, recall and f-1 score values for each and every DM

technique. The research presented through the Confusion Matrix for Multi-Class Classification for the final evaluation and accuracy, precision, recall and f-1 score values are used to get final evaluation.

The accuracy describes overall accuracy of the model, i.e. the percentage of all samples correctly classified by the classifier. The (TP+TN)/ (TP+TN+FP+FN) formula is used to calculate the accuracy. The precision shows the percentage of predictions for the positive class that were actually positive and it is calculated using the TP/ (TP+FP) formula. The recall shows the percentage of all positive samples that were correctly predicted as positive by the classifier. Also called true positive rate (TPR), sensitivity, or probability of detection. The TP/ (TP+FN) is used as a formula to calculate that value. The f-1 score provides a single measure combining precision and recall. The 2 × (precision × recall) / (precision + recall) formula is used to calculate f-1 score. The meaning of the TP is true positive. The meaning of the TN is true negative. The meaning of the FP is false positive and FN is false negative.

4.5 Features

It was one of the main challenging parts of the analysis and this research required data, which are directly related to the analysis of churn reasoning. For that, the features selection for the data set is done by using previous studies, technical expertise of the industry and existing systems. The literature review was done to get knowledge from previous studies and more features could be identified from that. Some of the features are unable to get to consideration due to, that features are specific to the country of service providers. The technical expertise of the industry has lots of knowledge about the features. Sometime they have ideas best features which are need to get consideration. Because of, they have worked within the industry with most important features and experience churn problem. Further, most of the telecommunication industries have their own big data analysis teams and already, they could be identify features which are related to the churn prediction. Having a domain knowledge of the industry was more helpful to identify the features. Because of then, features can be analyzed having a good knowledge with it.

Table 10.1 shown the features which are consider in the whole study.

No	Column
1	gender
2	seniorCitizen
3	tenureMonths
4	multipleLines
5	currentPackage
6	contract
7	paymentMethod
8	averageVoiceMonthlyCharges
9	averageDataMonthlyCharges
10	churnReason

Table 4.1 Feature

The gender describe whether the customer male or female. The Senior Citizen describe whether the customer senior citizen or not. The data set is represented this using Yes or No values. The Tenure Months describe number of months the customer has stayed with the company. This value has been calculated using date of sim card is purchased to customer deactivated date. The multiple lines describe whether the customer has multiple lines or not and it is denoted using the value of yes or no. The current package feature describe whether existing package is data or voice-data. This is already calculated feature through the package type. The contract feature describe how long the connection is exist. The payment method feature has four values and it is describe the way of payment proceed. The average voice monthly charges is describe amount to voice calls for the customer monthly. The average data monthly charges is describe amount to data for the customer monthly. The churn reason is the class variable of the data set columns. Mainly four values are considered.

4.6 Summary

This chapter presented the overview of our approach to analyze churn data to identify the churn reasoning. In this scenario, it is highlighted that our approach offers an efficient and precise solution for churn reason analysis using data mining techniques. The next chapter provides the design of our approach presented here.

Chapter 05

Analysis and Design

5.1 Introduction

The previous chapter has represented the Churn Reasoning Analysis Approach and it described about the approach to give and idea how this study goes on. This chapter provides analysis and design we have done. Further, this presents top level design, data gathering, data preprocessing and modeling techniques used.

5.2 Top Level Design of the Proposed Classification Model

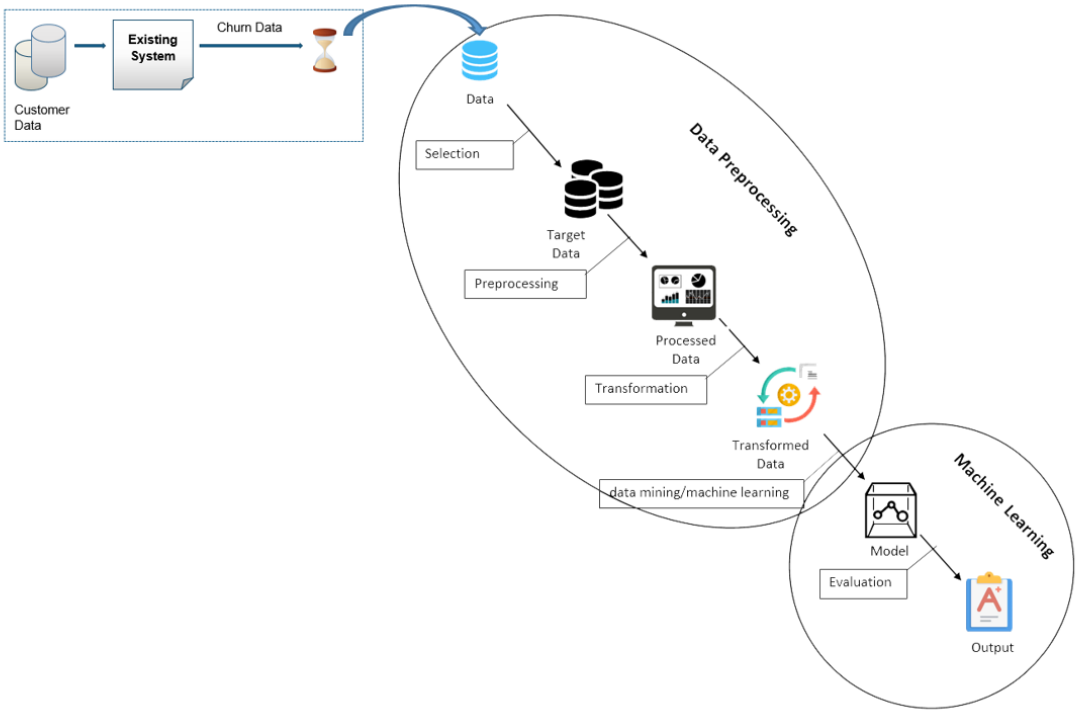


Figure 5.1 Classification Model

5.2.1 Data Gathering

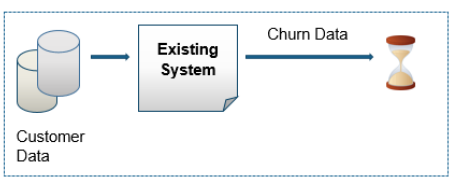


Figure 5.2 Data Source

The telecommunication industry has already adapted data mining based churn analysis approaches and the dataset used in this study is scraped from the Data Warehouse of one of the leading telecommunication companies in Sri Lanka for the period 2016 – 2021. The dataset is stored in a csv file, and it contains 11331 records. This all records are gathered from the existing system as shown in Figure 5.2. According to Figure 5.1, the data selection part has been considered and it is defined as the way of determining and obtaining relevant data for analysis from data collection which exists in the initial csv file. Then after, that process produces target data which is used for further analysis. The target data used to preprocess the part and processed data will be generated. This processed data is used in the transformation process. It is defined as the procedure of transforming data into suitable form required by mining procedure. Then after transformed data is generated. DM techniques are used to transform data to get the model expected. The DM is defined as sophisticated techniques applied to extract potentially useful patterns. Identifying strictly increasing patterns that represent knowledge on a given scale can be defined as Evaluation. The mean score, Confusion Matrix measurements, summarization and visualization can be used for that process.

5.2.2 Data Preprocessing

Data preparation is the most time-consuming phase in data analysis or data mining processes. As a first step, the future selection of the dataset is done according to the technical expertise of the industry, existing systems and existing research. Also, the idea of the underline business mechanism was helpful to confirm the described three points and get target data. The target dataset consists of 10 variables and churn reasoning is a class variable of the data set.

No	Column	Data Type
1	gender	object
2	seniorCitizen	object
3	tenureMonths	int
4	multipleLines	object
5	currentPackage	object

6	contract	object
7	paymentMethod	object
8	averageVoiceMonthlyCharges	float
9	averageDataMonthlyCharges	float
10	churnReason	object

Table 5.1 Churn Customer Variables and Data Types

The dataset is exit 7 categorical variables. Since the data mining model completely works on mathematics and numbers, this categorical variable may create trouble while building the model, therefore this categorical variable is converted to the numerical. The outlier removal was the next step of the data preprocessing. An outlier is a data point that is unusually different from other data points in a data set. This can distort analysis and violate our prediction. Therefore IQR (Interquartile Range) is used to remove outliers from the dataset. The imbalanced nature of dataset variables is the barrier towards a prediction model. The churn reasoning is identified as imbalanced class variable in the data set, and it was handled before the analysis using the data augmentation for the minority class technique and is referred as the SMOTE (The meaning of it is Synthetic Minority Oversampling Technique) [7] and used imbalanced-learn python library is used for achieving it. According to the steps above, a transformed dataset is created, and that dataset is used to develop a model.

5.2.3 Modeling

The data mining classification techniques are used mainly in the modeling module. The mention techniques are Logistic Regression, Naive Bayes, Random Forest, DT(Decision Tree), KNN (K-Nearest Neighbor), SVM (Support Vector Machine), Gradient Boost Classifier and Hybrid Model (Decision Tree Classifier, Random Forest Classifier and Gradient Boosting Classifier). The next step was to train the model using a training dataset and use test data to understand how well the model performs. There are various techniques available based on the dataset. Cross-validation is one of the techniques which is validating the model efficiency by training it on the subset of the provided data and testing on a previously unseen subset of the provided data. There are five cross-

validation techniques available and K-fold cross-validation technique is used in this study, and it divides the input data into five groups of samples (folds) of equal size.

Finally, evaluation is done through the output of model accuracy and reasoning prediction for a given input.

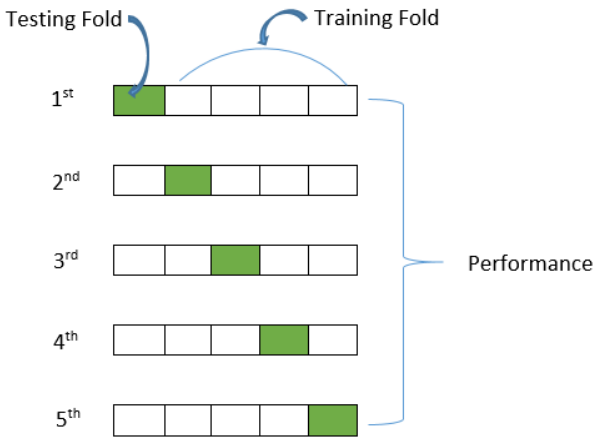


Figure 5.3 K-fold Cross-Validation Technique

5.3 Summary

This chapter is about analysis and designing part of the research under the topic's architecture diagram, data gathering, data preprocessing and modeling. This could be more help full to get some understanding of initial design and proceed with the next chapter. Implementation of the study will be discussed in the next chapter.

Chapter 06

Implementation of the Solution

6.1 Introduction

The previous chapter has described the Analysis and Design of Customer Churn Study and implementation of the solution will be discussed in this chapter. This contained the data gathering, data analysis and visualization and data reprocessing as main topics. The implementation of model is described in the section of Applying Appropriate Classification Technique which is contained under data preprocessing main topic.

6.2 Data Gathering

6.2.1 Loading Data

Personal computer local path is used to store the CSV file, which is extracted from the existing system. The churned customer dataset is imported to the Jupyter Notebook local installed python development web tool as shown in Figure 6.1.

```
df = pd.read_csv('chrun.csv')
```

Figure 6.1 The CSV Data File Import

6.3 Data Analysis and Visualization

	count	mean	std	min	25%	50%	75%	max
tenureMonths	11330.0	19.601589	20.425001	1.0	3.0	11.00	32.0	72.00
averageVoiceMonthlyCharges	11330.0	394.260494	267.034778	0.0	0.0	562.65	586.5	617.45
averageDataMonthlyCharges	11330.0	6618.991077	2006.163489	5019.1	5115.1	5762.10	7335.3	13109.80

Figure 6.2 The Statistics of Numeric Variables

Variable	Definition
count	The number of non-empty rows in a feature.
mean	The mean value of that feature.
std	The standard deviation value of that feature.
min	The minimum value of that feature.
25%, 50%, and 75%	The percentile/quartile of each feature. This quartile information helps us to detect Outliers.
max	The maximum value of that feature.

Table 6.1 Numeric Variables Description

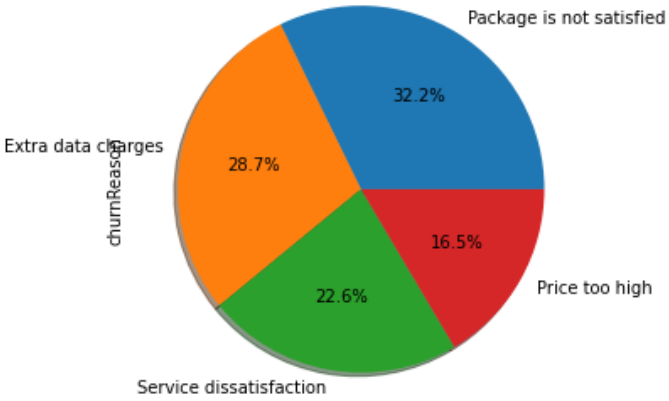


Figure 6.3 The Churn Reason Frequency

As shown in Figure 6.3 the “package is not satisfied” reason has churned most frequency with a percentage 32.2% and “price too high” reason has lowest frequency with a percentage 16.5%. Other two class variables are in middle of the described class variables. Those are extra data charges and service dissatisfaction. The percentage numerical values of both class variables are 28.7% and 22.6%

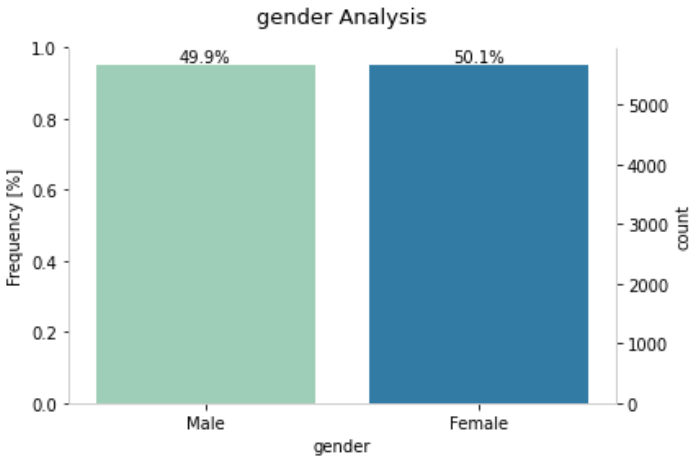


Figure 6.4 Analysis of Gender

According to Figure 6.4, female customers have churned slightly more than male customers, with a percentage of 50.1% have. However, this could be consider as little bit minor variation compared with both.

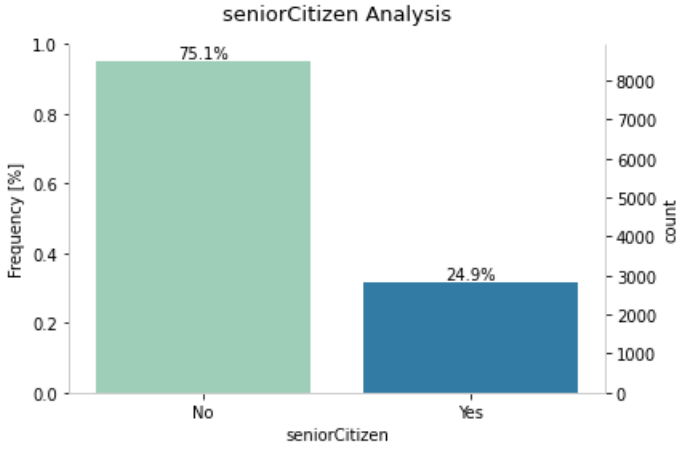


Figure 6.5 Analysis of Senior Citizen

As shown in Figure 6.5, not senior citizens have churned more than senior citizens, with a percentage of 75.1%. The variation is little bit high.

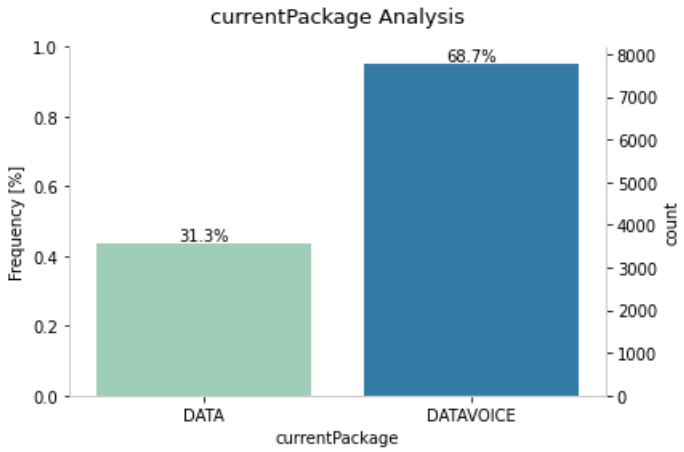


Figure 6.6 Analysis of Current Package

From Figure 6.6, the customers who have churned often are equipped with a current package with a percentage of 68.7. %. The variation is little bit high.

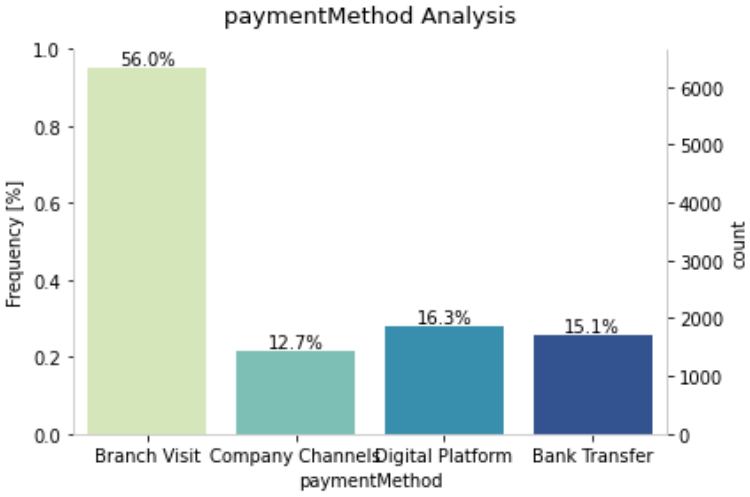


Figure 6.7 Analysis of Payment Method

According to Figure 6.7, the customers who have churned often have paid by visiting branch service with a percentage of 56%. The company channel factor is the lowest which show 12.7 percentage. Other two factors are digital platform and bank transfer. The values of both factors are 16.3% and 15.1%.

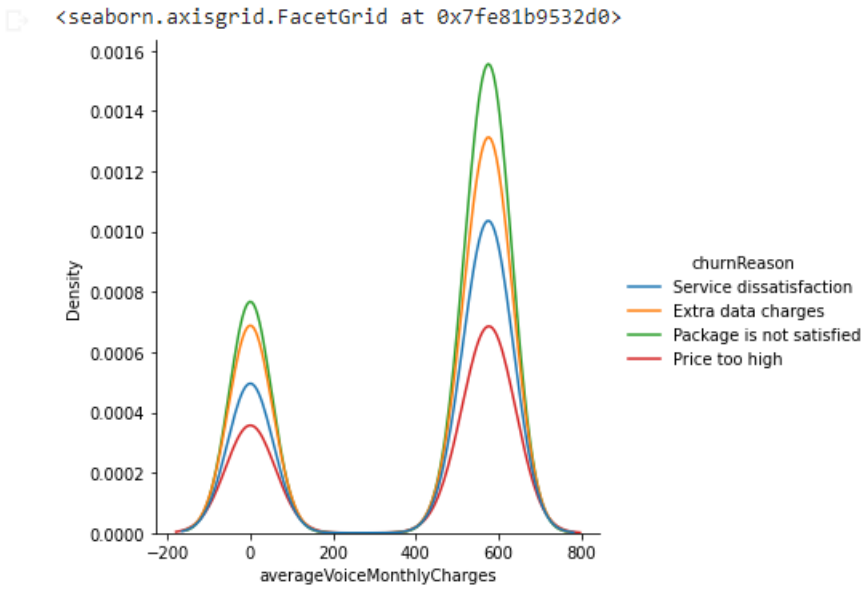


Figure 6.8 Analysis of Distribution Curves of Average Voice Monthly Charges

Figure 6.8 shows the distribution curves for average voice monthly charges for each churn reason. The customers have churned often-paid average voice monthly charges in the range of 500-700, for every churn reason.

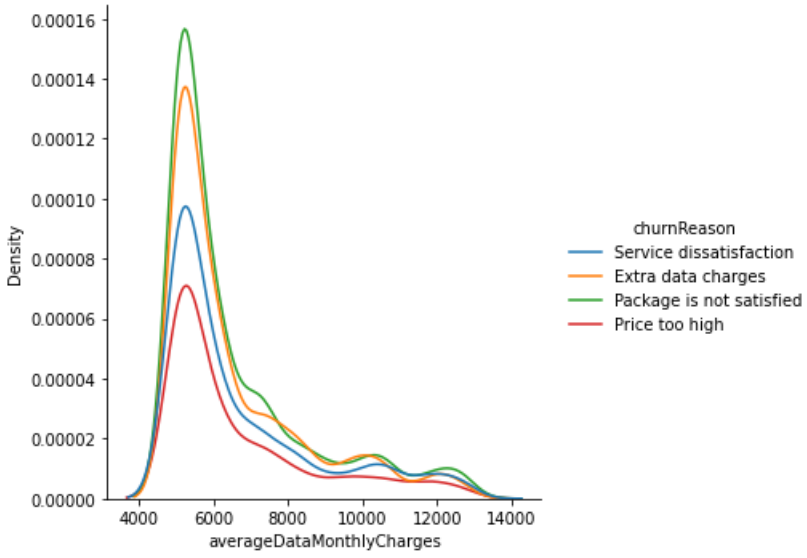


Figure 6.9 Analysis of Distribution Curves of Average Data Monthly Charges

Figure 6.9 shows the distribution curves for average data monthly charges for each churn reason. The customers have churned often-paid average data monthly charges in the range of 5000-6000, for every churn reason.

6.4 Data Preprocessing

Data preprocessing is one of the most important processes in the research study, which increases the accuracy of the dataset. To get best results from classification techniques, we have to have the particular dataset as an effective one.

6.5 Checking for Missing Values

It is required to check missing values as a step of data preprocessing. Because of it will impact to the final result of the evaluation. So, this has been done according to the Figure 6.10.

```
df.isnull().sum()
```

```
gender                          0
seniorCitizen                   0
tenureMonths                    0
multipleLines                   0
currentPackage                  0
contract                        0
paymentMethod                   0
averageVoiceMonthlyCharges      0
averageDataMonthlyCharges       0
churnReason                     0
dtype: int64
```

Figure 6.10 Result of Missing Values

By looking at the numbers we can understand that there are no missing values in each column of the data. That means dataset is good for now to proceed the next step of data preprocessing.

6.5.1 Remove Outliers

An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution errors. So, it's required to remove the outliers. According to Figure 6.11, outliers are removed using the Therefore IQR (Interquartile Range) technique.

```
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   gender                       10538 non-null  category
 1   seniorCitizen                10538 non-null  category
 2   tenureMonths                 10538 non-null  category
 3   multipleLines                10538 non-null  category
 4   currentPackage               10538 non-null  category
 5   contract                     10538 non-null  category
 6   paymentMethod                10538 non-null  category
 7   averageVoiceMonthlyCharges   10538 non-null  float64
 8   averageDataMonthlyCharges    10538 non-null  float64
 9   churnReason                  10538 non-null  category
 10  gender_Cat                   10538 non-null  int8
 11  seniorCitizen_Cat            10538 non-null  int8
 12  tenureMonths_Cat             10538 non-null  int8
 13  multipleLines_Cat            10538 non-null  int8
 14  currentPackage_Cat           10538 non-null  int8
 15  contract_Cat                 10538 non-null  int8
 16  paymentMethod_Cat            10538 non-null  int8
 17  churnReason_Cat              10538 non-null  int8
dtypes: category(8), float64(2), int8(8)
memory usage: 415.3 KB
```

Figure 6.11 IQR Technique Summary

The data set contained 11331 number of data and it has been reduced after the IQR technique is applied and final amount of data is 10538.

6.5.2 Categorical to Numerical Conversion

According to Figure 6.12, eight categorical variables are identified.

```
Data columns (total 10 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   gender                       10538 non-null  object
 1   seniorCitizen                10538 non-null  object
 2   tenureMonths                 10538 non-null  int64
 3   multipleLines                10538 non-null  object
 4   currentPackage               10538 non-null  object
 5   contract                     10538 non-null  object
 6   paymentMethod                10538 non-null  object
 7   averageVoiceMonthlyCharges   10538 non-null  float64
 8   averageDataMonthlyCharges    10538 non-null  float64
 9   churnReason                  10538 non-null  object
```

Figure 6.12 Categorical Variables

After applying the categorical to numerical conversion algorithm, the eight variables are converted to numerical. It is indicated in Figure 6.13.

```
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   gender                       10538 non-null  category
 1   seniorCitizen                10538 non-null  category
 2   tenureMonths                 10538 non-null  category
 3   multipleLines                10538 non-null  category
 4   currentPackage               10538 non-null  category
 5   contract                     10538 non-null  category
 6   paymentMethod                10538 non-null  category
 7   averageVoiceMonthlyCharges   10538 non-null  float64
 8   averageDataMonthlyCharges    10538 non-null  float64
 9   churnReason                  10538 non-null  category
 10  gender_Cat                   10538 non-null  int8
 11  seniorCitizen_Cat            10538 non-null  int8
 12  tenureMonths_Cat             10538 non-null  int8
 13  multipleLines_Cat            10538 non-null  int8
 14  currentPackage_Cat           10538 non-null  int8
 15  contract_Cat                 10538 non-null  int8
 16  paymentMethod_Cat            10538 non-null  int8
 17  churnReason_Cat              10538 non-null  int8
dtypes: category(8), float64(2), int8(8)
memory usage: 415.3 KB
```

Figure 6.13 Numerical Conversion Summary of Dataset

Further, The X and y variables are initialized before move to the next step of data preprocessing, according to the Figure 6.14,

```
X = df[['gender_Cat', 'seniorCitizen_Cat','tenureMonths_Cat','multipleLines_Cat','currentPackage_Cat','contract_Cat','paymentMeth
y = df['churnReason_Cat']
```

Figure 6.14 X and y Variables Initialized

6.5.3 Class Imbalance to Balance

The class imbalance occurs when one class is higher than in other classes. In machine learning Class Imbalance is a common problem, especially in classification problems. Imbalance records is laid low with the accuracy of the model and may impede our model accuracy. According to Figure 6.15, the classes are imbalanced initially. The initial number of data for the class variables are, Extra data charges 3070, Package is not satisfied 3404, Price too high 1722 and Service dissatisfaction 2342. The Package is not satisfied shows the maximum percentage compare with other three class variables. It is 32.302% percentage value. The lowest value shows Price too high class variable. It is 16.341 percentage value.

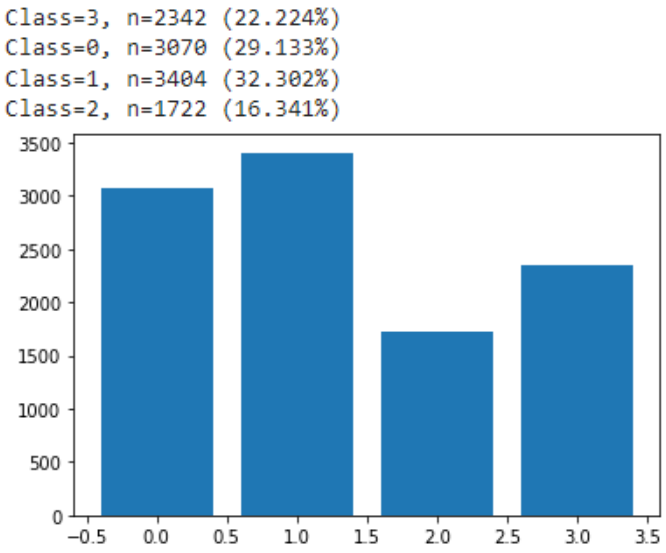


Figure 6.15 Class Imbalance Behavior

After applying the SMOTE algorithm [12], the classes are balanced. It is indicated in Figure 6.16.

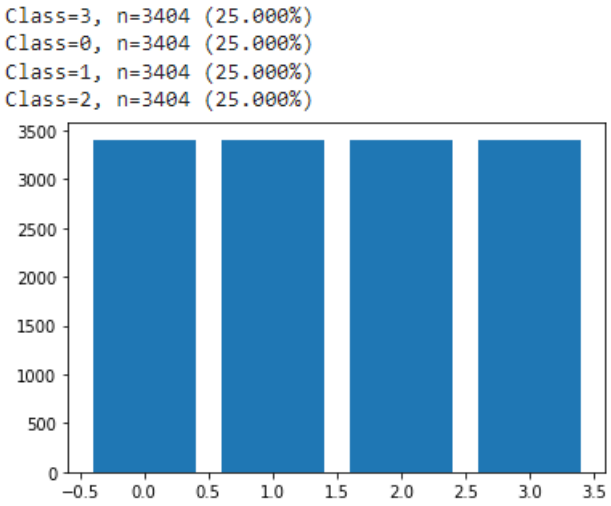


Figure 6.16 Balanced Class Variable Summary

The variables Class 0 describe about Extra data charges. The variable Class 1 describe about Package is not satisfied. The variable Class 2 describe about Price too high. The variable Class 3 describe about Service dissatisfaction. The value of balance classes are 3404 and it is 25% percentage.

6.5.4 Applying Appropriate Classification Technique

The Logistic Regression, Naive Bayes, Random Forest, DT (Decision Tree), KNN (K-Nearest Neighbor), SVM (Support Vector Machine), Gradient Boost Classifier and Hybrid Model (Decision Tree Classifier, Random Forest Classifier and Gradient Boosting Classifier) classification algorithm are used. For that, inbuilt DM classification used in python language libraries. The Voting Classifier is used to build hybrid model. The models training and testing is done using K-fold cross-validation technique (APPENDIX A).

In the code, a classifier variable is created and assign String values to it. It is an array variable and it contains the names of data mining techniques that are used in the study. Also, "y_pos" and "score" variables are created to save the result and to show it in relevant places. Then, each and every inbuilt algorithm is taken into consideration one by one. The inbuilt data mining algorithm is assigned to the variable and the "fit" function is used to pass X, y variables prepared. The cross_val_score function is used to train and test the model. End of these steps, a trained and tested model can be taken. This model is used to show the mean score and to get predicted results by passing sample values without including class variables. Because, the class variable is the predicated feature of this study. Also, feature importance is plotted and confusion matrix is plotted by using the final model. The Logistic Regression implementation has been shown in the APPENDIX C.

6.6 Summary

The implementation of the study was discussed in this chapter. The loading to the web tool has been shown and data analysis and visualization result described. The Logistic Regression, Naive Bayes, Random Forest, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Gradient Boost Classifier and Hybrid Model have been used for implementation. The next chapter will be explaining about the evaluation of the study.

Chapter 07

Evaluation

7.1 Introduction

The previous chapter has described about implementation of the study. It mainly describe the data gathering, data analysis and visualization, data preprocessing and classification DM technique. The way of model creation part included into that section. The Evaluation will be discussed in this chapter.

7.2 Evaluation for Classification

The confusion matrix and mean score are two accuracy representation. In this study used both two representations for the evaluation and study is proceed according to the evaluation measurement of confusion matrix.

7.2.1 Mean Score

The scores mean where calculated to each and every DM classification technique which are used in this study to compare accuracy. The result is shown in Table 7.2

No	DM Classification Technique	Accuracy (%)
1	Logistic Regression	26.15
2	Naive Bayes	26.15
3	**Random Forest**	**90.94**
4	**Decision Tree**	**92.44**
5	K-Nearest Neighbor	54.57
6	Support Vector Machine	25.12
7	Gradient Boost Classifier	80.81
8	Hybrid Model	56.58

Table 7.1 Classification Performance Numerical Mean Scores

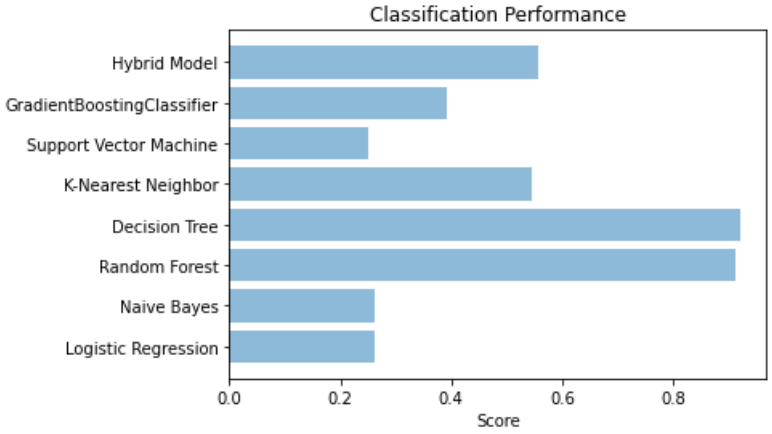


Figure 7.1 Classification Performance Graphical Mean Scores

Figure 7.1 shows the graphical representation of the performance for mean score. According to Figure 7.1 and Table 7.2 decision tree and random forest classifiers show maximum performance compare with other classifiers. The decision tree shows higher value according to the mean score.

7.2.2 Confusion Matrix

Confusion matrix is one of the important outputs from the classifications (Shown in APPENDIX B). Using the confusion matrix, we can find a number of evaluation factors such as accuracy, precision, recall and, F-1 score to evaluate data mining classification models. These measurements and their definition are given in the following Table 7.2.

Measure	Relevant Formula
Accuracy	(TP+TN)/ (TP+TN+FP+FN)
Precision	TP/ (TP+FP)
Recall	TP/ (TP+FN)
F1-score	2TP/(2TP+FP+FN)

Table 7.2 Evaluation Measurements for Classifiers

Therefore, confusion matrix is calculated for each classification technique and calculated each measurement in that Table 7.2

The accuracy, precision, recall and, F-1 score evaluation measurements considered to make the evaluation furthermore in this study.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.27	0.26	0.26	0.26
Naive Bayes	0.29	0.29	0.29	0.29
K-Nearest Neighbor	0.54	0.73	0.73	0.73
Support Vector Machine	0.25	0.25	0.25	0.25
Decision Tree	0.92	0.99	0.99	0.99
Random Forest	**0.93**	**0.99**	**0.99**	**0.99**
Gradient Boost Classifier	0.39	0.45	0.45	0.45
Hybrid Model	0.56	0.60	0.60	0.60

Table 7.3 Classification Performance Numerical Scores Based on Confusion Matrix

In order to evaluate the models, the values of Table 7.3 are considered and according to it, decision tree and random forest classifiers show maximum performance compare with other classifiers and random forest classification shows higher value according to Table 7.3 values.

7.2.3 Feature Importance

Score of all input features for a given model are calculated using feature importance techniques and that score represent the “importance” of each feature. The feature has a larger effect on the model shows higher score. The decision tree and random forest classifiers feature importance are calculated due to it shows the higher accuracy value for in this study.

According to the Table 7.4, Figure 7.2 and Figure 7.3 the large effective feature is Average Data Monthly Charges. The features Average Voice Monthly Charges and Tenure Months shows higher feature importance score as second and third feature.

No	Feature	Random Forest Score	Decision Tree Score
0	Gender	0.01573	0.01033
1	Senior Citizen	0.01085	0.00297
2	Tenure Months	0.24179	0.22760
3	Multiple Lines	0.01581	0.01552
4	Current Package	0.01361	0.01190
5	Contract	0.02232	0.02986
6	Payment Method	0.05694	0.06711
7	Average Voice Monthly Charges	0.22963	0.23429
8	Average Data Monthly Charges	0.39332	0.40043

Table 7.4 Feature Importance Numerical Scores

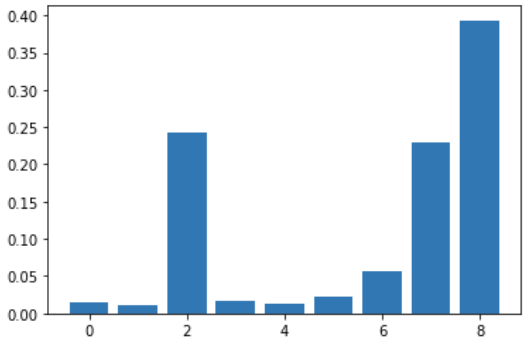


Figure 7.2 Feature Importance Graphical Scores for Random Forest

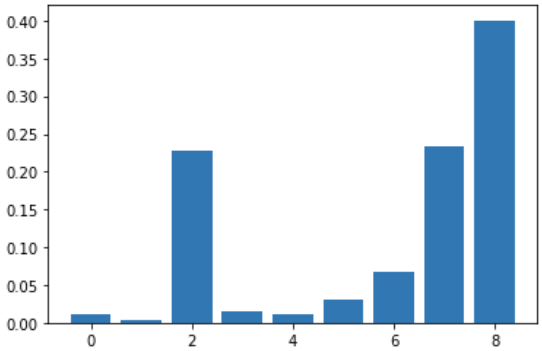


Figure 7.3 Feature Importance Graphical Scores for Decision Tree

7.3 Summary

The Mean score and confusion matrix measurement have been conceded to the evaluation and the accuracy, precision, recall and, F-1 score evaluation measurements considered to make the evaluation furthermore in this study. Acceding to the results, decision tree and random forest classifiers show maximum performance compare with other classifiers and random forest classification shows higher value. The next chapter discussed about conclusion and future works of this study.

Chapter 08

Conclusion and Future Work

8.1 Introduction

The previous chapter has described the evaluation of the study. It describe about two evaluation factors which are mean score and Confusion Matrix. But, evaluation decision has been taken through the classification performance numerical scores based on Confusion Matrix. It compared the DM technique result according to the accuracy, precision, recall and, F-1 score values. The conclusion and future works will be discussed in this chapter.

8.2 Conclusion

According to the previous research topics analysis, most of the studies are based on customer churn prediction and few research are focused on the customer churn reasoning part. Therefore, this study tried to touch that gap.

The dataset has been gathered from one of the leading telecommunication industries and we have considered main 10 features such as gender, senior citizen, tenure months, multiple lines, current package, contract, payment method, average voice monthly charges, average data monthly charges and, churn reason. The churn reason was the class variable of the dataset. Finally, based on this feature set proceed the study using DM eight classification techniques. The decision tree and random forest classifiers showed best results and random forest classifiers were chosen as suitable classifiers from both classifiers. To evaluate that confusion matrix was used and accuracy, precision, recall and F-1 score evaluation measurements considered.

The confusion matrix result is shown in the Appendix B for each DM classification techniques.

The future importance was calculated for decision tree and random forest classifiers and Average Data Monthly Charges was large effective feature from all ten input. The Average Voice Monthly Charges and Tenure Months shows higher feature importance score as second and third feature.

8.3 Limitation

The telecommunication industry data are directly connected with customers. Therefore, most of the data are confidential and unable to access. As example customers used packages (voice or data), CDR data of customers, customer's current status (IDD, NIDD, OGBD or TEMP), etc. Hence, feature selection can be limited. In other side, limitation in annotated data set and identifying the actual churn reason are challengers of this study.

8.4 Future Developments

Some of the valuable features are not considered for the initial dataset due to the limitation of the industry. As example, specific data packages, weekly average recharge, weekly average data usage, feature subscription, etc. In future, more features can be considered. Further, this study has been narrowed down to the data package. Therefore, voice packages can also be considered in future. This study is based on four class variables (churn reasoning) and class variables can be increased in future to study this area further. The deployment part has not considered in this study and it can be considered to develop in future. Then user can use this implementation for their future works.

8.5 Summary

This chapter concluded the conclusion, limitation and extended future work. According to the conclusion of the study, they show random forest classifiers were chosen as suitable classifiers according to the selected features of the dataset.

Reference

[1] A.Keramati, R. Marandi, M.Aliannejadi, I.Ahmadian, M.Mozaffari and U.Abbasi, "Improved churn prediction in telecommunication industry using data mining techniques," Science Direct, vol. 24, p. 994–1012, 2014.

[2] S. Y. Hung, D. C.Yen and H. Y. Wang, "Applying data mining to telecom churn management," Expert Systems with Applications, vol. 31, p. 515–524, 2006.

[3] b. O. Khalida , b. M. S. Sunarti , A. H. Norazrina and . b. A. B. Faizin, "Data Mining in Churn Analysis Model for Telecommunication Industry," Statistical Modeling and Analytics, vol. 1, pp. 19-27, 2010.

[4] N. P. P. T. Naren , "Customer Churn Prediction Using Big Data Analytics," 2016.

[5] M. Vishal , M. Richa and M. Renuka , "Data Analysis Techniques and Strategies," Review on factors affecting customer churn in telecom sector, vol. 9, p. 122–144, 2017.

[6] S. Farid and M. Mahbobeh , "A big data analytics model for customer churn prediction in the retiree segment," Information Management, vol. 48, pp. 238-253, 2019.

[7] . H. Bingquan, T. K. Mohand and . B. Brian, "Customer churn prediction in telecommunications," Expert Systems with Applications, vol. 39, p. 1414–1425, 2012.

[8] H. Nabgha, A. B. Naveed and I. Dr.Muddesar , "Customer Churn Prediction in Telecommunication A Decade Review and Classification," International Journal of Computer Science Issues (IJCSI), p. 271, 2013.

[9] A. Adnan , A. Sajid , A. Awais , . N. Muhammad, A. Khalid , H. Amir and H. Kaizhu , "Customer churn prediction in the telecommunication sector using a rough set approach," Science Direct, vol. 237, pp. 242-254, 2017.

[10] A. Q. Saad , S. R. Ammar , M. Q. Ali and K. Aatif , "Telecommunication Subscribers' Churn Prediction Model Using Machine Learning," Eighth international conference on digital information management (ICDIM 2013), 2013.

[11] D. Kiran and B. Surbhi , "Customer Churn Analysis in Telecom Industry," pp. 1-6, 2015.

[12] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hal and W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Artificial Intelligence Research, vol. 16, p. 321–357, 2002.

[13] H. Monireh and S. Mostafa , "New approach to customer segmentation based on changes in customer value," Journal of Marketing Analytics, pp. 110-121, 2015.

[14] T. Landis, "Customer Retention Marketing vs. Customer Acquisition Marketing," 12 04 2022. [Online]. Available: https://www.outboundengine.com/blog/customer-retention-marketing-vs-customer-acquisition-marketing/.

[15] Mrs. Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS," Computer Science and Engineering, vol. 1, pp. 301-305.

[16] H. Ying , Q. H. Bing and K. M. Tahar , "A Rule-Based Method for Customer Churn Prediction in Telecommunication Services," vol. 6634, pp. 411-422, 2011.

APPENDIX A

Logistic Regression Classifier

The inbuilt LogisticRegression function is used which is exist inside the sklearn.linear_model library of python. And passed X, and y variables to the fit function. Then used cross_val_score function which is exist under sklearn.model_selection library.

```
clf = LogisticRegression(multi_class='multinomial', solver='lbfgs', penalty='none')
clf.fit(X,y)
scores = cross_val_score(clf, X, y,cv=5)
score.append(scores.mean())
```

Naive Bayes Classifier

The inbuilt GaussianNB function is used which is exist inside the sklearn.naive_bayes library of python. And passed X, and y variables to the fit function. Then used cross_val_score function which is exist under sklearn.model_selection library.

```
# Fit the model on train
gnb = GaussianNB()
clf.fit(X,y)
scores = cross_val_score(clf, X, y,cv=5)
score.append(scores.mean())
```

Random Forest Classifier

The inbuilt RandomForestClassifier function is used which is exist inside the sklearn.ensemble library of python. And passed X, and y variables to the fit function. Then used cross_val_score function which is exist under sklearn.model_selection library.

```
clf = RandomForestClassifier(n_estimators=10)
clf.fit(X,y)
scores = cross_val_score(clf, X, y,cv=5)
score.append(scores.mean())
```

Decision Tree Classifier

The inbuilt DecisionTreeClassifier function is used which is exist inside the sklearn.tree library of python. And passed X, and y variables to the fit function. Then used cross_val_score function which is exist under sklearn.model_selection library.

```
clf = DecisionTreeClassifier()
clf.fit(X, y)
scores = cross_val_score(clf, X, y,cv=5)
score.append(scores.mean())
```

K-Nearest Neighbor Classifier

The inbuilt KNeighborsClassifier function is used which is exist inside the sklearn.neighbors library of python. And passed X, and y variables to the fit function. Then used cross_val_score function which is exist under sklearn.model_selection library.

```
clf = KNeighborsClassifier()
clf.fit(X,y)
scores = cross_val_score(clf, X, y,cv=5)
score.append(scores.mean())
```

Support Vector Machine Classifier

The inbuilt svm.SVC function is used which is exist inside the sklearn library of python. And passed X, and y variables to the fit function. Then used cross_val_score function which is exist under sklearn.model_selection library.

```
from  sklearn import svm
clf=svm.SVC(probability=True)
clf.fit(X,y)
scores = cross_val_score(clf, X, y,cv=5)
score.append(scores.mean())
```

Gradient Boost Classifier

The inbuilt GradientBoostingClassifier function is used which is exist inside the sklearn.ensemble library of python. And passed X, and y variables to the fit function. Then used cross_val_score function which is exist under sklearn.model_selection library.

```
from sklearn.model_selection import cross_val_predict
from sklearn.ensemble import GradientBoostingClassifier
clf=GradientBoostingClassifier()
clf.fit(X,y)
scores = cross_val_score(clf, X, y,cv=5)
score.append(scores.mean())
```

Hybrid Model

- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier

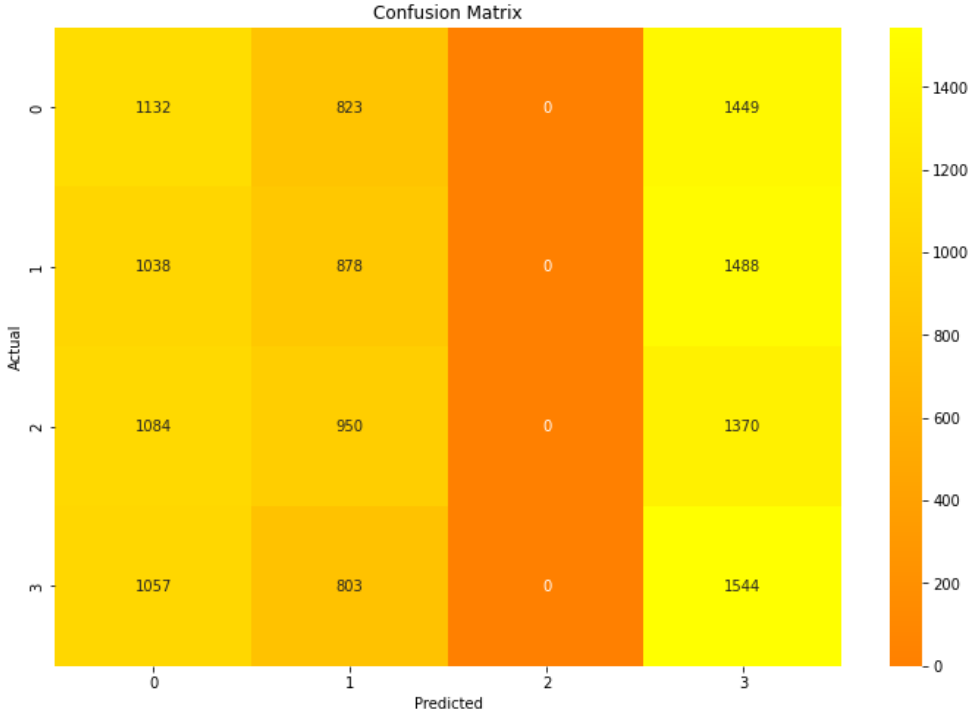
The inbuilt VotingClassifier function is used which is exist inside the klearn.ensemble library of python. And passed X, and y variables to the fit function. Then used cross_val_score function which is exist under sklearn.model_selection library.

```
# Defining the ensemble model
clf = VotingClassifier(estimators)
clf.fit(X, y)
scores = cross_val_score(clf, X, y,cv=5)
score.append(scores.mean())
```

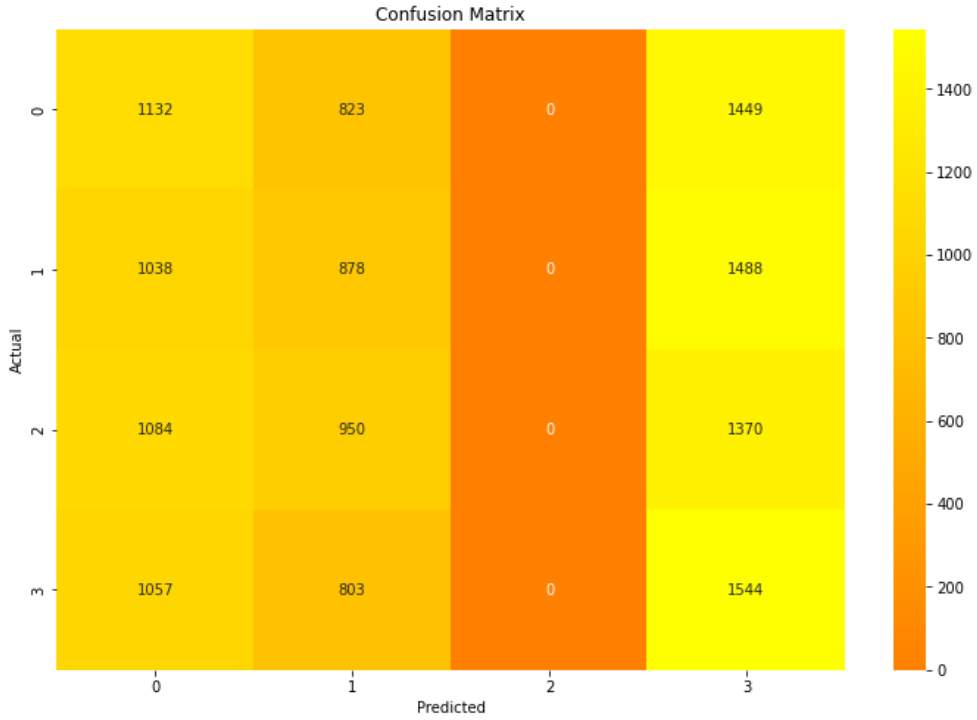
APPENDIX B

This appendix shows confusion matrix each and every data mini techniques.

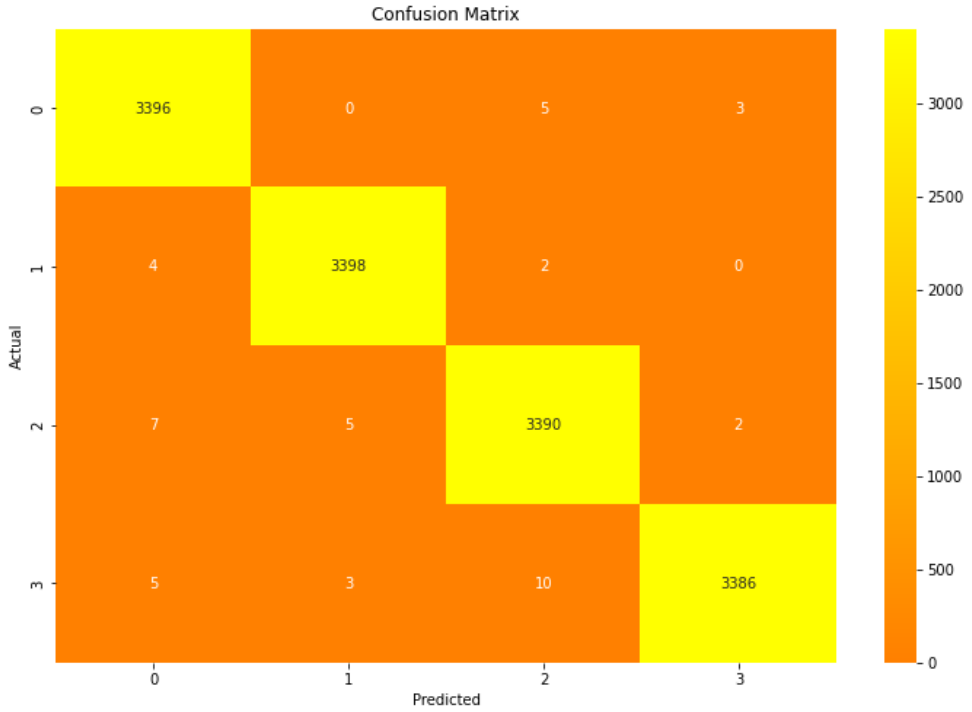
Logistic Regression Classifier



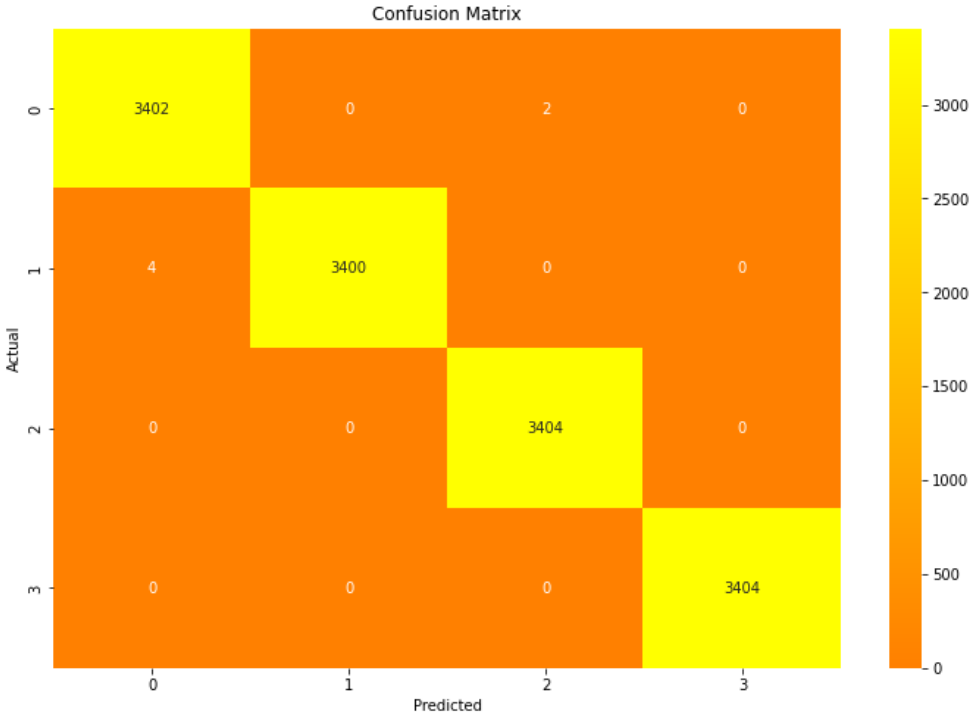
Naive Bayes Classifier



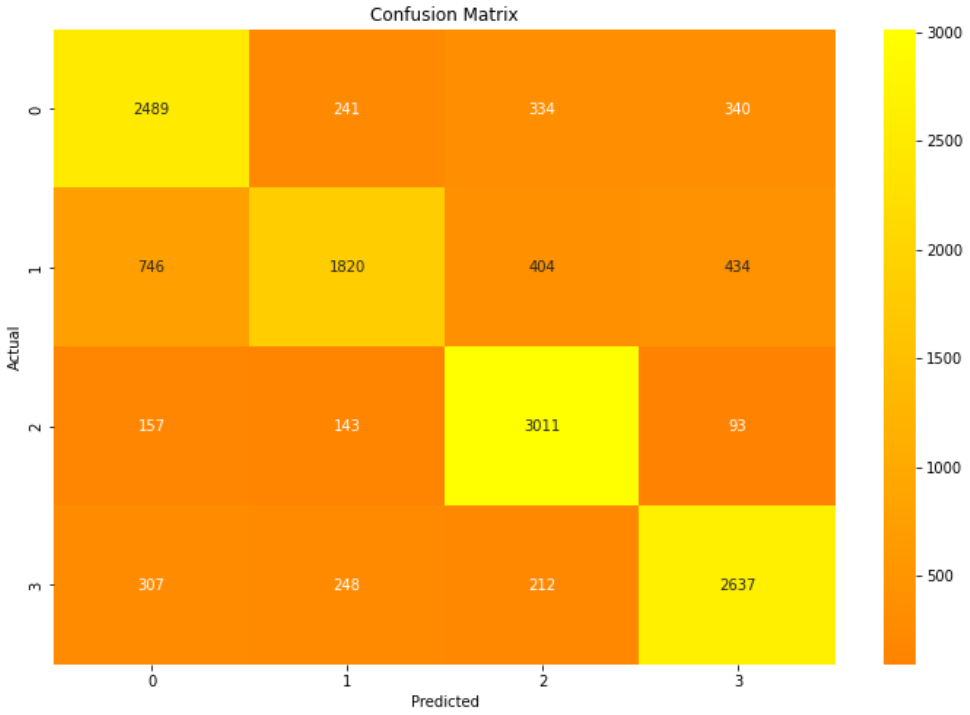
Random Forest Classifier



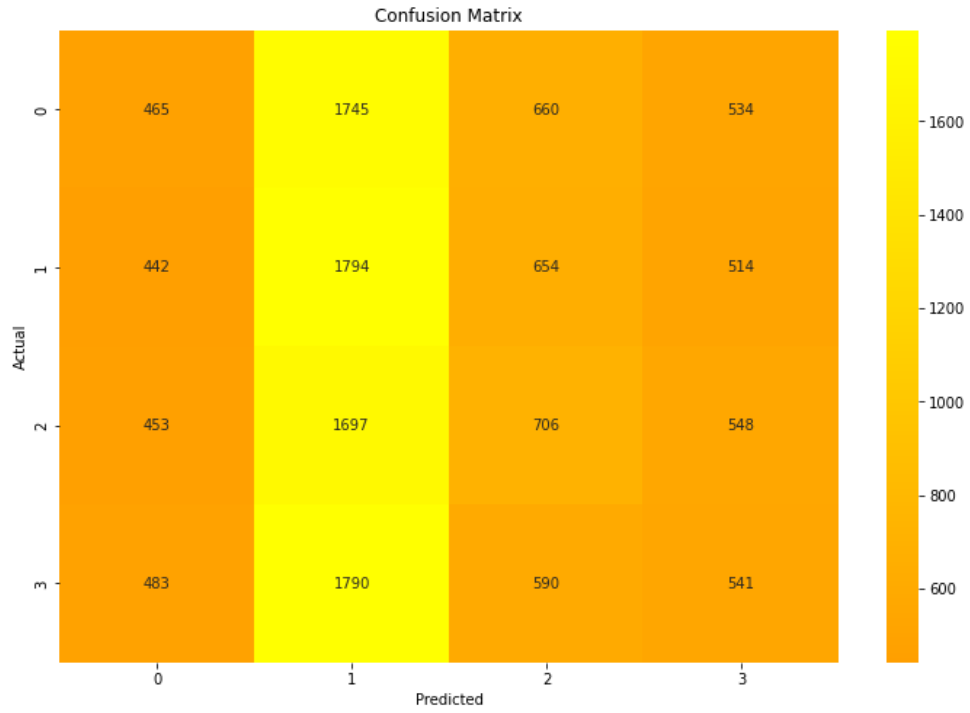
Decision Tree Classifier



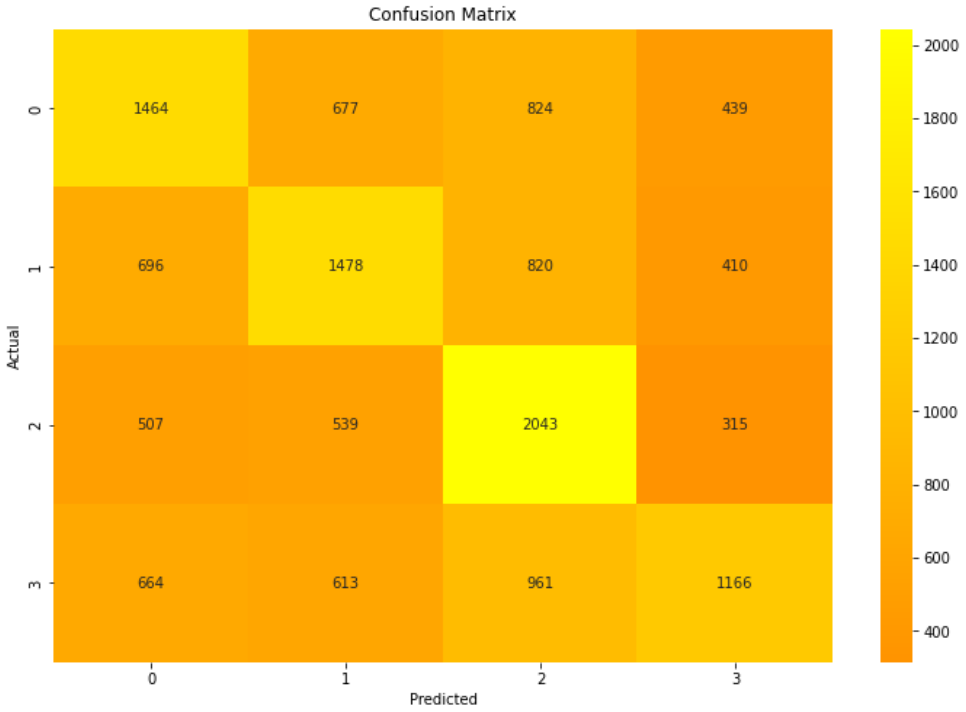
K-Nearest Neighbor Classifier



Support Vector Machine Classifier

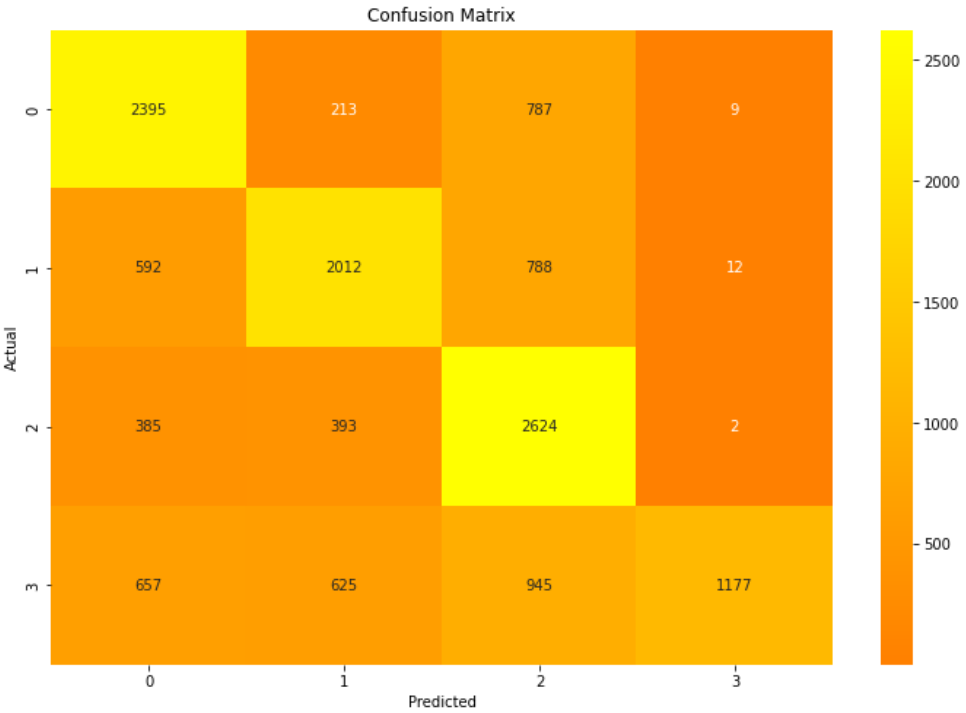


Gradient Boost Classifier



Hybrid Model

- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier



APPENDIX C

```
clf = LogisticRegression(multi_class='multinomial', solver='lbfgs', penalty='none')
clf.fit(X,y)
scores = cross_val_score(clf, X, y,cv=5)
score.append(scores.mean())
print('The accuration of classification is %.2f%%' %(scores.mean()*100))
```

```
C:\Users\Kasun\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn\linear_model\_logistic.py:444: ConvergenceWarn
ing: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

```
The accuration of classification is 26.15%
```

```
from numpy import array

Xnew = array([[0,1,18,0,0,1,0,100.75,11171.15]])
ynew = clf.predict(Xnew)
# show the inputs and predicted outputs
print("X=%s, Predicted=%s" % (Xnew[0], ynew[0]))
```

```
X=[0.000000e+00 1.000000e+00 1.800000e+01 0.000000e+00 0.000000e+00
 1.000000e+00 0.000000e+00 1.007500e+02 1.117115e+04], Predicted=0
```