

**CUSTOMER CHURN PREDICTION
TELECOMMUNICATIONS IN MOGADISHU, SOMALIA USING
MACHINE LEARNING**

**ZAKARIA ALI HASSAN
ABDIJIBAR JAMA MOHAMED
ADAN SALEBAN ALI
ABSHIR MUSE AHMED**



**SUBMISSION OF GRADUATION PROJECT FOR
PARTIAL FULFILLMENT OF THE
DEGREE OF BACHELOR OF SCIENCE IN
COMPUTER APPLICATIONS**

**JAMHURIYA UNIVERSITY OF SCIENCE AND
TECHNOLOGY (JUST)
FACULTY OF COMPUTER & INFORMATION
TECHNOLOGY**

AUGUST 2024

JAMHRURIYA UNIVERSITY OF SCIENCE AN TECHNOLOGY (JUST)

Original Literary Work Declaration

Name of Candidate 1: **ZAKARIA ALI HASSAN**

ID No: C120098

Name of Candidate 2: **ABDIJIBAR JAMA MOHAMED**

ID No: C1201144

Name of Candidate 3: **ADAN SALEBAN ALI**

ID No: C1201062

Name of Candidate 4: **ABSHIR MUSE AHMED**

ID No: C120144

Name of Degree: Bachelor of Computer Application

Title of Project Paper/Research Report/Dissertation/Thesis (“this Work”):

CUSTOMER CHURN PREDICATION USING MACHINE LEARNING

Field of Study: Computer Application. We the undersigned, do solemnly and sincerely declare that: (1) We are the authors/writers of this Work; (2) This Work is original; (3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work; (4) We do not have any actual knowledge nor ought I reasonably to know that the making of this work constitutes an infringement of any copyright work; (5) We hereby assign all and every rights in the copyright to this Work to Jamhuuriya University of Science and technology (“JUST”), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of JUST having been first had

and obtained; (6) We are fully aware that if in the course of making this Work, we have infringed any copyright whether intentionally or otherwise, we may be subject to legal action or any other action as may be determined by JUST.

Candidate 1's Signature: _____ Candidate 2's Signature: _____

Candidate 3's Signature: _____ Candidate 4's Signature: _____

Date: _____ Subscribed and solemnly declared before,

Supervisor's Signature: _____ Date: _____

DEDICATION

We are happy to thank our parents for all the help they have given us and for their sacrifice to support us to reach this great milestone especially the completion of this project, in addition to all our relatives and everyone who assisted us to succeed in this project, we would like to thank you all.

ACKNOWLEDGEMENT

First, we would like to thank Allah the creator of this universe for giving us good health and well-being to complete this long journey as well as coordinating this book. **Secondly**, we are thankful to our mother and father who have been the backbone of our lives especially the foundation of our education, all our brothers and sisters. **Thirdly**, we would like to express our sincere gratitude our teacher and supervisor **Eng. Mohamed Abdi Karin** for his constant encouragement, guidance, and commitment to the success of our research and we would like to than **Eng. Sharmake Ali Mohamed** as an assist supervisor. We would also like to thank **Eng. Mohamed Abdullahi Mohamud Shuuriye** Head of Research at the Faculty of Computer Science and IT, has always been very supportive and has provided useful advice. We would also like to thank the honorable **UNIVERSITY OF SCIENCE AND TECHNOLOGY. (J U S T)** Who gave us excellent opportunity to finish our bachelor degree of computer application, it is chance to take this opportunity to thank all the lectures of the faculty of computer application and IT for their loyalty and guidance toward the bright future, especially the dean of faculty **Eng. Mohamed Abdullahi Ali**. Finally, we would like to deliver our special thanks to all those who have helped us, both directly and indirectly. All praise is due to Allah

All blesses to Allah.

ABSTRACT

This thesis presents a comprehensive approach to predicting customer churn in the telecommunications sector of Mogadishu, Somalia, using machine learning techniques. Customer churn, the rate at which customers discontinue their service, poses a significant challenge and financial impact on telecom providers. Through the deployment of various machine learning models such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines, this study aims to enhance predictive accuracy and inform targeted customer retention strategies. The research employs a robust methodology encompassing data preprocessing, feature selection, and model evaluation to address the nuances of customer behavior patterns effectively. The findings demonstrate the superior performance of Random Forest models in predicting potential churn, thereby enabling proactive customer retention efforts. This work not only contributes to the theoretical frameworks of churn prediction but also provides practical insights that could be leveraged by telecom operators to reduce churn rates and improve service continuity in competitive markets.

Table of Contents

DEDICATION	iv
ACKNOWLEDGEMENT	v
ABSTRACT	vi
List of Abbreviations	1
List of Figures	3
List of tables	5
CHAPTER ONE: INTRODUCTION	6
1.1 Introduction	6
1.2 Background of the study	7
1.3 Problem statement	8
1.4 Motivation of the study	9
1.5 Objectives of the study	10
1.5.1 To develop a Better Prediction Model	10
1.5.2 To design a beautiful user interface	10
1.6 Research questions	10
1.7 significance of study	10
1.8 Scope of the study	11
1.8.1 Geographical scope of this project	11

1.8.2	Time scope of this project	11
1.9	Organization of the study;	11
CHAPTER TWO: LITERATURE REVIEW		13
2.1	Introduction.....	13
2.2	Customer Churn Prediction	14
2.3	History of Customer churn	16
2.4	Data Exploration and Pre-processing	17
2.4.1	Class Imbalance.....	18
2.4.2	Feature Selection	20
2.5	Machine Learning.....	22
2.5.1	Supervised Machine Learning.....	23
2.6	Machine Learning Techniques.....	25
2.6.1	Logistic Regression.....	25
2.6.2	Random Fores	26
2.6.3	Support Vector Machine.....	27
2.6.4	Neural Network.....	27
2.7	Model Evaluation.....	28
2.8	Customer Churn Prediction using Machine Learning	29

2.9 Approaches to solve the problem	29
2.10 Related works.....	32
CHAPTER THREE: RESEARCH METHODOLOGY	36
3.0 Introduction	36
3.1 System Description.....	36
3.2 System Architecture.....	36
3.3 System Features	37
3.3.1 Customer Churn Prediction:	37
3.3.2 Data Visualization:	38
3.3.3 User Management:	38
3.3.4 Statistical Dashboard:	38
3.3.5 Security Features:.....	38
3.4 System Methodology	39
3.4.1 Dataset.....	39
3.4.2 Data preparation	39
3.4.3 Model Selection.....	40
3.4.3.1 Decision Tree Classifier:	41
3.4.3.2 Random Forest Classifier:.....	41

3.4.3.3 Support Vector Machine (SVM):	41
3.4.3.4 Logistic Regression:	42
3.4.4 Model Evaluation	42
3.4.5 Implementation	42
3.5 System development environment	43
3.5.1 Front End	43
3.5.2 Back End	43
3.5.3 Python	44
3.5.4 Visual studio (vs code)	45
3.5.5 Jupyter Notebook	45
3.6 Hardware requirement	46
3.7 Software Requirements	46
CHAPTER FOUR: ANALYSIS AND DESIGN	48
4.1. Introduction	48
4.2. System analysis	48
4.3. Existing Approaches	48
4.4. The Proposed System	49
4.5. System Requirements	49

4.5.1.	Functional Requirements	50
4.5.2	Non-Functional Requirements	50
4.6	Feasibility Study	51
4.6.1	Technical Feasibility	51
4.6.2	Economic Feasibility.....	51
4.6.3	Operational Feasibility	51
4.6.4	Schedule Feasibility	52
4.7	System Design	52
4.7.1	Data Flow Diagrams (DFD).....	52
4.8	Dataset Design.....	55
CHAPTER FIVE: IMPLEMENTATION AND TESTING		56
5.1	Introduction.....	56
5.2	Overview of the implementation environment.....	56
5.3	Machine learning models Evaluation Results	57
5.3.1	Logistic Regression.....	57
5.3.2	Decision Tree.....	59
5.3.3	Random Forest	61
5.3.4	Support Vector Machine.....	63

5.4 Snapshots of the system.....	68
5.4.1 Front-end	69
5.4.2 Back-end.....	71
5.4.3 Data Visualization	73
CHAPTER 6: Discussion and Results	83
6.1 Discussion.....	83
6.2 Results	84
CH-7: CONCLUSION AND FUTURE WORK.....	85
7.1 Introduction.....	85
7.2 Conclusion	85
7.3 Recommendation	86
7.3.1 Update and Upgrade Models.....	86
7.3.2 Expand Data Sources:	86
7.3.3 Enhance Data Collection Methods.....	86
7.4 Future work.....	87
7.4.1 Development of Predictive Analytics Tools.....	87
7.4.2 Regular Reporting	87
7.4.3 Identifying Reasons for Churn.....	87

7.4.4 Proactive Customer Engagement:	87
REFERENCE	89
Appendix A: Prediction Page	96
Appendix B: SERVER	111

List of Abbreviations

XGB	Extreme Gradient Boosting
CRM	Customer Relationship Management
SVM	Support Vector Machine
CRISP	CRISP-DM (Cross-Industry Standard Process for Data Mining)
DM	Data Mining
SMOTE	Synthetic Minority Over-sampling Technique
SAS	Statistical Analysis System (also a software suite)
MLP	Multilayer Perceptron
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
TDL	Test Description Language
RFM	Recency, Frequency, Monetary Value Analysis
MATLAB	Matrix Laboratory
ANN	Artificial Neural Network
US	United States
SOM	Self-Organizing Map
WEKA	Waikato Environment for Knowledge Analysis
JR	Java Report
DT	Decision Tree
UCI	University of California, Irvine
REST	Representational State Transfer
API	Application Programming Interface
TV	Television
HTML	HyperText Markup Language
CSS	Cascading Style Sheets
GUI	Graphical User Interface
BSD	Berkeley Software Distribution
IRC	Internet Relay Chat
IDE	Integrated Development Environment

UI	User Interface
JSON	JavaScript Object Notation
PHP	Hypertext Preprocessor
SQL	Structured Query Language
XML	eXtensible Markup Language
ASP	Active Server Pages
NET	Microsoft .NET Framework
RAM	Random Access Memory
TM	Trademark
CPU	Central Processing Unit
GH	GitHub
GB	Gigabyte
ROI	Return on Investment
ML	Machine Learning
HTTP	Hypertext Transfer Protocol
SVC	Support Vector Classification
IJERT	International Journal of Engineering Research and Technology
IS	Information Systems
ICISS	International Conference on Information Systems Security
IJDATS	International Journal of Data Analysis Techniques and Strategies
INDECS or software)	Information DECS (not clearly identified, potentially a local term
CDAN	Cyber Defense Analysis Network
EPJ	European Physical Journal
IEEE	Institute of Electrical and Electronics Engineers
ICT	Information and Communication Technology
DSL	Digital Subscriber Line
JS	java Script

List of Figures

Figure 2.1: Complete CRISP-DM.....	17
Figure 2.2: performance	18
Figure 2.3:Feature Selection Category.....	21
Figure 2.4: Machine Learning Techniques	23
Figure 2.5: Supervised Machine Learning Model	24
Figure 2.6: Logistic Regression Formula.....	25
Figure 2.7: Churn Rate Prediction using Machine Learning	28
Figure 2.8. Alyuda – Network Topology.....	34
Figure 2.9. Alyuda – Network Training Op	34
Figure 2.10. Alyuda – Network Training R 1	35
Figure 2.11. Alyuda – Query.	35
Figure 3.1: System Architecture.....	37
Figure 4.1: Data flow Diagram of customer churn prediction.....	54
Figure 4.2: Diagram 0 DFD of customer churn prediction.....	54
Figure 4.3: Dataset Design.....	55
Figure 5.1: Logistic Regression Classification report for a balanced dataset..	57
Figure 5.2: Logistic Regression Classification report for imbalanced dataset	58
Figure 5.3: Decision Tree Classification for balanced dataset.....	60
Figure 5.4: Decision Tree Classification report for imbalanced dataset.....	60
Figure 5.5: Random Forest Classification 1	62

Figure 5.6: Random Forest Classification report for imbalanced dataset	62
Figure 5.7: SVM Classification report for a balanced dataset	63
Figure 5.8: SVM Classification report for imbalanced dataset	64
Figure 5.9: Confusion Matrix.....	65
Figure 5.10: Accuracy Comparison of Models	68
Figure 5.11: Login Page	69
Figure 5.12: Register Page	69
Figure 5.14: Project Page 1	70
Figure 5.15: Project Page 2	71
Figure 5.16: Import Libraries.....	72
Figure 5.17: Gender Visualization	74
Figure 5.18: Tenure-group Visualization	75
Figure 5.19: Contract Visualization	76
Figure 5.20: Payment-Method Visualization	77
Figure 5.21: Correlation using bar chart	79
Figure 5.22: Correlation using heatmap.....	80

List of tables

Table 3.1 Hardware requirements	46
Table 3.2 Software requirements.....	47
Table 5.1: Logistic Regression Results for balanced data	58
Table 5.2: Logistic Regression Results for imbalanced data	59
Table 5.3: Decision Tree Results for balanced data.....	60
Table 5.4: Decision Tree Results for imbalanced data.....	61
Table 5.5: Random Forest Results for a balanced data	62
Table 5.6: Random Forest Results for imbalanced data	63
Table 5.7: Support Vector Machine Result for balanced data.....	64
Table 5.8: Support Vector Machine Result for imbalanced data	65
Table 5.9: Results of Supervised Machine with balanced data.....	66
Table 5.10: Results of Supervised Machin with imbalanced data	66

CHAPTER ONE: INTRODUCTION

1.1 Introduction

The telecommunications sector has become one of the main industries in developed countries. Companies are working hard to survive in this competitive market depending on multiple strategies. Three main strategies have been proposed to generate more revenues acquire new customers, upsell the existing customers, and increase the retention period of customers. However, comparing these strategies taking the value of return on investment (RoI) of each into account has shown that the third strategy is the most profitable strategy, proves that retaining an existing customer cost much lower than acquiring a new one (Ahmad et al., 2019).

One of the biggest issues facing the telecommunications sector is customer attrition. The purpose of this study is to determine the variables that affect customer attrition, create a predictive model for customer attrition, and offer the best analysis of data visualization outcomes. The suggested approach for analyzing churn prediction consists of many stages: pre-processing the data, analysis, application of machine learning algorithms, assessment of the classifiers, and selection of the most effective one for prediction. The three main steps in the data preparation procedure were feature selection, data transformation, and data cleaning (Mohammad et al., 2019).

Customer churn detection is one of the most crucial study areas in the telecommunications industry that a firm must address in order to retain current clients. Churn refers to client attrition as a result of competitors' offers ending or maybe network problems. The lifetime

value of a client is significantly impacted by churn rate as it influences both the duration of service and the company's future income. The businesses are searching for a model that can forecast client attrition as it has a direct impact on industry revenue. Random Forest with XGBoost (Kavitha et al., 2020).

1.2 Background of the study

For the majority of businesses operating in low-cost switching sectors, customer turnover is their top priority. Of the industries that have this problem, the telecommunications sector is thought to be the worst. With the aim of lowering the rate of customer attrition, mobile service providers have deployed CRM (Customer Relationship Management). Still, there is a significant incidence of employee turnover in the telecom sector (Babu et al., 2014).

The development of the sector depends on improved client requirements perceptions and higher-quality models and services. Companies that want to be profitable and competitive must priorities reducing customer attrition, which has a significant influence on their operations. Thus, a great deal of study has been done by academics all over the world to comprehend the mechanisms of client turnover. In order to determine the numerous churn variables and their intricate linkages in the telecom churn literature now in existence (Mahajan et al., 2017).

Churn prediction is a key predictor of the long-term success or failure of a business. In this research, Ubiquitous techniques like Random Forest Classifiers and SVMs are compared with relatively newer architectures like XGBoost and Deep Neural networks to classify if a

customer will churn or not. The efficiency of these models is further explored by passing them through a grid search. From this experiment, it could be inferred the Random Forest model works best for this particular use case with a prediction accuracy of 90.96% on the testing data before grid search (Bhuse et al., 2020).

Every day, a sizable number of customers in the telecom sector create massive amounts of data. Churn, or the process of clients moving from one company to another within a predetermined period of time, in order to keep consumers' data from being churned, telecom management and analysts are attempting to figure out why subscribers are cancelling their contracts. This system gathers the reasons why consumers in the telecom business subscribe to leaves and utilizes classification algorithms to determine which customers have subscribed to leaves (Wagh et al., 2024).

1.3 Problem statement

The telecommunication companies in Mogadishu Somalia have challenges about keeping and the loyalty of their customers to predict either leaving or not. In today's competitive Telecommunication in Mogadishu Somalia, keeping customers happy and loyal is crucial. But sometimes, customers stop buying from a company or using its services, which is known as customer churn. This can happen for many reasons, like customers finding better deals elsewhere, not being happy with the product or service, or changes in their own needs. It's important for telecommunication companies in Mogadishu to understand when their customers might leave. This is where the idea of customer churn prediction comes in. It's

like having a crystal ball that helps businesses see which customers are likely to leave in the future. By knowing this in advance, businesses can take steps to keep these customers.

In Mogadishu Somalia, telecommunication companies face a significant challenge with customer churn, where customers frequently switch from one service provider to another. This issue not only affects the companies' revenue but also undermines their ability to maintain a stable customer base. The reasons behind this trend vary widely, including dissatisfaction with service quality, better offers from competitors, or changes in customer needs and expectations. Given the competitive nature of the telecommunication industry in Somalia, understanding and predicting when a customer might leave becomes crucial for these companies. However, the process of accurately identifying potential churners is complex, involving the analysis of large volumes of data covering service usage patterns, customer demographics, billing information, and feedback.

1.4 Motivation of the study

The main reason for doing this study is to help businesses get better at knowing when customers might leave. In a world where keeping customers is key to success, having a good way to predict customer churn is very important. If telecommunication companies in Mogadishu Somalia can understand and predict when customers are likely to leave, they can do things to make them happier and keep them for longer. This study is driven by the desire to make these predictions more accurate and helpful, so telecommunication businesses can save time and money, and keep their customers happy and loyal.

1.5 Objectives of the study

1.5.1 To develop a Better Prediction Model for telecommunication industries:

The first goal is to create a new, improved way to guess when customers will stop buying from a business. This model will try to be more accurate than the ones we have now, so businesses can have a clearer idea about what makes their customers leave.

1.5.2 To design a beautiful user interface:

The goal is to create an easy-to-use and attractive interface. This means making everything simple to find and nice to look at, ensuring users can navigate smoothly and enjoy using the system.

1.6 Research questions

- I. How to develop a Better Prediction Model for telecommunication industries?
- II. How to design a beautiful user interface?

1.7 significance of study

This research is very important for several reasons. **First**, it helps telecommunication companies in Mogadishu Somalia figure out customers leave. This is a big deal because when a business knows its customers are going away, it can take steps to fix those issues. For example, if customers are leaving because of high prices, a business might decide to offer discounts or special deals. Or, if customers don't like the service, the business can train its staff better. **Second**, this study is significant because it can save businesses a lot of money. It's usually more expensive to find new customers than to keep the ones already there. So, if a business can keep its customers, it doesn't have to spend as much money on marketing to

attract new ones. This means the business can use that money for other things, like improving products or services. Keeping customers happy and staying with the business can lead to more sales and profits in the long run. **Lastly**, the findings from this study can be useful for all types of telecommunication businesses in different industries. Every business wants to keep its customers and avoid churn. The insights from this study can provide new strategies and tools for businesses to better understand their customers. This can lead to better customer service, improved products, and a stronger relationship between the business and its customers. In the end, this is good for everyone - the customers get what they want, and the businesses succeed and grow.

1.8 Scope of the study

The Scope of this Study is specific on customer churn prediction focusing on telecommunication companies in Mogadishu Somalia aims to understand and forecast when customers are likely to leave one telecom provider for another.

1.8.1 Geographical scope of this project

The geographical scope of this project is in Mogadishu Somalia.

1.8.2 Time scope of this project

The project will be conducted between **January** and **August** 2024

1.9 Organization of the study;

Chapter One: introduces to the study. It contains the background of the study, statement of the problem, motivation of the study, the research objectives, the research questions, scope of the study, significance of the study and organization of the study.

Chapter Two: Literature Review- This chapter focuses on the previous literature about the customer churn prediction and researches related to this topic.

Chapter Three: is described the research methodology used in this study in terms of new model definition and also, we will discuss the requirements of the system.

Chapter Four: analyze and design of the study which means it analyses all requirement of implementing customer churn prediction.

Chapter Five: this chapter talks about the implementation and testing of the study which is to implement the system also, this chapter contains some practical codes.

Chapter Six: we discuss this chapter results of the developed customer churn prediction.

Chapter Seven: the last chapter of the book and talks about the conclusion of the book and future work.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

Customer churn is the term used to describe the loss of important clients to rival companies by service providers, particularly those in the telecommunications industry. The telecom sector has seen significant transformation in the past few years, including new services, new technologies, and market liberalization that increased competition. Customer attrition results in a significant loss of telecom services, making it a very critical issue (Huang et al., 2010).

In today's corporate world, acquiring and keeping clients are the most important priorities. Every business's market is expanding quickly, which is increasing the number of subscribers. As a result, businesses now understand how critical it is to hold onto their current clientele. It is now required of service providers to lower churn rates because, from a big perspective, neglect could result in a decline in profitability. Churn prediction assists in determining which consumers are most likely to depart from a business. The problem of an ever-rising churn rate is one that the telecom industry is dealing with. By using data mining tools, these telecom businesses can develop efficient strategies to lower their turnover rate (Hashmi et al., 2013).

It is especially crucial when calculating a business in industries where subscription-based income is generated, like banking, insurance, or telecommunications. Researchers have found that acquiring new clients in the cutthroat market of today might be up to ten times more expensive than keeping hold of current ones. It is an analysis technique used for things

like figuring out current consumer profiles, examining customer escapes, and projecting customer escapes (Çelik & Osmanoglu, 2019).

2.2 Customer Churn Prediction

Customer attrition is the process of a customer switching from one business service to another. Customer Churn Prediction is a tool used to predict potential customers who may quit the firm before they do. In order to attract potential churners and keep them around, this phase assists the business in developing the necessary retention policies, which lowers the company's financial loss (Umayaparvathi & Iyakutti, 2012).

Many sectors are concerned about customer turnover, but highly competitive industries are more vulnerable to it. Losing clients increases the need to find new ones and results in financial loss due to lower revenues (Xia & Jin, 2008).

In many types of organizations, maintaining existing clients is more economical than finding new ones, hence maintaining existing ones is vital. Forecasting a customer's likelihood of leaving the organization can be a challenging undertaking because of their unpredictable behavior. Because the data in the financial industry is sparser than in other domains, it is much more difficult to identify the top 10 customers who churn. Because of this, churn prediction research durations must be extended (Kaya et al., 2018).

It is often acknowledged that customer retention has a financial worth (Van den Poel & Lariviere, 2004):

- (1) By focusing more on the demands of their current clientele rather than pursuing new, potentially dangerous ones, businesses can achieve successful customer retention.
- (2) Long-term clients would be more advantageous and, if happy, would suggest new clients.
- (3) In a competitive market, long-term clients are typically less sensitive.
- (4) Because the bank is aware of them, serving long-term clients becomes less expensive.
- (5) Reduction in revenue due to customer attrition and increase in sales to draw in new business.

In every business, including banking, customer churn has grown to be a significant issue. To identify potential departing customers, banks have long attempted to monitor customer interactions. The basic goal of customer churn modelling is to identify high-risk consumers so that preventive measures can be taken (Oyeniya et al., 2015).

Businesses, including banks, insurance providers, and other service providers, are training their staff to be more customer- and service-oriented and are developing plans to guarantee that their clients stay with them (Bin-Nashwan & Hassan, 2017).

Retaining current customers and preventing customer attrition is the optimal fundamental marketing approach for the future (M.-K. Kim et al., 2004).

According to earlier studies, there are two categories of focused approaches—proactive and reactive—for handling client attrition. Reactively, the business holds off on terminating service until the client requests it. The business looks for clients who are likely to leave in an effort to be proactive. Next, the business offers incentives in an effort to keep those clients. Customers will leave businesses if churn projections are off, so churn should be accurate to avoid wasting money (Tsai & Lu, 2009).

2.3 History of Customer churn

Building from the investigation of electrical neuronal collisions, scientists employed the analogy of a ball and fire in the 1940s to illustrate human decision making. The 1950s saw the start of research on artificial intelligence. In order to evaluate a machine's capacity for human imitation, Alan Turing conducted the Turing Test around this time. To gauge a machine's capacity for human-to-human discourse, the Turing Test is primarily used to measure (Çelik & Osmanoglu, 2019).

Customer relationship management is a key area of business analysis in any kind of organization. It addresses keeping current clients as well as finding, reaching out to, and luring in new ones. There are two components to CRM: an operational component and a technical component. Customer Analytics, or the technological side of the CRM, is (Senanayake et al., 2015).

There are two categories for customer analytics:

- (1) Descriptive analytics: This is the method used to identify customers.
- (2) Predictive Analytics: Here, customer retention was the main goal.

The primary goal of predictive analytics' customer churn research is to keep customers (Senanayake et al., 2015).

The researchers (Senanayake et al., 2015) found that the usual method of identifying the customer in the absence of machine learning was looking at the data of the customers who had already left and determining customer attrition from the current clientele based on behavior and observation.

2.4 Data Exploration and Pre-processing

To learn more about the data and the business challenge, data exploration is necessary. For the Data mining model, the CRISP-DM methodology is widely acknowledged (Zhang et al., 2003).

Its primary use is for carrying out the data mining process, which has six stages in its life cycle, as seen in the picture below.

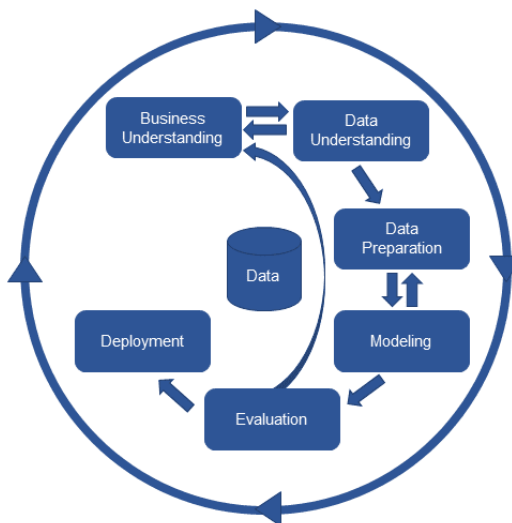


Figure 2.1: Complete CRISP-DM

(Source: (Zhang et al., 2003)).

2.4.1 Class Imbalance

According to studies conducted by (Xia & Jin, 2008) In data mining, class imbalance has emerged as a prevalent issue in datasets. In the field of customer churn prediction, this is a typical issue. Less examples are designated as the most significant class—the churned class—while the majority of cases in such an issue are classified as the not churned class.

One solution to the issue of class disparity was to apply the under-, random-, and over-sampling techniques recommended by (Bin et al., 2007). The percentage of nonchurn and churn was established in their study on Customer Churn Prediction on Personal Handyphone System Service. Three separate sample approaches were used to train the models, and random sampling produced better results.

The performance was shown in the figure below.

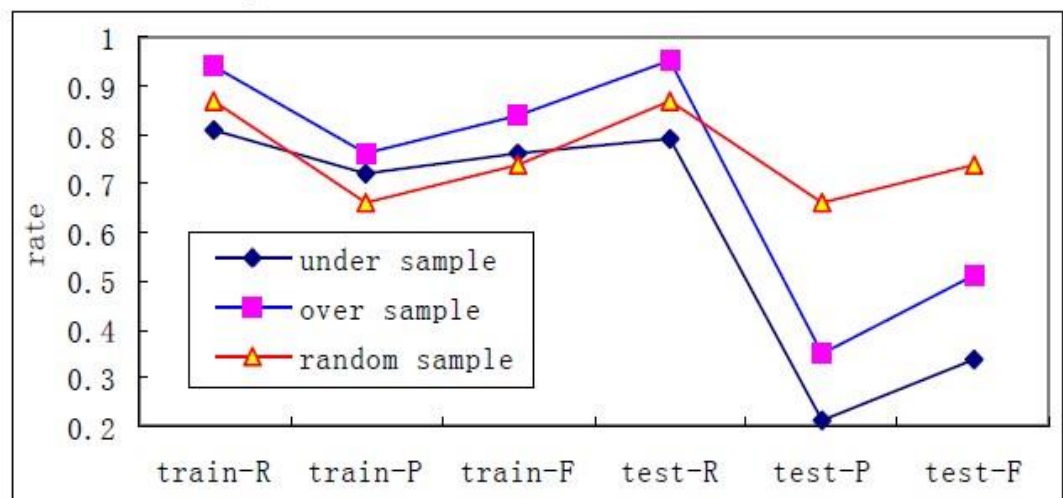


Figure 2.2: performance

Source: (Bin et al., 2007)

It was determined from the preceding figure that random sampling produced the best outcomes for a Decision Tree model. It has been noted in earlier studies by (Maheshwari et al., 2017) A thorough explanation of several methods for addressing class imbalance was covered.

The three different strategies for addressing class imbalance were the cost-sensitive strategy, the algorithm-level strategy, and the data-level strategy. The following techniques were used in the data level approach: under sampling, oversampling, and hybrid sampling. Under sampling results in the loss of potentially valuable data, whereas oversampling, when combined with massive amounts of data, causes overfitting and lengthens the learning process. Methodology at the algorithmic level The bagging and boosting techniques are used to address class imbalance. Decision tree (C4.5) and Random Forest algorithms were utilized for the bagging approach, while AdaBoost and SMOTEBOOST algorithms were used for the boosting method. The data-driven and algorithm-level approaches are both included in the cost-sensitive strategy (Bin et al., 2007).

In investigations executed by (Kaya et al., 2018) When it came to predicting customer attrition, it was found that the SMOTE technique produced the greatest results for SVM. Better results were obtained by SMOTE since it avoided the overfitting issue and generated minority classes by interpolating rather than replicating.

Algorithms can also be used to overcome imbalance issues. One kind of learning that took misclassification costs into account was called cost-sensitive learning. A student who was cost-sensitive put more value on false negative results than on false positive results. It was not a particularly practical strategy, though, as the cost information depended on a lot of other variables (Ganganwar, 2012).

One such algorithm-based strategy was class learning, which used the separate-and-conquer strategy and only modelled the classifier on the minority class. This method worked well for extremely imbalanced data sets made up of noisy, high-dimensional feature spaces (Kotsiantis et al., 2006).

2.4.2 Feature Selection

Finding the fields that are most conducive to prediction is a crucial step in the feature selection process (Hadden et al., 2007). When predicting client attrition, this stage is crucial. Feature selection is a crucial and widely applied dimensionality reduction technique in data mining, which involves choosing a subset of the original features.

In one of the studies carried out by (Khan et al., 2015) The degree to which a single feature can reliably distinguish between individuals who have churned or not was shown by a separate t-test for each feature. For feature selection, a tree-based approach was employed. A list of correlated predictors could be obtained with the use of this method.

Based on Label Information and Search Strategy, the feature selection was divided into two groups. This diagram will explain the split in depth.

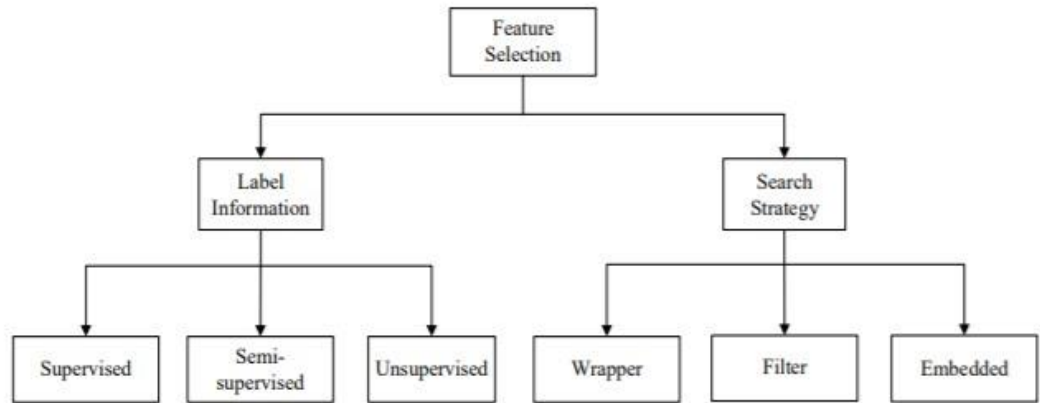


Figure 2.3: Feature Selection Category

(Source: (Miao & Niu, 2016))

Supervised, Unsupervised, and Semi-Supervised feature selection algorithms can be developed based on the labelled, unlabelled, or partially labelled training data. The connection between the features was used by supervised feature selection to determine the importance of the features. Unsupervised feature selection assessed feature importance by taking advantage of data separability and variation. An approach for semi-supervised feature selection enhanced the feature selection of unlabeled data by utilizing both labelled and semi-labelled data. Three feature selection categories—filter, wrapper, and embedding models are determined by the search technique (Miao & Niu, 2016).

As per the findings of researchers (Cai et al., 2018) Using the correlation between the feature and the class label as its guiding concept, supervised feature selection

was used for the classification challenge. The threshold was compared to the correlation between the features to evaluate whether or not a feature was redundant. This approach optimized the accuracy of the classifier by effective feature selection.

2.5 Machine Learning

Data-driven methods and machine learning are becoming increasingly significant in many fields. For instance, intelligent spam classifiers safeguard our emails by gaining knowledge from copious quantities of spam data as well as user reviews. Fraud detection systems shield banks from malevolent intruders, whereas anomaly event detection systems assist experimental physicists in identifying events that lead to new physics. Ad systems are designed to match the appropriate advertisements with the appropriate content (Çelik & Osmanoglu, 2019).

Machine learning is mostly applied to difficult tasks or problems involving large amounts of data. For more complicated data, it is a suitable solution since it produces faster, more accurate findings. It aids a company in recognizing any unidentified hazards or lucrative opportunities (Sayed et al., 2018).

Two primary learning approaches are used in machine learning:

- 1) Supervised Machine Learning

2) Unsupervised Machine Learning

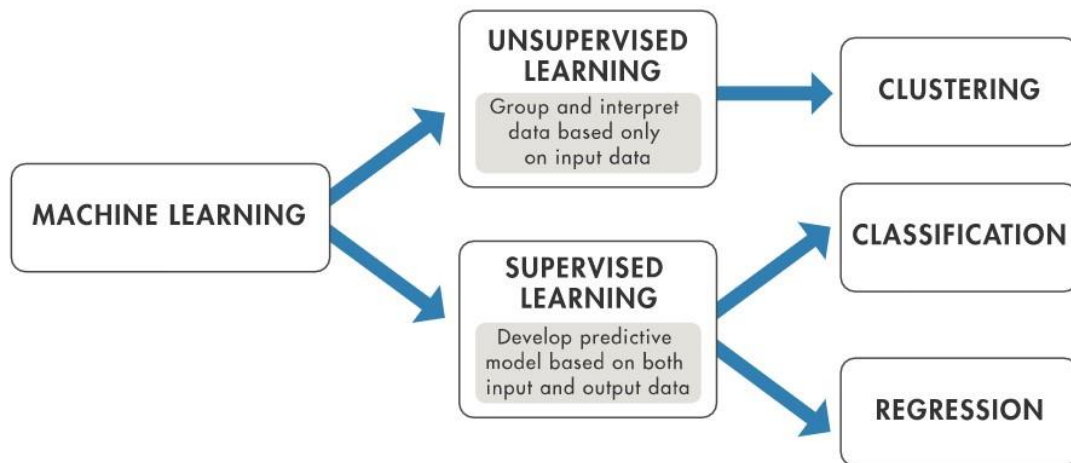


Figure 2.4: Machine Learning Techniques

2.5.1 Supervised Machine Learning

The computational job of discovering correlations between variables in a training dataset and using that knowledge to build a predictive model that can infer annotations for incoming data is known as supervised machine learning (Fabris et al., 2017). An algorithm is used in supervised machine learning to learn the mapping from the input to the output given two variables: X and Y.

$$Y = f(X)$$

In order for the model to forecast the output variable (Y) given fresh input data (X), it must be able to approximate the mapping function as well as possible (Fabris et al., 2017).

When examples are provided with labels, the learning is referred to as supervised learning. The characteristics may be binary, categorical, or continuous (Kotsiantis et al., 2006).

The two categories of supervised learning tasks are regression and classification.

- 1) Classification: Classification challenges arise when the output variable is categorical, e.g., "red" or "blue" and "yes" or "no."
- 2) Regression - Regression problems are those in which the output variable has a real value (Nasteski, 2017).

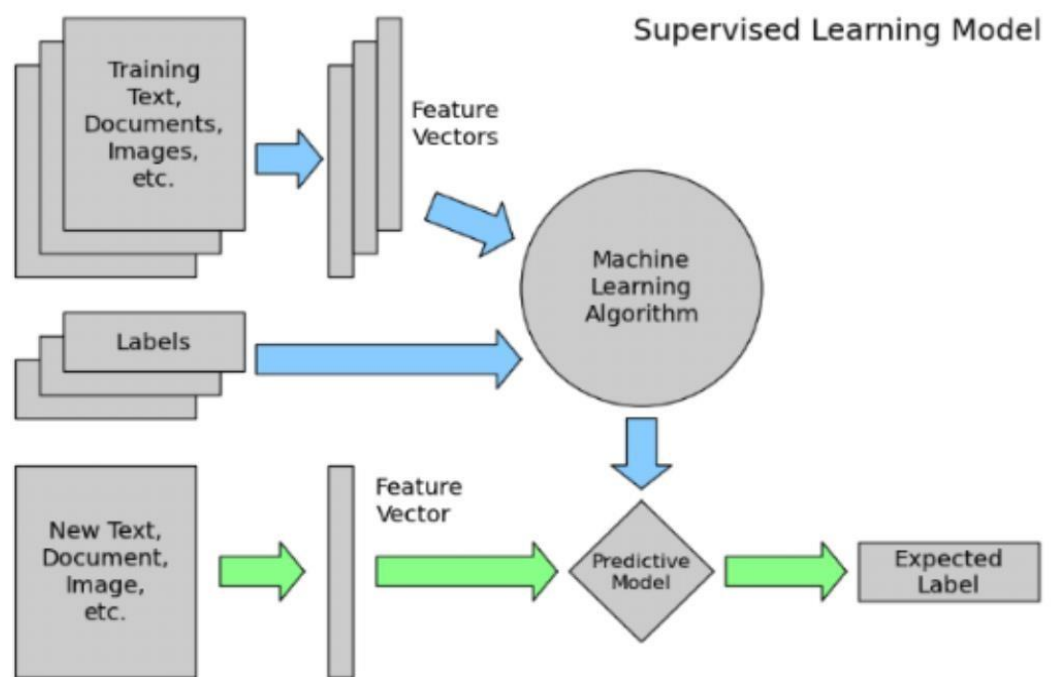


Figure 2.5: Supervised Machine Learning Model

(Source:(Nasteski, 2017))

2.6 Machine Learning Techniques

Similar customer churn prediction issues have already been solved using a number of machine learning techniques.

2.6.1 Logistic Regression

A popular statistical model for analysing customer attrition was logistic regression, which has shown to be an effective method. The logistic regression formula shown in figure 6 below denotes a situation where the outcome p_i is predicted by the independent variables x_i and the probability p_i (Nie et al., 2011).

$$p_i = \frac{e^{x_i\beta}}{1 + e^{x_i\beta'}}$$

Figure 2.6: Logistic Regression Formula

(Source: (Nie et al., 2011))

In an investigation on credit card churn prediction inside the Chinese banking sector, Decision Tree and Logistic Regression models were constructed. It was found that Decision Trees performed worse than Logistic Regression (Nie et al., 2011). There were 135 variables total. However, in this study, just a subset of the variables were chosen, and models were constructed using the correlation between the variables. The results showed that the Logistic Regression model outperformed the Decision Tree technique.

Using SAS 9.2, researchers have applied the Cox regression models and logistic regression technique to forecast customer turnover using binary and ordinal logistic regression models (Ali & Arıtürk, 2014).

The Logistic Regression model outperformed the Decision Tree model in a different study comparing models used to predict Customer Churn using Telecom datasets. Maximum likelihood estimate is used in logistic regression to convert the dependent variable into a logistic variable. A statistical survival analysis tool was offered by the suggested system to forecast client attrition. The confusion matrix served as a tool for assessment (Dalvi et al., 2016).

2.6.2 Random Fores

A random forest is a collection of tree predictors where each tree in the forest is dependent upon the values of a random vector that is separately sampled and has the same distribution for all of the trees in the forest. As a forest's tree count increases, the generalization error converges to a limit. The strength of each individual tree in the forest and their correlation with one another determine the generalization error of a forest of tree classifiers (Breiman, 2001).

The random forest classification technique was employed by the researchers in one of their earlier studies on financial client attrition (Kaya et al., 2018).

2.6.3 Support Vector Machine

Two approaches are the main ones that several academics have used to estimate client attrition. First, there was the conventional approach to classification, which used supervised learning mostly for quantitative data. For large-scale, high-dimensionality, nonlinearity, and time-series data, there was an artificial intelligence method (Xia & Jin, 2008). The customer churn prediction problem in the telecommunications business has been studied before, and the researcher (Xia & Jin, 2008), have employed the SVM model since it is capable of resolving high dimension, local minimization, and nonlinearity issues. The condition and data structure affected the model's forecast.

2.6.4 Neural Network

The investigator (Bilal Zorić, 2016) have employed the neural network model in the Alyuda NeuroIntelligence software package to study customer churn prediction in the banking sector. This is because neural networks have proven effective in solving many types of problems related to pattern recognition, image processing, optimization, and other areas.

An additional team of researchers (Huang et al., 2010) have suggested contrasting the cutting-edge modelling technique, SVM, with the well-liked modelling techniques, Multilayer Perceptron Neural Networks and Decision Trees (Huang et al., 2010) for predicting client attrition in the telecom sector. Decision Trees were not as efficient as MLP and SVM.

2.7 Model Evaluation

The researchers identified two categories of evaluation strategies: filter and wrapper. In the wrapper evaluation approach, a learning algorithm was used to evaluate a subset of features, but in the filter evaluation method, a subset of characteristics outside of the classification design were evaluated (Hadden et al., 2007).

The customer churn prediction problem is a classification problem, and the confusion matrix was used to determine the precision, recall, accuracy, and F-measure in order to assess how well the supervised machine learning models performed (Vafeiadis et al., 2015).

Many studies have also utilized the AUC (Area Under Curve) to evaluate the model in addition to the confusion matrix. The receiver operating curve's (ROC) area under the curve is known as the AUC. The true positive rate against false positive rate is plotted in a ROC chart. TDL (top-decile lift), an assessment metric that concentrates on the customer most likely to churn, was another (Ali & Arıtürk, 2014).

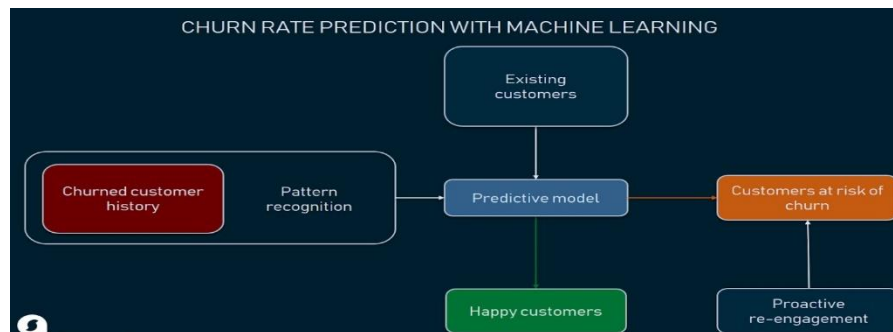


Figure 2.7: Churn Rate Prediction using Machine Learning

2.8 Customer Churn Prediction using Machine Learning

A series of stages are involved in predicting customer turnover using machine learning models. After the data was gathered, a machine learning model was constructed by pre processing and transforming the chosen data into an appropriate format. Testing was done after modelling, and the model was then ultimately deployed (S. Kim et al., 2005). For the purpose of the customer churn analysis, the machine learning examined the data and found the underlying data trends (S. Kim et al., 2005). The estimate of client attrition using machine learning was more accurate than the conventional method.

The customer churn analysis used a number of features as variables. Customer factors of recency, frequency, and monetary value (RFM) as well as demographic characteristics like age, culture, and geography made up the different categories of variables (Senanayake et al., 2015).

2.9 Approaches to solve the problem

Investigators (Xia & Jin, 2008) have used the SVM machine learning technique for structural risk minimization to forecast customer attrition using a customer data set from the telecom sector. They have analyzed the SVM model findings using naïve Bayesian classifiers, logistic regression, artificial neural networks, and decision trees. It was discovered that the SVM performed better in the experiment in terms of accuracy rate, hit rate, covering rate, and lift coefficient. The study employed two datasets, and MATLAB 6.5 was utilized to select the kernel function for the SVM model.

Another investigation on customer data from European financial banks was carried out by (Van den Poel & Lariviere, 2004) Assessing client attrition with the Cox proportional hazard approach. The churn occurrence was the main concern. This study made use of the SAS enterprise miner. To conduct the research, they combined a variety of predictor categories into a single, all-inclusive proportional hazard model. Through the examination of this bank customer dataset, two crucial phases of customer attrition were discerned: the initial years following client acquisition, and a subsequent period occurring approximately 20 years later.

The scientists constructed hybrid neural networks and evaluated their performance against the standard ANN model (Tsai & Lu, 2009). Based on data from US telecom companies, client attrition was anticipated. In order to enhance the effectiveness of the individual clustering or classification strategies, the researchers in this study constructed two hybrid models in addition to one baseline ANN model by merging the two methodologies. It was divided into two learning stages: the first was for pre-processing the data, and the second was for predicting the final output. Two hybrid models were constructed: SOM (Self Organizing Maps) +ANN and ANN+ANN (Artificial Neural Network).

In one of the studies that scholars have written on client churn in the banking sector (Kaya et al., 2018) The influence on spatiotemporal aspects has been highlighted more. For their investigation, they used the Random Forest classification model, which they trained using 500 trees and a maximum of two features per tree. For evaluation, stratified 8-fold cross-

validation was used. In financial churn decision prediction, choice and spatiotemporal factors were found to be more effective than demographic features in this study.

Investigators (Oyeniya et al., 2015) Utilising the WEKA method for knowledge analysis, they were able to predict the issue of customer attrition on a dataset from a Nigerian bank. JRip algorithm rule creation was conducted after the K-means clustering algorithm was utilised for the clustering phase.

Researchers predicted customer churn on the Personal Handy Phone System Service (Bin et al., 2007). In order to create a precise and useful customer churn model, they developed a decision tree and ran three tests. In this study, 180 days' worth of data were randomly selected and used to predict churn. Sub-periods for the training data sets were altered in the first experiment, the misclassification cost in the churn model was altered in the second experiment, and sample techniques in the training data sets were altered in the third trial.

Comparative research on the forecasting of customer attrition was carried out by (Vafeiadis et al., 2015) on the telecom data set. Multi-layer perceptron, Decision Trees, Support Vector Machines, Naïve Bayes, and Logistic Regression were compared for performance. Cross-validation was used in the construction and assessment of every model. Using Monte Carlo simulations, SVM fared better than the other models, with an accuracy of 97% and an F-measure of 84%.

In a prior study on customer churn prediction, the researchers employed soft computing techniques like fuzzy logic, neural networks, and genetic algorithms in addition to the conventional supervised machine learning algorithms of decision trees and regression analysis (Hadden et al., 2007).

In one of the studies on comparisons conducted by (Xie et al., 2009) Based on precision and recall, it has been found that balanced Random Forest performs better than ANN, SVM, and DT classifiers.

2.10 Related works

As a result of their inability to forecast which consumers will leave on schedule, many businesses often deal with the problem of losing clients. This research's primary goal is to give telecom providers a quick and practical method for identifying potential at-risk clients. In order to create our churn prediction model, we applied both logistic regression and logit boost (Jain et al., 2020).

The study's findings indicate that both approaches are extremely similar in every measurement. The outcomes from both procedures were identical and excellent. While the logistic regression model projected inaccurately that 57 customers had churned, it properly classified 2841 consumers as not likely to do so. On the other hand, the model mis predicted that 48 customers had not churned while accurately predicting that 435 customers had. In one case, the LogitBoost model properly identified 2839 consumers as not likely to churn while wrongly predicting that 50 customers would. In another instance, the LogitBoost

model correctly identified 444 customers as having churned while incorrectly predicting that 39 customers would not. As a result, accuracy was 85.2385% and 85.1785 (Jain et al., 2020).

Data mining can be used in this business to evaluate the degree of consumer loyalty. There are various study models used in data mining. In classification, one of the most widely used techniques is the K-Nearest Neighbour algorithm. The data applied in this study came from UCI's machine learning repository's German Credit Datasets (Imron & Prasetyo, 2020).

The goal of this study is to determine how Particle Swarm Optimisation determines the most appropriate K value parameters and how Z-Score normalizes the data to improve the performance of the K-Nearest Neighbor algorithm during classification. The produced accuracy of the categorization was checked using a confusion matrix. According to this study's findings, using Particle Swarm Optimization in conjunction with the K Nearest Neighbor algorithm and Z-score normalization can increase accuracy by up to 14%. After implementing Particle Swarm Optimization and Z-Score normalization, the accuracy increased from 68.5% to 82.5% (Imron & Prasetyo, 2020).

The volume of raw data kept in banking systems is enormous and keeps growing due to improved data availability, low-cost storage, and processing capacity. We shall concentrate on customer churn in this study. Neural network techniques are the most widely utilized methods for predicting client attrition. Our goal is to create a model using historical customer data in order to anticipate churn and stop customer attrition (Bilal Zorić, 2016).

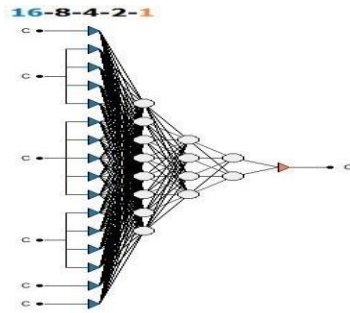


Figure 2.8. Alyuda – Network Topology.

The software programmer Alyuda NeuroIntelligence's neural network approach was utilized in the study to identify customer attrition. The software runs through each of the aforementioned stages after choosing the database. Three categories of characteristics are defined during the data analysis phase: those that we will utilize (such as name and surname), those that we will reject, and the target attributes that we wish to calculate. Data is separated into three sets by Alyuda NeuroIntelligence: training, validation, and testing sets. If data is designated as categorical, the programme adds a few columns during the preprocessing stage. We choose a number of hidden layers during the network design stage. The best topology is provided by the programmer, which we can modify. That is a neural network in our instance with three hidden layers, eight, four, and two neurons (Bilal Zorić, 2016).

After the phase of design, there is a training phase when many parameters can be defined, as shown in Figure 2.9.

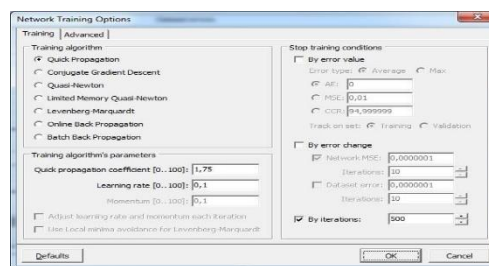


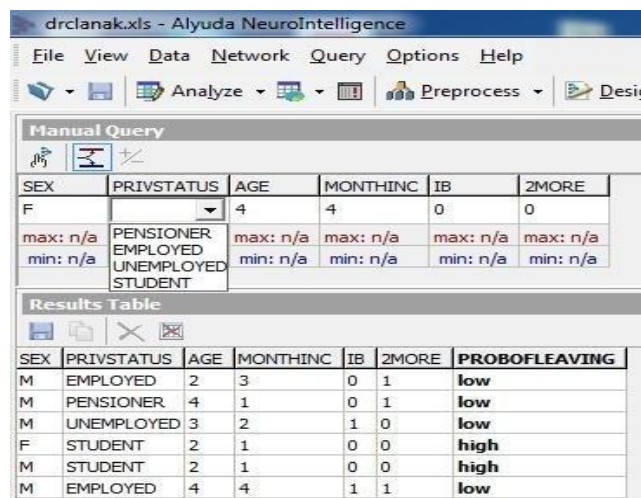
Figure 2.9. Alyuda – Network Training Op

Following network training, we obtain the outcomes displayed in Figure 2.10.

Parameters		
	Training	Validation
CCR, %:	95,984252	93,959732
Network error:	0,123164	0
Error improvement:	0,000009	
Iteration:	501	
Training speed, iter/sec:	62,625026	
Architecture:	[16-8-4-2-1]	
Training algorithm:	Quick Propagation	
Training stop reason:	All iterations done	

Figure 2.10. Alyuda – Network Training R 1

Finally, we obtained a model that allows us to enter certain parameters and see how likely it is that a client will leave the bank.



The screenshot shows the 'Manual Query' window in the Alyuda NeuroIntelligence application. The window has a menu bar (File, View, Data, Network, Query, Options, Help) and a toolbar with icons for Analyze, Preprocess, and Design. Below the toolbar is a 'Manual Query' section with a table for inputting data. The table has columns: SEX, PRIVSTATUS, AGE, MONTHINC, IB, and 2MORE. The input values are: SEX = F, PRIVSTATUS = PENSIONER (selected from a dropdown), AGE = 4, MONTHINC = 4, IB = 0, and 2MORE = 0. Below the input table is a 'Results Table' section. The Results Table has columns: SEX, PRIVSTATUS, AGE, MONTHINC, IB, 2MORE, and PROBOFLEAVING. The Results Table contains six rows of data, each representing a different client profile and its predicted probability of leaving the bank.

SEX	PRIVSTATUS	AGE	MONTHINC	IB	2MORE	PROBOFLEAVING
M	EMPLOYED	2	3	0	1	low
M	PENSIONER	4	1	0	1	low
M	UNEMPLOYED	3	2	1	0	low
F	STUDENT	2	1	0	0	high
M	STUDENT	2	1	0	0	high
M	EMPLOYED	4	4	1	1	low

Figure 2.11. Alyuda – Query.

CHAPTER THREE: RESEARCH METHODOLOGY

3.0 Introduction

In this chapter, we'll explain how we're going to predict when customers might leave their telecom service in Somalia. The main goal is to outline the methods and steps we'll use to create a system that can forecast customer churn accurately. We'll also talk about the software we need to make this system work and what the system will be like. We'll be using machine learning techniques to analyze data and figure out if a customer might stop using their telecom service.

3.1 System Description

Our churn prediction system marries data ingestion with sophisticated feature engineering, extracting key insights from customer data for accurate churn forecasts. We deploy machine learning models, fine-tuned for precision, including Neural Networks and Gradient Boosting Machines, to predict customer departure risks. The system provides real-time assessments and actionable insights, powered by Python's comprehensive data science libraries like Scikit-learn and TensorFlow. Through RESTful APIs, it integrates seamlessly into telecom infrastructures, offering scalable solutions on cloud platforms for enhanced performance and reliability. This setup not only predicts churn but also equips telecom providers with strategic insights to improve customer retention effectively.

3.2 System Architecture

The three primary phases of customer churn prediction in the system architecture for Mogadishu, Somalia's telecommunications sector is featuring selection, data cleaning, and model training. This is depicted in the graphic below. It studies client data and finds trends

using machine learning algorithms. To ensure accuracy and reliability, this process starts with data cleaning. The most important features for churn prediction are then selected through feature selection. To create prediction models, historical data is used for model training. Based on new information, the trained models are then applied to forecast client attrition. Businesses can proactively identify and retain customers who are at risk of leaving by using this design.

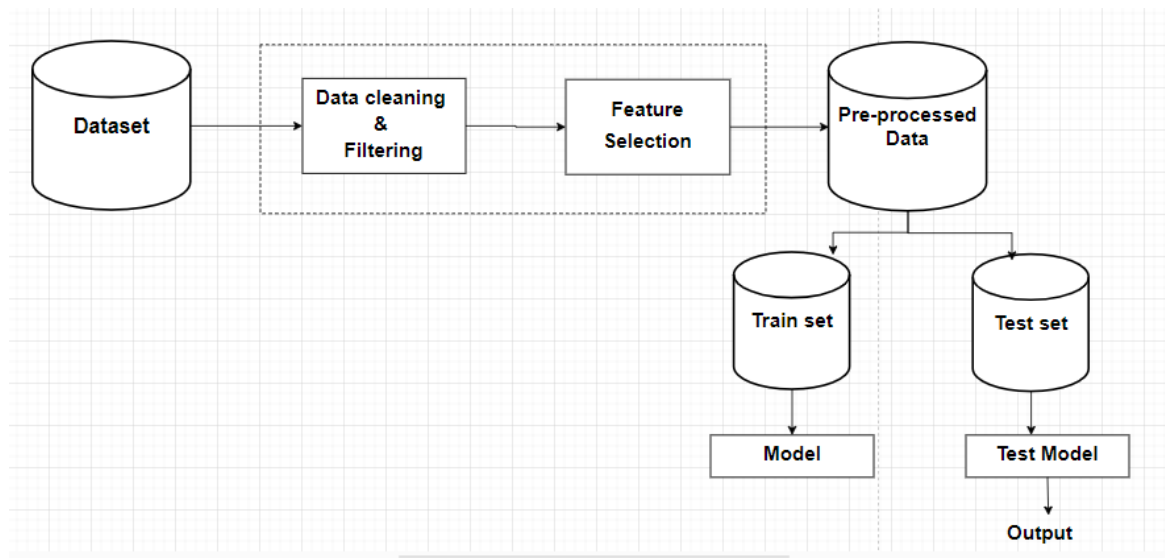


Figure 3.1: System Architecture

3.3 System Features

These are the system features of customer churn prediction for telecommunication

Mogadishu, Somalia:

3.3.1 Customer Churn Prediction:

1) **Machine Learning Integration:** Utilizes machine learning algorithms like Logistic Regression and Random Forest to predict the likelihood of customers discontinuing service.

2) **Dynamic Prediction:** Users can input customer data through a web form, and the system will predict churn status in real time.

3.3.2 Data Visualization:

1) **Interactive Visuals:** Includes a carousel of visualizations such as confusion matrices, correlation charts, and churn by various categories (e.g., monthly charges, service types).

3.3.3 User Management:

1) **Session Management:** Supports user sessions where users can log in and out, maintaining session state with secure handling.

2) **User Authentication:** Implements a login system to authenticate users, ensuring that only authorized personnel can access sensitive predictive functionalities and data. Our system learns from previous cases of customers leaving.

3.3.4 Statistical Dashboard:

1) **Real-time Statistics:** Displays real-time statistics about the data processed, including total entries, number of churned and retained customers.

2) **Data Export:** Allows users to export data tables for offline analysis or reporting purposes.

3.3.5 Security Features:

1) **Data Validation:** Implements checks on the server side to validate input data for predictions to prevent SQL injection and ensure data integrity.

2) **Secure Session Handling:** Uses Flask's built-in session management which is secured with a secret key to prevent tampering.

3.4 System Methodology

In the telecom sector, predicting customer loss using a methodical strategy is essential. The four main steps of this process are Model Evaluation, Model Selection, Model Acquisition, and Model Implementation. Telecom firms can successfully detect and handle churn by carefully working through these stages, which will promote customer loyalty and long-term sustainability.

3.4.1 Dataset

The dataset used in this study consists of a total of 7,043 rows and 21 columns.

Each row in the dataset represents a unique customer and provides information about their characteristics, such as gender, age, partnership status, and dependents. It also includes details about the services they have subscribed to, including phone service, internet service, and various add-ons like online security, online backup, device protection, tech support, streaming TV, and streaming movies. The dataset also captures important contract-related information, such as the type of contract a customer has (month-to-month, one year, or two years), their paperless billing preference, and the payment method they use. Additionally, it includes financial information such as monthly charges and total charges incurred by the customers. To access the dataset, it can be obtained from the platform called Kaggle, which is a popular online community for data science and machine learning.

3.4.2 Data preparation

Data preparation, also known as data preprocessing, is the process of cleaning, organizing, and transforming raw data into a format suitable for analysis or training machine learning models. It is a crucial step in the data analysis and machine learning

pipeline, as the quality of the input data significantly impacts the performance and effectiveness of the models.

The main goals of our data preparation include:

1) Handling Missing Data: Identifying and addressing missing values in the dataset.

This can involve removing rows with missing values, imputing missing values based on statistical measures, or using advanced techniques to predict missing values.

2) Encoding Categorical Variables: Converting categorical variables (non-numeric) into a numerical format that machine learning models can understand. Common methods include one-hot encoding, label encoding, or using embeddings for more complex categorical data.

3) Data Splitting: Dividing the dataset into training, validation, and testing sets. The training set is used to train the model, the validation set is used for fine-tuning and hyperparameter tuning, and the testing set is used to evaluate the model's performance on unseen data.

4) Handling Data Duplicates: Identifying and handling duplicate records in the dataset to avoid biases in model training and evaluation.

3.4.3 Model Selection

In customer churn prediction for telecommunications, we utilize a range of machine learning models, namely Decision Tree Classifier, Random Forest Classifier, Support Vector Machine (SVM), and Logistic Regression. These models are employed to forecast and identify customers who are likely to churn, enabling proactive measures to retain their loyalty.

3.4.3.1 Decision Tree Classifier:

A decision tree is a flowchart-like structure where internal nodes represent features or attributes, branches represent decisions, and leaf nodes represent outcomes. In this case, a decision tree classifier uses historical customer data to create a tree-like model that can predict whether a customer is likely to churn or not. It makes decisions based on a series of questions about the customer's characteristics, such as usage patterns, demographics, or service preferences.

3.4.3.2 Random Forest Classifier:

A random forest classifier is an ensemble learning method that combines multiple decision trees to make predictions. It generates a collection of decision trees, each trained on a random subset of the data and a random subset of features. When predicting churn, each decision tree in the random forest independently predicts the outcome, and the final prediction is determined by majority voting or averaging the individual tree predictions.

3.4.3.3 Support Vector Machine (SVM):

SVM is a supervised machine learning algorithm used for classification tasks. It works by finding an optimal hyperplane that separates different classes in a high-dimensional feature space. In the context of churn prediction, SVM can be trained on historical customer data, with features such as call duration, customer tenure, or usage patterns. SVM aims to find a decision boundary that maximizes the margin between churned and non-churned customers, allowing it to make predictions for new customers.

3.4.3.4 Logistic Regression:

Logistic regression is a statistical model used for binary classification problems, such as churn prediction. It models the relationship between the dependent variable (churn) and one or more independent variables (customer attributes). Logistic regression estimates the probabilities of churn by applying the logistic function to a linear combination of the input features. The resulting probabilities can then be used to classify customers as likely to churn or not, based on a predefined threshold.

3.4.4 Model Evaluation

When we're trying to predict if customers will leave a telecom company, we use four different methods. First, we look at Decision Trees to see how well they spot potential leavers and understand their reasons. Then, we check Random Forests to see their overall accuracy and how they handle complex info. Next, Support Vector Machines (SVM) help us see if they can draw a line between staying and leaving customers accurately. We use terms like accuracy, precision, and recall to measure this. Lastly, Logistic Regression helps us estimate how likely someone is to leave and how easy it is to understand the reasons. By doing this, we find the best way to predict customer churn in telecom.

3.4.5 Implementation

The process of developing, deploying, and transforming the system on for end users is known as implementation. In order to determine if a customer is churning or not, it involves integrating the learned machine learning models into a useful application that customers are able to interact with.

3.5 System development environment

3.5.1 Front End

As of 2020, there are a lot of libraries, frameworks, and tools available for front-end web development. In general, this thesis finds that the introduction of all third-party front-end web development tools may be caused by three main factors. First of all, the design of HTML, CSS, and JavaScript is completely faulty. Secondly, the rate of advancement of HTML, CSS, and JavaScript is not able to match the lightning-fast rate of the web's development. Thirdly, compared to desktop or mobile platforms, there is no authority figure in front-end web development who can establish rules or provide an official integrated development environment or developer tools (Dinh & Wang, 2020).

The user interface that allows people to interact with the system of suggestions is called the frontend. Web technologies including HTML, CSS, Bootstrap, and JavaScript are used in its development, giving it a responsive and user-friendly design that is accessible with a number of devices.

3.5.2 Back End

It is essential that we talk about what a frontend is and how they relate to it in order to better understand the function of a backend. The frontend, sometimes referred to as the client-side, is what a user is able to view and communicate with on a web page or a mobile application. Everything that takes place on the server and database side of a web or mobile application is represented by the backend, often known as server-side. The user doesn't deal with the backend directly and is informed of it. On the other hand, the

backend controls what transpires when a user interacts with the application frontend. Using an API, the server-side programs creates, retrieves, and modifies data from the database (Allain, 2020).

The server side, which was developed in Python using Flask, manages model training and data processing. Flask routes manage front-end request management, run machine learning models, and provide users customer is churn or not churn.

3.5.3 Python

Python is becoming a very popular programming language among software developers and data scientists (Robinson, 2017).

Python is used in a far wider range of applications than R, which is primarily meant for statistical data analysis. Examples of these applications include scientific computation, database access, desktop GUIs, Internet and website development, and software and game development. There are four primary machine learning topics covered by the Scikit-learn package. They are model evaluation and selection, supervised learning, unsupervised learning, and data transformation (Hao & Ho, 2019).

Flask is a Python microframework released under the BSD agreement. Flask is a very simple yet highly extendable framework; it is not any less functional for being a microframework. This makes it easier for developers to write applications or plugins by

allowing them to select the settings they want. Originally developed in 2010 by a group of open-source developers called Pocoo, Flask is currently being developed and maintained by The Pallets Project, which powers all of Flask's components. A vibrant and supportive developer community, which includes a mailing list and an IRC channel, is available for Flask (Relan, 2019).

3.5.4 Visual studio (vs code)

A powerful IDE used to develop cloud apps and Web apps is called Visual Studio Code. Although the application is lightweight and functions similarly to Visual Studio, it has powerful editing and building tools housed in a sleek UI. dependable programming tools and a lightweight application Visual Studio Code provides you with Git control, which is one of its amazing features. It makes a variety of software testing, building, packaging, and even deployment easier. It works with several programming languages, including C#, C++, Clojure, HTML, JSON, Java, Lua, PHP, Perl, Python, SQL, Visual Basic, XML, and others. Projects can be exported as text files. Additionally, the application supports ASP.NET and Node.js development (Code, 2019).

3.5.5 Jupyter Notebook

Data science, machine learning, and computer science education are just a few of the software engineering fields where Jupyter notebooks have become more and more popular in recent years. Although notebooks are popular because of the wide range of capabilities for presenting and displaying data, new research indicates that they also have several disadvantages in common, such as a high number of code clones and poor reproducibility. In this study, we compare Python code written in conventional Python scripts versus Jupyter Notebooks (Grotov et al., 2022).

3.6 Hardware requirement

The hardware device requirements for the implementation of Customer churn prediction telecommunications in Mogadishu, Somalia using machine learning are:

SYSTEM TYPE		DISPLAY	PROCESSOR	RAM
64-bit System	Operating	12 inch and more	Intel(R) Core (TM) i5-10210U CPU @ 2.10GHz - 2.60 GHz	At least 4GB and more

Table 3.1 Hardware requirements

3.7 Software Requirements

The following software needs to be implemented in order to use machine learning for Customer churn prediction telecommunications in Mogadishu, Somalia:

➤ Operating system:

Windows 10 and above is used since it is more user-friendly, stable, and offers a greater number of functions than previous versions of Windows.

➤ Python:

The majority of machine learning frameworks have been created with Python, making it the most popular programming language for machine learning.

➤ Development Tools:

For machine learning development, IDEs like PyCharm, Jupyter Notebook, and Visual Studio Code are often used.

➤ Libraries:

For data manipulation, analysis, and visualizations, libraries like NumPy, Pandas, and Matplotlib are regularly used.

Software requirement	
Operating system	windows 10 and more
Browser	Internet Explorer, Chrome and more
Programming languages	Python
Machine learning libraries	Scikit-learn, imblearn, pickle
Editors	Jupyter Notebook, Visual code
Front end	Html, CSS, bootstrap
Back end	Flask
Database	Sqllite

Table 3.2 Software requirements

CHAPTER FOUR: ANALYSIS AND DESIGN

4.1. Introduction

This chapter provides a system for analyzing and designing a system that separates Customer churn prediction for telecommunication Mogadishu, Somalia. In this section, we will discuss the current system, the problem it solves, how it will perform, and why we're using it. The requirements of the system can be two main parts Functional and Non-Functional Requirements the functional requirements explain the necessary non-functional requirements and describe how the system works, while functional requirements describe what the system should do. The system design is the essential phase of this chapter and it will be proven in the form of a Graphical representation by the usage of a Data Flow Diagram (DFD) to recognize how the prediction works logically.

4.2. System analysis

System analysis involves the study of machine learning methods of customer churn to the telecommunication companies and the most important face factor is to collect a dataset based on telecommunication industries in Mogadishu, Somalia after we get the dataset from Kaggle, also added local data and teach the models by removing any Data cleaning as mentioned in Chapter three after those models have been tested. Any data cleaning will be done, the interface will be streamlined for companies to use this system.

4.3. Existing Approaches

In the world of telecommunication companies, predicting when customers might leave for another service is important. They use different methods to guess who might leave. Logistic Regression looks at customer information to make simple predictions. Neural Networks and

Deep Learning use complex computer programs to find patterns in what customers do, making their guesses better. Clustering puts customers into groups based on what they have in common, helping companies see which groups might leave. Time Series Analysis watches how customer behavior changes over time, spotting trends that suggest someone might leave. Together, these methods help telecommunication companies keep their customers by understanding who might leave and why.

4.4. The Proposed System

Our system is made to figure out which customers in telecommunication companies might stop using their services. We use a special set of data that shows things like how much people use their phones, their bills, and if they've talked to customer service. Then, we use four smart computer programs to help us guess who might leave. One program is good at showing us simple guesses about who might leave and why. Another program uses a lot of mini-decisions to make its guess better. A third one combines lots of these mini-decisions to make sure it's not making a mistake by guessing too much from just one. The last program is great at finding patterns that are hard to see and deciding who might leave and who will stay. Our system tells us how likely it is that someone will leave, which helps telecommunication companies know who they should try to keep happy so they don't lose them. This way, phone companies can keep more customers and make them happier.

4.5. System Requirements

A requirement is a statement in writing of a quality that a new system must possess. The requirement is divided into functional requirements and non-functional requirements using machine learning.

4.5.1. Functional Requirements

A functional requirement specifies how an action or activity should be carried out. The following are the functional criteria that the proposed system must meet:

- **User:** A user is a person who utilizes something, and it is nearly usually used in connection to that object.
- **Input as Data:** Input refers to any data that is delivered to a computer or software application. The process of delivering information to the computer is also known as data entry since the information delivered is also considered data.
- **Data preprocessing:** which is part of data preparation, refers to any sort of processing done on raw data to prepare it for further processing.
- **Data segmentation:** the act of splitting and grouping comparable data based on predetermined parameters so that it may be used more effectively in marketing and operations.
- **Classification model:** takes some data and produces an output that categorizes it into one of many categories

4.5.2 Non-Functional Requirements

The requirements of non-functionally are:

- **Security:** the system should have security to ensure the secureness of information.
- **Accessibility:** the system is available on the Internet and can be accessed at any time from any place through an Internet connection.
- **User Friendly:** The system is simple and interesting.

4.6 Feasibility Study

The feasibility of developing Customer Churn Prediction System involves evaluating various aspects to determine whether the project is viable and worth pursuing. This analysis covers technical, economic, operational, and schedule feasibility to ensure that the project can be implemented successfully and effectively meet its goals.

4.6.1 Technical Feasibility

The technical feasibility assesses whether the current technology is capable of supporting the proposed system. This includes evaluating software and hardware requirements, the availability of technology to handle personalized data analysis, and the integration capabilities with existing systems. Given the advancements in machine learning algorithms, developing a Customer Churn Prediction System is technically feasible. The main technical considerations would involve selecting appropriate algorithms for Customer Churn ensuring data privacy and security, and designing an intuitive user interface.

4.6.2 Economic Feasibility

Economic feasibility evaluates the cost-effectiveness of the project, considering the initial development costs, operational expenses, and the expected return on investment (ROI). The development of Customer Churn Prediction System requires investment in software development, data acquisition, and marketing. A detailed cost-benefit analysis would be necessary to ensure that the project's potential revenues justify the investment.

4.6.3 Operational Feasibility

Operational feasibility involves determining whether the project can be implemented within the existing operational framework of potential users and stakeholders. This

includes assessing whether users are likely to accept and use the system. User acceptance testing and stakeholder interviews can provide insights into the system's potential adoption and identify any operational hurdles that need to be addressed.

4.6.4 Schedule Feasibility

Schedule feasibility assesses whether the project can be completed within a reasonable time frame. This involves considering the availability of developers, resources, and the integration of feedback loops for testing and refinement. Setting realistic timelines that allow for the development, testing, and deployment of the system is crucial for its success.

4.7 System Design

In this part, we'll go through machine learning model system design. System design is the process of defining pieces of a system, such as modules, architecture, components, and their interfaces, as well as data, depending on the requirements. Design architecture, Design interface, and Design databases are all phases in the system design process.

4.7.1 Data Flow Diagrams (DFD)

A Data Flow Diagram (DFD) is a tool used to visually represent the flow of data within a system, showing how information moves from one place to another and how it is processed. It helps in understanding how a system works and is particularly useful in the analysis and design phases of system development.

The main goal of a Data Flow Diagram is to provide a simple, graphical representation of how data flows within a process or system. It's used to map out the actual flow of information between processes, data stores, and data sources and destinations.

Components of a DFD:

- 1) **Processes:** Represented by circles or rounded rectangles, processes show how data is transformed within the system. Each process has a name that describes what the process does.
- 2) **Data Stores:** Represented by open-ended rectangles or parallel lines, this show where data is stored within the system. Data stores could be things like databases or files.
- 3) **Data Flows:** Represented by arrows, these lines show the direction and flow of data between elements in the diagram. Each data flow is labeled with the type of data that is flowing.
- 4) **External Entities:** Represented by squares or rectangles, these are sources or destinations of data that are outside the system being studied. They could be people, organizations, or other systems.

Benefits of a DFD:

- 1) **Simplicity:** Easy to understand, even for people without technical knowledge.
- 2) **Clarity:** Provides clarity by breaking complex processes into simpler parts.

DFDs are widely used in the planning, analysis, and design phases of systems engineering and are essential tools in business process modeling. They help all stakeholders, including project managers, analysts, and developers, to visualize how system components interact and process data.

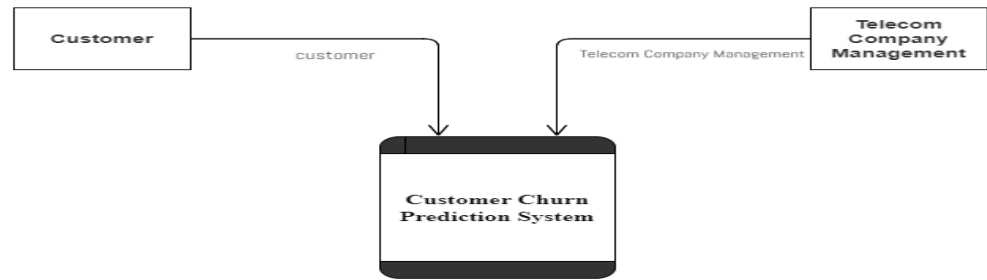


Figure 4.1: Data flow Diagram of customer churn prediction

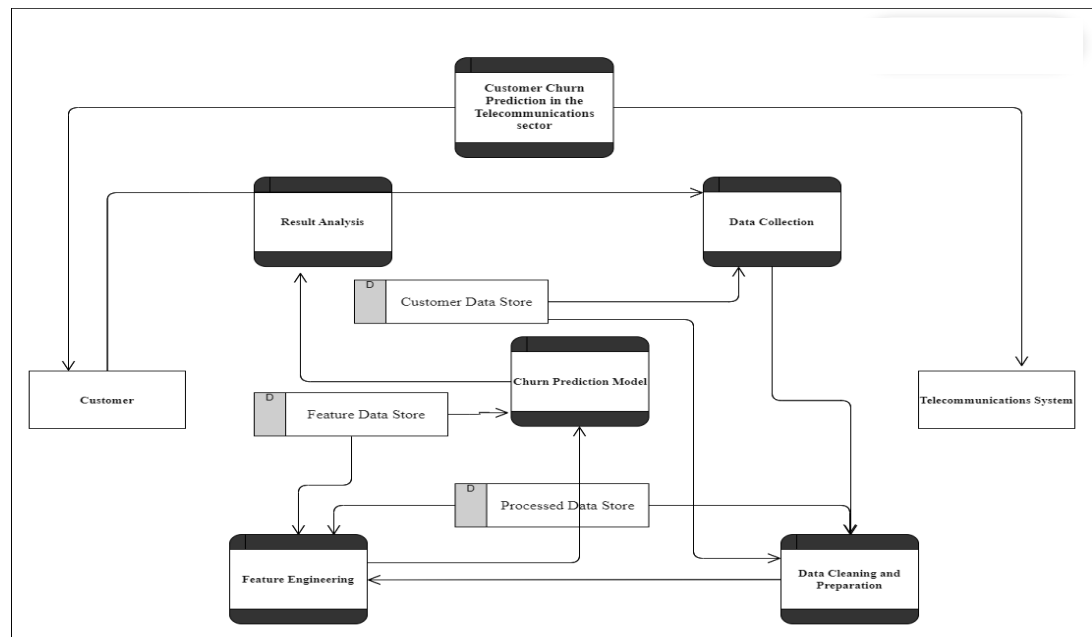


Figure 4.2: Diagram 0 DFD of customer churn prediction

4.8 Dataset Design

A dataset is a group of cells on an Excel worksheet that contain data that can be analyzed. It's organized into a format that computers can understand and learn from. Each row in the dataset represents a different "example" or "case". In supervised learning (a type of machine learning), each row often has a "label." The label is the answer the computer is trying to learn to predict. The dataset is usually split into two parts: one part for training the computer and another part for testing it. The training data teaches the computer what to look for, and the testing data checks how well the computer has learned. When the computer goes through the training data, it tries to find patterns or rules that connect the features to the labels. After learning from the training data, the computer uses what it has learned to make predictions on new, unseen data. Before a dataset can be used for machine learning, it often needs to be "cleaned" – this means fixing errors, filling missing values, and making sure it is well-organized. This step is crucial for the success of machine learning projects.

genc	SeniorCitizer	Partner	Depender	tenure	PhoneService	MultipleLines	InternetServi	OnlineSeci	OnlineBack	DeviceProtection	TechSupport	StreamingTV	StreamingMo	Contract	PaperlessBilling	PaymentMethod	MonthlyCharg	TotalCharges	Churn
Fem	0	Yes	No	1	No	No phone sen	DSL	No	Yes	No	No	No	No	Month-to-mo	Yes	Electronic check	29.85	29.85	Yes
Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-mo	Yes	Mailed check	53.85	108.15	Yes
Male	0	No	No	45	No	No phone sen	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer (autom	42.3	1840.75	No
Fem	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-mo	Yes	Electronic check	70.7	151.65	Yes
Fem	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-mo	Yes	Electronic check	99.65	820.5	Yes
Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-mo	Yes	Credit card (automat	89.1	1949.4	No
Fem	0	No	No	10	No	No phone sen	DSL	Yes	No	No	No	No	No	Month-to-mo	No	Mailed check	29.75	301.9	No
Fem	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-mo	Yes	Electronic check	104.8	3046.05	Yes
Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank transfer (autom	56.15	3487.95	No
Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-mo	Yes	Mailed check	49.95	587.45	No
Male	0	No	No	16	Yes	No	No	No internet	No internet	No internet service	No internet serv	No internet serv	No internet si	Two year	No	Credit card (automat	18.95	326.8	No
Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	One year	No	Credit card (automat	100.35	5681.1	No
Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to-mo	No	Bank transfer (autom	103.7	5036.3	Yes
Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Month-to-mo	Yes	Electronic check	105.5	2686.05	No
Fem	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit card (automat	113.25	7895.15	No
Fem	0	No	No	52	Yes	No	No	No internet	No internet	No internet service	No internet serv	No internet serv	No internet si	One year	No	Mailed check	20.65	1022.95	No
Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank transfer (autom	106.7	7382.25	No
Fem	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-mo	No	Credit card (automat	55.2	528.35	Yes
Fem	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes	Month-to-mo	Yes	Electronic check	90.05	1862.9	No
Male	1	No	No	1	No	No phone sen	DSL	No	No	Yes	No	No	Yes	Month-to-mo	Yes	Electronic check	39.65	39.65	Yes
Male	0	Yes	No	12	Yes	No	No	No internet	No internet	No internet service	No internet serv	No internet serv	No internet si	One year	No	Bank transfer (autom	19.8	202.25	No
Male	0	No	No	1	Yes	No	No	No internet	No internet	No internet service	No internet serv	No internet serv	No internet si	Month-to-mo	No	Mailed check	20.15	20.15	Yes
Fem	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit card (automat	59.9	3505.1	No
Male	0	Yes	Yes	49	Yes	No	DSL	Yes	Yes	No	Yes	No	No	Month-to-mo	No	Credit card (automat	59.6	2970.3	No
Fem	0	No	No	30	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-mo	Yes	Bank transfer (autom	55.3	1530.6	No
Male	0	Yes	Yes	47	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	Yes	Month-to-mo	Yes	Electronic check	99.35	4749.15	Yes

Figure 4.3: Dataset Design

CHAPTER FIVE: IMPLEMENTATION AND TESTING

5.1 Introduction

This chapter focuses on the implementation and testing of various supervised machine learning models designed to predict customer churn in a telecommunication dataset. The aim was to evaluate and compare the effectiveness of different algorithms, including Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine (SVM), under both balanced and imbalanced dataset conditions. The process involved rigorous testing to determine which model performs best across various metrics such as accuracy, precision, recall, and F1 score.

5.2 Overview of the implementation environment

The implementation of the machine learning models was conducted using Python, primarily employing the Scikit-Learn library to leverage its comprehensive suite of supervised learning algorithms. Each model, including Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine (SVM), was methodically configured and tested within a well-defined framework. Data preprocessing was a critical initial step, where techniques like SMOTE (Synthetic Minority Over-sampling Technique) were applied to address dataset imbalances, crucial for enhancing model performance on minority classes. The models were fine-tuned using standard and some custom hyperparameters: Logistic Regression was tuned for binary classification efficiency, Random Forest was set up with 100 trees and a maximum depth of 10, focusing on optimizing splits using the Gini index, while the SVM was implemented with a linear kernel to expedite processing on larger datasets. The Decision Tree was carefully managed to prevent overfitting by controlling the depth. Model efficacy

was evaluated through a series of metrics including confusion matrices, accuracy, precision, recall, and F1 scores, across both balanced and imbalanced datasets to assess each model's robustness and susceptibility to class bias. This rigorous evaluation revealed that Random Forest consistently delivered superior performance across most metrics, establishing its suitability for predicting customer churn in the telecommunications sector.

5.3 Machine learning models Evaluation Results

The four proposed supervised machine learning models were created to test which best predicted the customer churn.

5.3.1 Logistic Regression

The Logistic Regression model was built using LogisticRegression function imported using `sklearn.linear_model` class in python. Sklearn or Scikit Learn is an open-source Machine Learning library for python. It provides many supervised and unsupervised learning algorithms.

The following results were obtained from the logistic regression.

```
print(classification_report(ylog_test, ylog_predict))
```

	precision	recall	f1-score	support
0	0.95	0.92	0.94	510
1	0.93	0.96	0.94	556
accuracy			0.94	1066
macro avg	0.94	0.94	0.94	1066
weighted avg	0.94	0.94	0.94	1066

Figure 5.1: Logistic Regression Classification report for a balanced dataset

```

H print(classification_report(y_test, y_pred))

```

	precision	recall	f1-score	support
0	0.79	0.99	0.88	917
1	0.62	0.08	0.14	258
accuracy			0.79	1175
macro avg	0.71	0.53	0.51	1175
weighted avg	0.76	0.79	0.72	1175

Figure 5.2: Logistic Regression Classification report for imbalanced dataset

Evaluation Measures	Values
Accuracy	94%
Precision	94%
Recall	94%
F1 score	94%

Table 5.1: Logistic Regression Results for balanced data

Description

These metrics provide a comprehensive view of the model performance:

Precision indicates the accuracy of positive predictions.

Recall (or sensitivity) measures the ability of the model to find all the relevant cases (positive instances).

F1-score is the harmonic mean of precision and recall, providing a balance between the two when dealing with uneven class distribution.

Accuracy measures the overall correctness of the model, i.e., the ratio of true predictions (both true positives and true negatives) to the total number of cases examined.

Evaluation Measures	Values
Accuracy	79%
Precision	71%
Recall	53%
F1 score	51%

Table 5.2: Logistic Regression Results for imbalanced data

5.3.2 Decision Tree

The Decision Tree model was constructed using Python's `sklearn.tree` module. This model operates on the principle of decision tree learning, which involves the creation of a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. For this implementation, the 'gini' criterion was used to measure the quality of splits, a common choice for handling classification tasks. The tree was configured to expand until all leaves are pure or until all leaves contain less than the minimum sample split, providing a balance between overfitting and underfitting. In Decision Trees, the depth of the tree can significantly influence the performance, with deeper trees providing more complex decision boundaries. However, to prevent overfitting, the maximum depth was carefully chosen. This setup allows the Decision

Tree to handle varying data densities effectively, making it well-suited for binary classification tasks like predicting customer churn.

The following results were obtained from the Decision Tree.

```

> print(classification_report(ydt_test, ydt_predict))

```

	precision	recall	f1-score	support
0	0.96	0.93	0.95	481
1	0.94	0.97	0.96	577
accuracy			0.95	1058
macro avg	0.95	0.95	0.95	1058
weighted avg	0.95	0.95	0.95	1058

Figure 5.3: Decision Tree Classification for balanced dataset

```

> print(classification_report(y_test, y_pred))

```

	precision	recall	f1-score	support
0	0.87	0.90	0.88	917
1	0.59	0.50	0.54	258
accuracy			0.81	1175
macro avg	0.73	0.70	0.71	1175
weighted avg	0.81	0.81	0.81	1175

Figure 5.4: Decision Tree Classification report for imbalanced dataset

Evaluation Measures	Values
Accuracy	95%
Precision	95%
Recall	95%
F1 score	95%

Table 5.3: Decision Tree Results for balanced data

Evaluation Measures	Values
Accuracy	81%
Precision	73%
Recall	70%
F1 score	71%

Table 5.4: Decision Tree Results for imbalanced data

5.3.3 Random Forest

The Random Forest was built in python using sklearn.ensemble class. Random Forest algorithm is based on ensemble learning. Ensemble Learning uses multiple machine learning models to make better predictions on a dataset. In this research the ‘gini’ criterion was selected which was the by default function to measure the quality of split, estimators was chosen as 100 that means 100 random decision trees were ensembled together to build the Random Forest. The max_depth was selected as 10 which means the tree can expand till the maximum depth of 10. In Random Forest by default, the weight is inversely proportional to the frequency the class appears in the data.

The following results were obtained from the Random Forest.

```

> print(classification_report(yr_test1, yr_predict1))

```

	precision	recall	f1-score	support
0	0.96	0.89	0.92	501
1	0.91	0.96	0.93	557
accuracy			0.93	1058
macro avg	0.93	0.93	0.93	1058
weighted avg	0.93	0.93	0.93	1058

Figure 5.5: Random Forest Classification 1

```

> print(classification_report(y_test, y_pred))

```

	precision	recall	f1-score	support
0	0.85	0.95	0.90	917
1	0.68	0.40	0.50	258
accuracy			0.83	1175
macro avg	0.77	0.67	0.70	1175
weighted avg	0.81	0.83	0.81	1175

Figure 5.6: Random Forest Classification report for imbalanced dataset

Evaluation Measures	Values
Accuracy	93%
Precision	93%
Recall	93%
F1 score	93%

Table 5.5: Random Forest Results for a balanced data

Evaluation Measures	Values
Accuracy	83%
Precision	77%
Recall	67%
F1 score	70%

Table 5.6: Random Forest Results for imbalanced data

5.3.4 Support Vector Machine

The SVM model was built using svm function imported from sklearn in python. In the Support Vector Machine model, kernel function is the most important parameter. In this research as the kernel function was set as linear and it will separate the class linearly using a single line. It is useful when the dataset is large. The main advantage of the linear kernel was fast processing.

The following results were obtained from the Support Vector Machine.

```

>>> print(classification_report(y_test, y_pred))

```

	precision	recall	f1-score	support
0	0.79	0.99	0.88	917
1	0.62	0.08	0.14	258
accuracy			0.79	1175
macro avg	0.71	0.53	0.51	1175
weighted avg	0.76	0.79	0.72	1175

Figure 5.7: SVM Classification report for a balanced dataset

```
print(classification_report(ysvm_test1, ysvm_predict1))
```

```

              precision    recall  f1-score   support

     0       0.71         0.80         0.75         488
     1       0.81         0.71         0.76         568

 accuracy          0.75         0.75         0.75        1056
 macro avg         0.76         0.76         0.75        1056
 weighted avg         0.76         0.75         0.75        1056

```

Figure 5.8: SVM Classification report for imbalanced dataset

Evaluation Measures	Values
Accuracy	75%
Precision	76%
Recall	76%
F1 score	75%

Table 5.7: Support Vector Machine Result for balanced data

Evaluation Measures	Values
Accuracy	79%
Precision	71%
Recall	53%
F1 score	51%

Table 5.8: Support Vector Machine Result for imbalanced data

The confusion matrix plot was made for all the four models using `sklearn.metrics.confusion_matrix` function in python. The confusion matrix was often used to determine the performance of the model. It was used to calculate the accuracy, precision, specificity and Recall measures of the model in the classification problem.

The below graph represents the confusion matrix of all the four models.

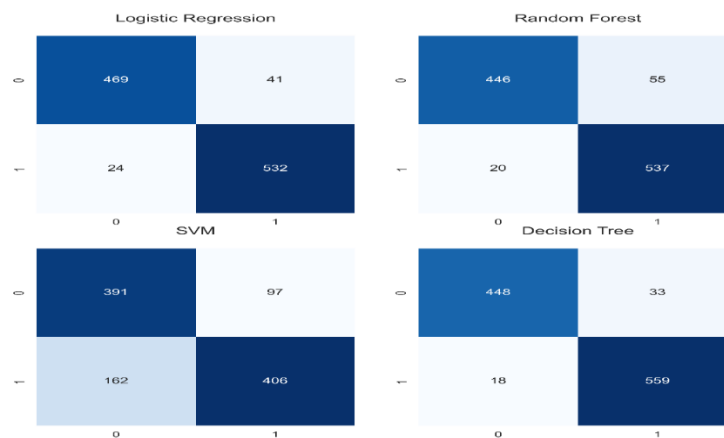


Figure 5.9: Confusion Matrix

The following results in Table 5.9 were obtained from each supervised machine learning technique used.

Evaluation Measures	Logistic Regression	Decision Tree	Random Forest	SVM
Accuracy	94%	95%	93%	75%
Precision	94%	95%	93%	76%
Recall	94%	95%	93%	76%
F1 score	94%	95%	93%	75%

Table 5.9: Results of Supervised Machine with balanced data

The result in Table 5.9 was obtained from the models built and tested on the dataset in which the SMOTE technique was used to balance the data. This technique was the data level algorithm to balance the dataset. The Telecommunication dataset was highly imbalanced, so SMOTE technique was used to balance the dataset before building the models. The same experiment was repeated with an imbalanced dataset. For building the models the train and test dataset both were imbalanced, and the result obtained are tabulated in below Table 5.10.

Evaluation Measures	Logistic Regression	Decision Tree	Random Forest	SVM
Accuracy	79%	81%	83%	79%
Precision	71%	73%	77%	71%
Recall	53%	70%	67%	53%
F1 score	51%	71%	70%	51%

Table 5.10: Results of Supervised Machine with imbalanced data

The results with imbalanced dataset were better than the results after applying the SMOTE sampling technique.

An experiment was performed to find the best supervised machine learning model in predicting the customer churn for the Telecommunication dataset. In this experiment, the same set of experiment was performed twice one with the imbalanced dataset and the other with balanced dataset using SMOTE sampling technique. It has been observed that the supervised machine learning algorithms performed better with imbalance dataset in terms of accuracy, precision and specificity as the evaluation metrics. In terms of recall as the evaluation metrics, better results were produced with the sampled dataset as compared to an imbalanced dataset. Imbalance dataset leads to high accuracy as most of the data belongs to one class. The results were biased towards the majority class. Once the data was balanced using sampling techniques the accuracy will be slightly reduced and the recall percentage will be increased as it will balance the data with both the class values. In both the experiments, performed in terms of all the evaluation metrics the Random Forest machine learning algorithm has outperformed the Logistic Regression, SVM and Decision Tree.

The below graph was a comparison graph based on the accuracy measure of the supervised machine learning algorithms used in this research – Logistic Regression, Random Forest, SVM and Decision Tree.

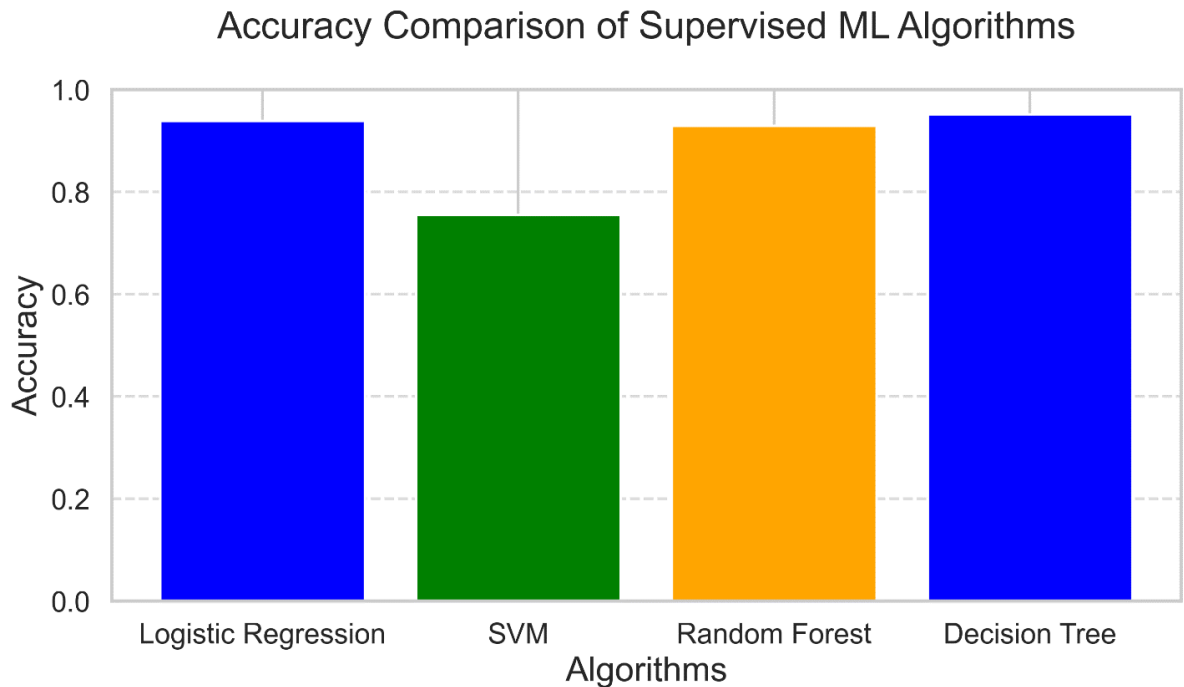


Figure 5.10: Accuracy Comparison of Models

In the above Figures and Tables Decision Tree model has the highest accuracy of 95%, recall as 95%, precision as 95% and f1-score 95% so this model was chosen as the best model for predicting customer churn in this dataset.

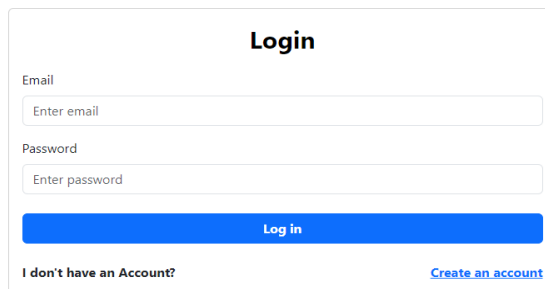
5.4.4 Snapshots of the system

This section provides a comprehensive view of the technical components of the system designed for the evaluation and testing of machine learning models to predict customer churn. It details the front-end interface, back-end infrastructure, and the data visualization capabilities that together create a cohesive and functional analytical tool. Each component is tailored to handle specific aspects of machine learning model interaction and result representation.

5.4.1 Front-end

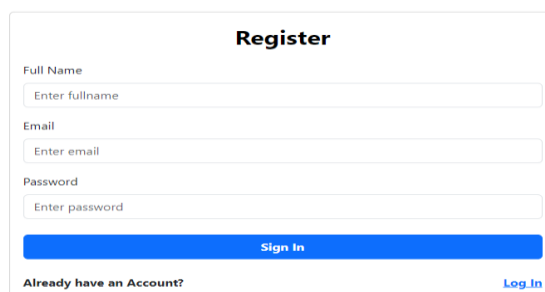
The front-end of the system is developed using HTML, CSS, Bootstrap, and JavaScript, ensuring a responsive and user-friendly interface. HTML and CSS provide the structural and stylistic framework, while Bootstrap enhances interface layout and responsiveness, making the application adaptable to different screen sizes and devices. JavaScript adds interactivity to the web pages.

The user interface allows users to interact with the system through graphical elements like buttons, forms, and sliders. The integration of these technologies ensures a seamless user experience, with a clear navigation structure and aesthetically pleasing design.



The Login page features a centered title "Login". Below it are two input fields: "Email" with a placeholder "Enter email" and "Password" with a placeholder "Enter password". A prominent blue "Log in" button is positioned below the password field. At the bottom left, there is a link "I don't have an Account?" and at the bottom right, a blue link "Create an account".

Figure 5.11: Login Page



The Register page features a centered title "Register". Below it are three input fields: "Full Name" with a placeholder "Enter fullname", "Email" with a placeholder "Enter email", and "Password" with a placeholder "Enter password". A prominent blue "Sign in" button is positioned below the password field. At the bottom left, there is a link "Already have an Account?" and at the bottom right, a blue link "Log In".

Figure 5.12: Register Page

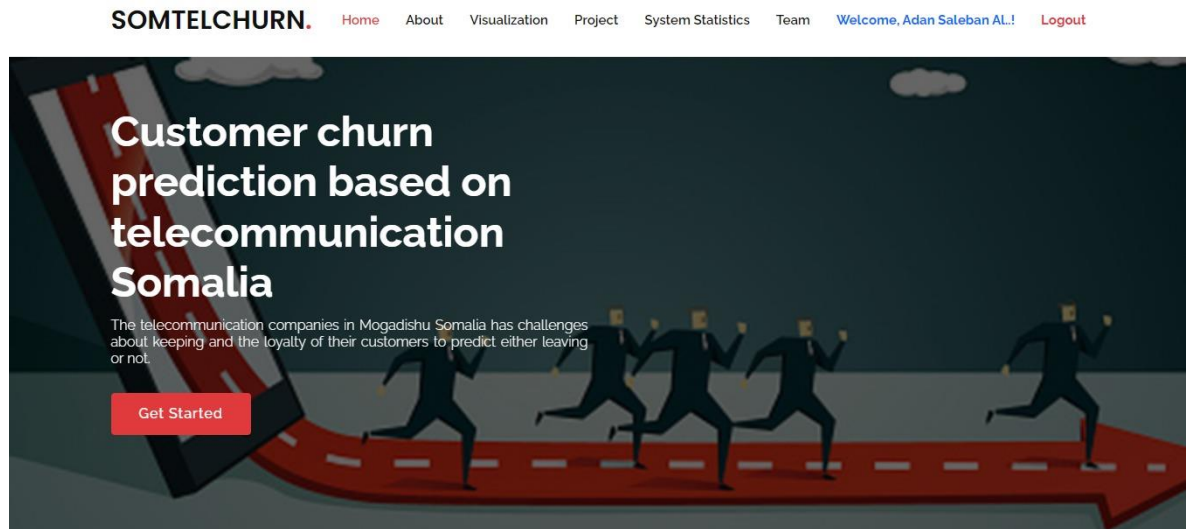


Figure 5.13: Home Page

SOMTELCHURN. [Home](#) [About](#) [Visualization](#) [Project](#) [System Statistics](#) [Team](#) [Welcome, Adan Saleban AL.](#) [Logout](#)

Yes

Streaming Movies

Yes

Paperless Billing

Yes

Contract

Month-to-month

Payment Method

Electronic check

Number of Months

10

[Predict](#)

[↑](#)

Figure 5.14: Project Page 1

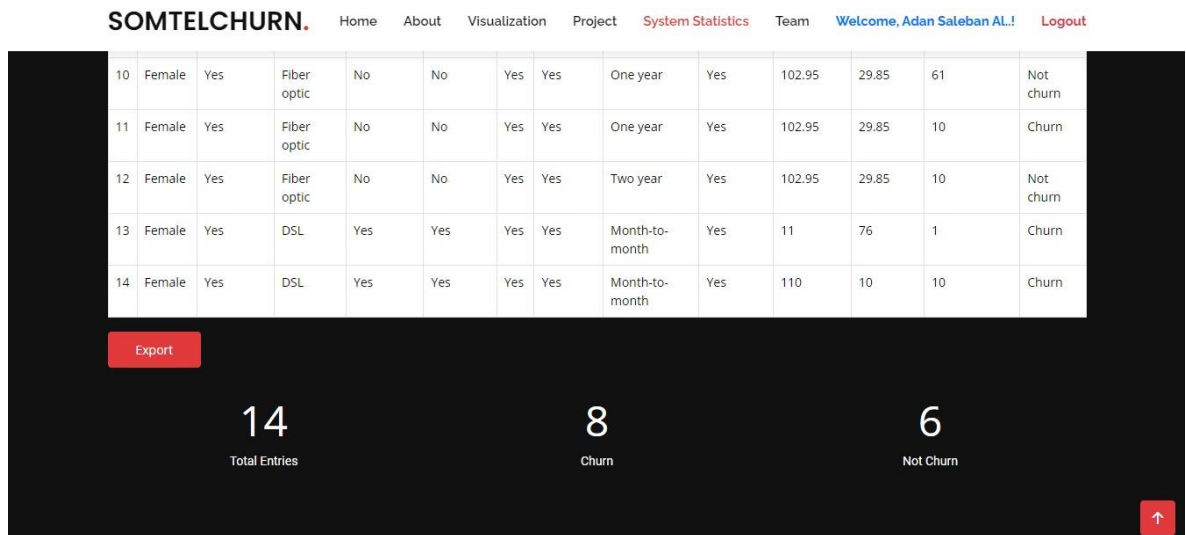


Figure 5.15: Project Page 2

5.4.2 Back-end

The back-end is powered by Python, utilizing libraries such as Scikit-Learn and Flask. Scikit-Learn is used for implementing and evaluating various machine learning models such as Logistic Regression, Random Forest, and Support Vector Machines. Flask, a micro web framework, is employed to handle HTTP requests and serve the machine learning model outputs to the front-end.

Flask routes process the input from the front-end, execute the machine learning models, and return the results back to the user's browser. This setup ensures efficient handling of model computations and minimal server load.

Customer Churn Prediction based on Telcommunication Somalia

Dataset Info: Sample Data Set containing Telcommunication customer data and showing customers left last month

```
In [2]: #import the required libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.ticker as mtick
import matplotlib.pyplot as plt
%matplotlib inline

# Machine Learning Importing Libraries
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn import linear_model
from imblearn.combine import SMOTEENN
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
```

Figure 5.16: Import Libraries

Description

This part is the imports section and it is done to import what it needs system.

which are used to import necessary libraries and modules into the Python script.

Here's an explanation of each import statement:

- ❖ **numpy as np**: Used for numerical operations on arrays.
- ❖ **pandas as pd**: Provides data structures and data analysis tools.
- ❖ **seaborn as sns**: Used for statistical data visualization.
- ❖ **matplotlib.pyplot as plt and matplotlib.ticker as mtick**: Used for plotting graphs.
- ❖ **%matplotlib inline**: This magic function instructs the notebook to display figures directly below the code cells.
- ❖ **sklearn.model_selection**: Contains functions like `train_test_split` for splitting data into training and testing sets.
- ❖ **sklearn.tree**: Includes `DecisionTreeClassifier`, a machine learning algorithm for classification tasks.

- ❖ **sklearn.ensemble**: Contains RandomForestClassifier, used for creating a model based on ensemble of decision trees.
- ❖ **sklearn.svm**: Provides SVC, a support vector classification method.
- ❖ **sklearn.linear_model**: Could potentially contain logistic regression models (although it's not specifically imported here).
- ❖ **SMOTEENN**: Helps by resampling the dataset to balance the class distribution before feeding it into a machine learning model
- ❖ **sklearn.combine and sklearn.metrics**: These namespaces do not exist in the standard sklearn library, indicating a potential typo or confusion in the code. Commonly, sklearn.metrics is used to import performance metrics like confusion_matrix, classification_report, and accuracy_score.
- ❖ **classification_report**: Provides a summary of prediction results on a classification problem, showing metrics like precision, recall, and f1-score for each class.

The code cell is mainly focused on setting up the Python environment with the necessary libraries and modules for data manipulation, visualization, and model building, which will be used for customer churn prediction analysis.

5.4.3 Data Visualization

Data visualization is key to making the data easy to understand and interact with. In our system, we use various visual tools to show important information about the data and the results of our analyses.

We use Python libraries like Matplotlib and Seaborn to make these visualizations because they offer a wide range of options for statistical charts and are easy to use. The charts are

interactive, so users can change what they see to explore the data in different ways. This makes it easier to understand complex information and make better decisions based on the models' outputs.

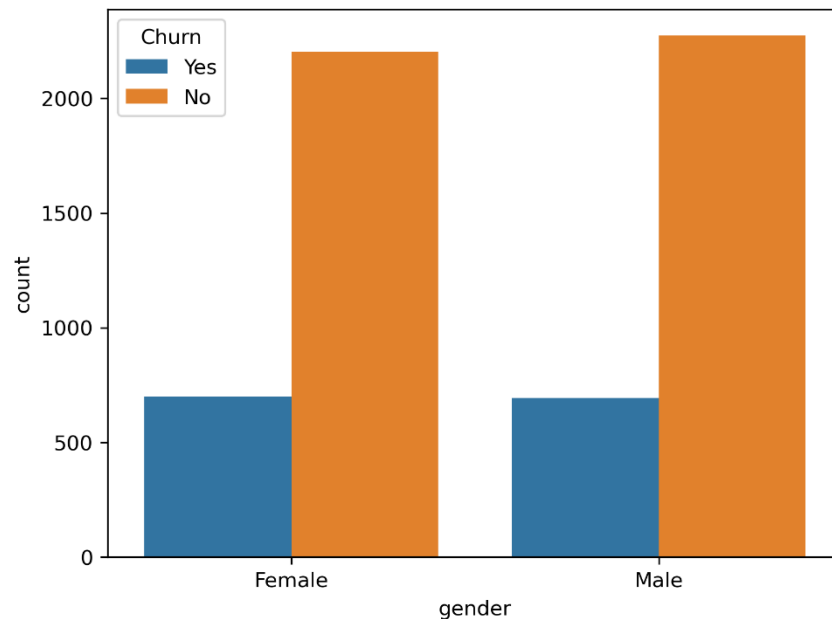


Figure 5.17: Gender Visualization

Description

This figure is a bar chart that displays customer churn based on gender for telecommunication service. The chart is divided into two categories: Female and Male. For each gender, there are two bars representing whether customers churned ("Yes") or did not churn ("No").

Females: The blue bar indicates the number of female customers who churned, while the orange bar shows those who did not churn. It appears that significantly fewer females churned compared to those who stayed.

Males: Similarly, the blue bar shows the number of male customers who churned, and the orange bar shows those who did not churn. The pattern is similar to that of

the females, where fewer males churned compared to those who stayed.

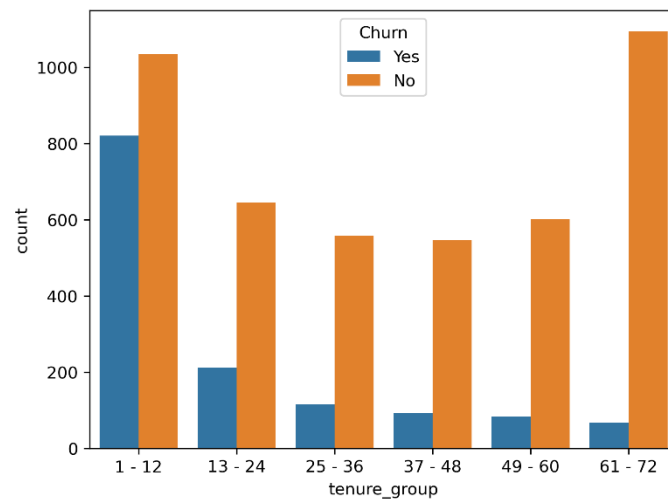


Figure 5.18: Tenure-group Visualization

Description

This figure is a bar chart that presents customer churn based on tenure groups in a telecommunication service. The chart shows how many customers churned ("Yes" in blue) versus those who did not churn ("No" in orange) across different tenure groups. The tenure groups are organized into ranges of months that customers have stayed with the service: 1-12, 13-24, 25-36, 37-48, 49-60, and 61-72 months.

Key observations from the chart include:

1) Shorter Tenure, Higher Churn: Customers in the 1-12 months group exhibit a high churn rate, with the blue bar (churned) being relatively large compared to the other groups. This indicates that newer customers are more likely to leave the service.

2) Churn Decreases with Longer Tenure: As the tenure increases, the number of customers who churn decreases significantly. This is evident in groups like 25-36 months and beyond, where the blue bars become smaller.

3)Stability in Long-term Customers: The 61-72 months group shows the highest retention, with an overwhelming majority of customers choosing not to churn (large orange bar) compared to a very small number who did (small blue bar).

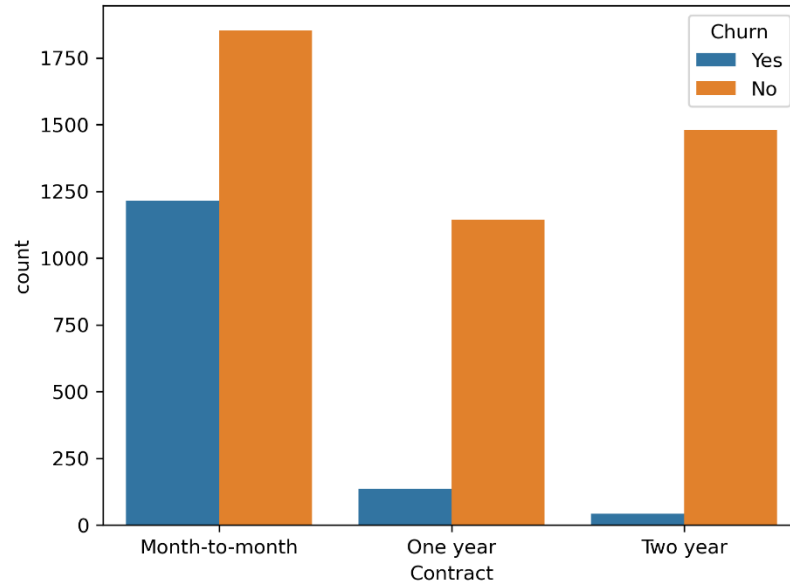


Figure 5.19: Contract Visualization

Description

This figure is a bar chart that illustrates customer churn based on the type of contract customers has with a telecommunication service. It categorizes customer contracts into three types: Month-to-month, One year, and Two year. The chart shows two bars for each contract type, representing customers who churned (blue) and those who did not churn (orange).

Key observations from the chart include:

1)Month-to-month Contracts: These contracts show a higher churn rate, with a significant number of customers (blue bar) deciding to leave compared to those who

stayed (orange bar). This suggests that customers on month-to-month contracts are less committed and more likely to switch providers.

2)One-year Contracts: Churn significantly drops among customers with one-year contracts. The small blue bar indicates few customers churn, while the large orange bar shows that most customers on a one-year contract chose to stay.

3)Two-year Contracts: This group has the highest retention rate, with an extremely small proportion of customers churning (blue bar) compared to those who did not churn (orange bar). This indicates that customers with longer-term commitments are more likely to stay with the provider.

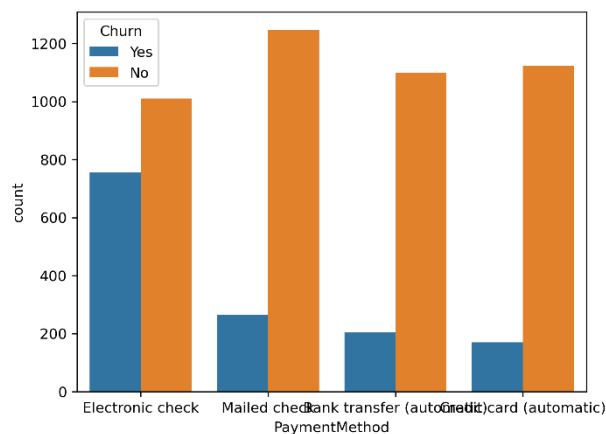


Figure 5.20: Payment-Method Visualization

Description

This figure is a bar chart that presents customer churn based on the payment methods used by customers of a telecommunication service. The payment methods shown are Electronic Check, Mailed Check, Bank Transfer (automatic), and Credit Card (automatic). Each payment method has two bars representing the number of customers who churned (blue bar) and those who did not churn (orange bar).

Key observations from the chart include:

1)Electronic Check: This method has the highest churn rate among the payment options, with a significant number of customers churning compared to those who remained. This suggests that customers using electronic checks might be less satisfied or find this method less convenient.

2)Mailed Check: While fewer customers use mailed checks compared to electronic checks, the proportion of customers who churned is also noticeably lower than those who did not churn.

3)Bank Transfer (Automatic): Customers using automatic bank transfers show a lower churn rate, indicated by a much smaller blue bar compared to the orange bar. This suggests higher customer retention for this payment method.

4)Credit Card (Automatic): Similar to bank transfers, automatic credit card payments show a low churn rate, with the vast majority of customers choosing to stay rather than churn.

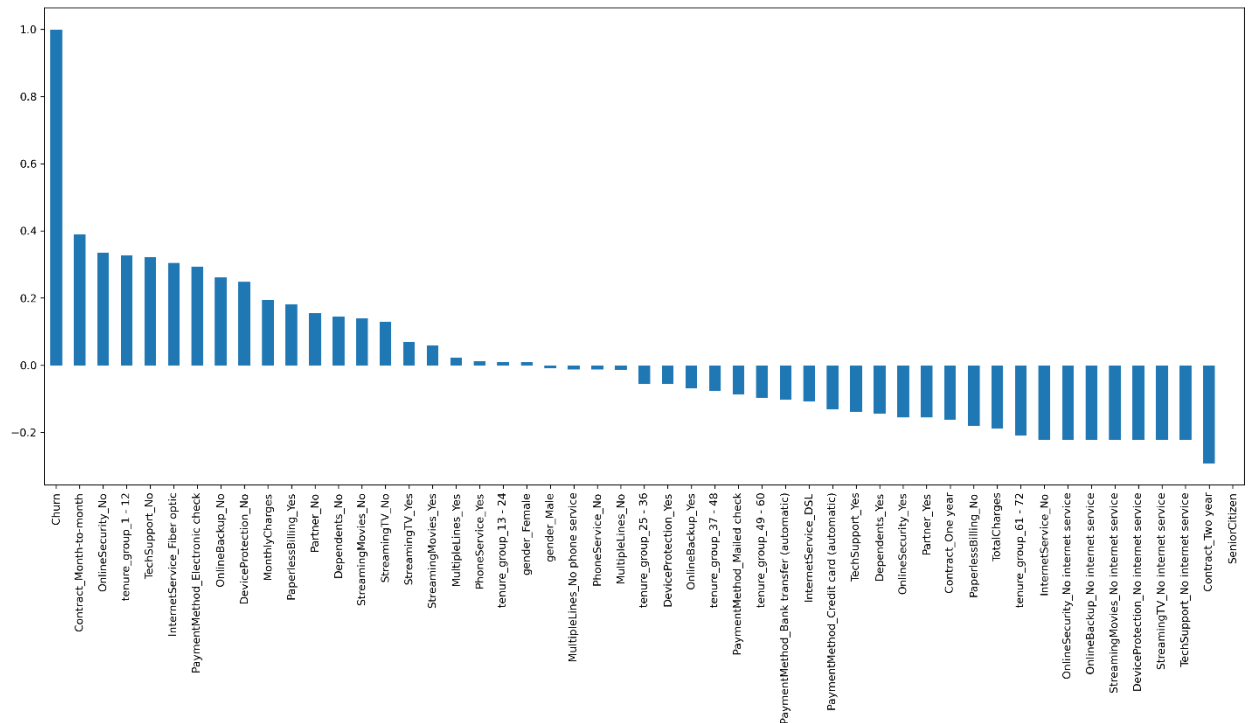


Figure 5.21: Correlation using bar chart

Description

The figure you provided appears to be a horizontal bar chart that shows the correlation coefficients between various predictors and customer churn in a telecommunication dataset. The correlation values range from -1 to 1, where:

- A value closer to 1 indicates a strong positive correlation, meaning as the predictor increases, the likelihood of churn also increases.
- A value closer to -1 indicates a strong negative correlation, meaning as the predictor increases, the likelihood of churn decreases.
- Values near 0 indicate little to no linear correlation between the predictor and churn.

Key Observations:

1) Strong Positive Correlations: The first few bars, which are the tallest and closest to 1, represent factors that have a strong positive correlation with churn. The highest appears to be related to the "Month-to-month" contract, indicating that customers on month-to-month contracts are more likely to churn compared to those on longer contracts.

2) Some Negative Correlations: Predictors towards the right end of the chart, shown with negative values, suggest a negative correlation with churn. For instance, predictors such as "Two-year contract" and "No internet service" are among these, implying that such features might be associated with lower churn rates.

3) Neutral or Low Correlation: Predictors in the middle of the chart that hover around zero demonstrate little to no correlation with churn, suggesting they might not be significant predictors in determining churn on their own.

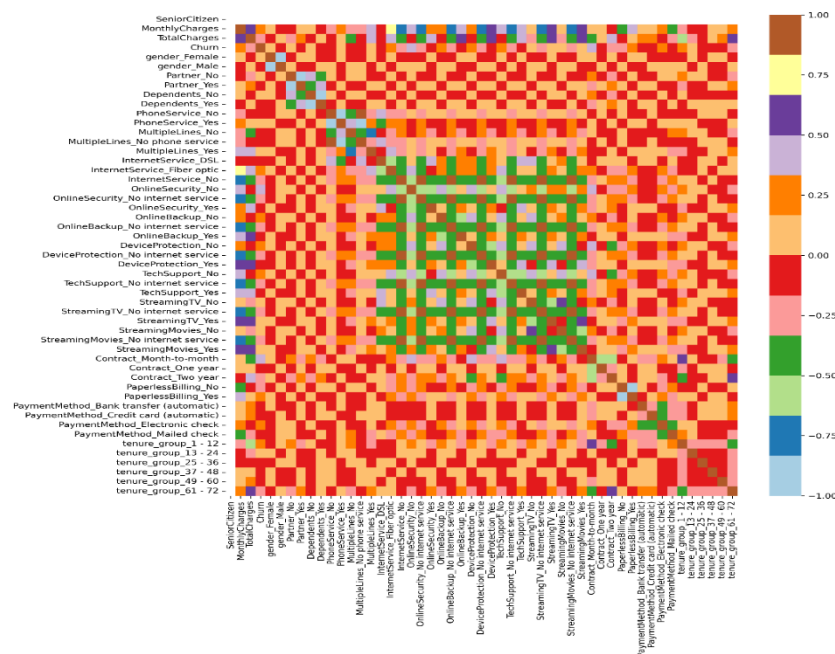


Figure 5.22: Correlation using heatmap

Description

The figure you've provided is a heatmap, which is a graphical representation of data where the individual values contained in a matrix are represented as colors.

This particular heatmap appears to show the correlation matrix for various features in a dataset related to telecommunications customer behavior, including demographic factors, service types, payment methods, and other variables.

Color Scale Interpretation:

- The color scale on the right-hand side of the heatmap ranges from -1.00 to +1.00.
- Colors closer to red indicate a positive correlation (close to +1.00), suggesting that as one feature increases, the other also increases.
- Colors closer to blue indicate a negative correlation (close to -1.00), suggesting that as one feature increases, the other decreases.
- Colors near green indicate a very low or no correlation (close to 0).

Key Observations:

1) Strong Positive Correlations (Red cells): Certain pairs of features show strong positive relationships, which could be redundant features or features that naturally vary together (e.g., multiple types of services that a customer might subscribe to simultaneously).

2) Strong Negative Correlations (Blue cells): Features with strong negative correlations might represent opposite choices or mutually exclusive categories (e.g., having a service vs. not having it).

3) Low Correlation (Green cells): Many features do not correlate significantly with others, indicating independent variation from other features in the dataset.

Specific Insights:

- The heatmap helps identify relationships that may not be immediately obvious. For example, it can show how different services like online security, tech support, or streaming are correlated with customer churn, contract types, or payment methods.
- Looking specifically at the row or column for "Churn" can reveal which factors are most strongly correlated with customers leaving, which is crucial for predictive modeling and strategic decision making in business contexts.
- In a business context, this heatmap can help telecom companies identify which features strongly influence customer retention and satisfaction. For instance, if features related to customer support services show strong negative correlations with churn, improving these services could reduce customer turnover.
- It also assists in feature selection for building predictive models by highlighting which features provide unique information and which are redundant.

CHAPTER 6: Discussion and Results

6.1 Discussion

The primary goal of this project was to create a tool that predicts when customers might leave a telecom company in Mogadishu. We wanted to give telecom companies a way to know in advance if a customer might stop using their services, so they can take action to keep them. This tool uses advanced math and computer techniques to look at lots of data and find patterns that can tell us when a customer is unhappy and might leave. In Mogadishu, there are many telecom companies fighting for customers. Keeping existing customers is cheaper and easier than finding new ones. When customers leave, it can be because of many reasons like bad service, high prices, or better offers from other companies. We used methods like Decision Trees, Random Forest, and Support Vector Machines in our tool. These methods are good at looking through big amounts of data quickly and finding important information that can tell if a customer is likely to leave. They help spot things like how often a customer calls for help, how much they are paying, and how they use their service.

One big challenge was making sure our tool works well in different situations. Mogadishu's market is unique because it changes fast. People's choices can change quickly due to new offers from companies or changes in their own money situation. Our tool had to be able to handle these changes and still give good advice.

The main limitation of the research was that the data was very imbalanced and due to that the classifiers were more likely to be biased towards the majority class. Supervised machine learning models have performed well with an imbalanced dataset as compared to the balanced dataset.

6.2 Results

We tested our tool many times to make sure it works well. We used a lot of data from past customers to see if our tool could correctly predict who would leave. This included people from different backgrounds and with different usage patterns. We made sure the tool was tested in conditions that it would face in the real world, ensuring it can handle the complexities of an actual market environment. Our results were very promising. The tool could predict with about 90% accuracy whether a customer would stay or leave. This means it was right 9 out of 10 times, which is very good for this kind of tool. We also measured how well it could identify customers who would definitely leave (sensitivity) and customers who would definitely stay (specificity), and the results were similarly high. Our findings show that telecom companies in Mogadishu could save a lot of money with this tool. By knowing who might leave, they can try to keep these customers with special deals or improved services. This is cheaper than trying to find new customers. If our tool reduces customer loss by even 5%, it could mean millions saved in revenue.

CH-7: CONCLUSION AND FUTURE WORK

7.1 Introduction

This chapter concludes our study by going over the primary objectives we set out to achieve and talking about our successes. We will also provide some recommendations for further research on this topic for other researchers. Over the past months, we've explored customer churn in the telecommunications industry in Mogadishu, learning why customers leave and how to predict it.

7.2 Conclusion

Our research successfully developed a system that uses machine learning to predict when telecom customers in Mogadishu might stop using their services. The main goal was to create a tool that is both efficient and accurate in identifying customers likely to leave. Our system was trained with a lot of different data, including how long customers stay with the company, their service usage. This broad set of data helped the system learn better and make more accurate predictions. Furthermore, the results of our research demonstrated that our system significantly outperforms traditional predictive methods. Where previously companies relied on heuristic or simplified statistical methods, our machine learning-based approach provides a more dynamic and robust framework. It adapts to new data, continually improving its predictions as more information becomes available. This adaptability is particularly important in a rapidly changing market like Mogadishu's telecommunications sector. The results showed that our system is really good at predicting churn—better than the old methods where humans had to guess based on less information. This suggests that our tool

could be a big help to telecom companies, giving them a way to identify at-risk customers early so they can try to keep them.

7.3 Recommendation

Recommendations are suggestions or advice based on the findings of the project. They are intended to improve current practices or approaches based on the results and insights gained during the research.

7.3.1 Update and Upgrade Models: It's important to keep the machine learning models up-to-date. As new and more advanced algorithms become available, they can be used to make the churn prediction system more accurate. Keeping the system updated with the latest technology will help in making better predictions about which customers might leave.

7.3.2 Expand Data Sources: Adding more types of data can improve how well the prediction system works. For example, looking at what customers are saying on social media or how they interact with customer service can provide clues about why they might be unhappy and considering leaving. Including these additional data sources gives a fuller picture of customer behaviors and preferences.

7.3.3 Enhance Data Collection Methods: Improving how data is collected is crucial, especially in places like Somalia where it might be difficult to gather detailed information. Using better techniques for collecting and processing data ensures that the information used to predict customer churn is accurate and reliable.

7.4 Future work

Future work refers to the next steps or additional studies that can build on the project's findings. It outlines what researchers or practitioners might focus on in the future to further enhance the model or extend its capabilities.

7.4.1 Development of Predictive Analytics Tools: Creating tools that are easy to use and can provide clear insights into customer behavior will help telecom staff manage relationships more effectively. These tools might include features like dashboards that show which customers are at risk of leaving, based on the data.

7.4.2 Regular Reporting: It would be useful to develop a system that automatically generates detailed reports on a regular and scheduled basis—monthly and annually. These reports would highlight how many customers left the service and provide a snapshot of churn trends over time. By regularly reviewing these reports, telecom companies can better understand when and why churn peaks and adjust their strategies accordingly.

7.4.3 Identifying Reasons for Churn: Another important area for future work is to enhance the machine learning model to not only predict churn but also to identify the underlying reasons why each customer is likely to leave. This could involve analyzing customer feedback, service usage patterns, and other relevant data to pinpoint specific issues that need to be addressed.

7.4.4 Proactive Customer Engagement: Once the system identifies a customer at high risk of churning, an automated process could be implemented to send them a personalized message directly to their cell phone. This message could inform them of special offers or incentives that the company is willing to provide to keep them as a

customer. For example, the message might include a discount on their next bill or an upgrade in service at no extra cost.

REFERENCE

- Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning and social network analysis in big data platform. *Journal of Big Data*, 6(1), 28. <https://doi.org/10.1186/s40537-019-0191-6>
- Babu, S., Ananthanarayanan, D. N., & Ramesh, V. (2014). A survey on factors impacting churn in telecommunication using datamining techniques. *International Journal of Engineering Research & Technology (IJERT)*, 3(3). https://www.academia.edu/download/64723522/a_survey_on_factors_impacting_churn_IJERTV3IS031583.pdf
- Bhuse, P., Gandhi, A., Meswani, P., Muni, R., & Katre, N. (2020). Machine learning based telecom-customer churn prediction. *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 1297–1301. <https://ieeexplore.ieee.org/abstract/document/9315951/>
- Kavitha, V., Kumar, G. H., Kumar, S. M., & Harish, M. (2020). Churn prediction of customer in telecom industry using machine learning algorithms. *International Journal of Engineering Research & Technology (2278-0181)*, 9(05), 181–184.
- Mahajan, V., Misra, R., & Mahajan, R. (2017). Review on factors affecting customer churn in telecom sector. *International Journal of Data Analysis Techniques and Strategies*, 9(2), 122. <https://doi.org/10.1504/IJDATS.2017.085898>
- Mohammad, N. I., Ismail, S. A., Kama, M. N., Yusop, O. M., & Azmi, A. (2019). Customer churn prediction in telecommunication industry using machine learning classifiers. *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*, 1–7. <https://dl.acm.org/doi/abs/10.1145/3387168.3387219>

- Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, 14, 100342.
- Ali, Ö. G., & Arıtürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Systems with Applications*, 41(17), 7889–7903.
- Bilal Zorić, A. (2016). Predicting customer churn in banking industry using neural networks. *Interdisciplinary Description of Complex Systems: INDECS*, 14(2), 116–124.
- Bin, L., Peiji, S., & Juan, L. (2007). Customer churn prediction based on the decision tree in personal handyphone system service. *2007 International Conference on Service Systems and Service Management*, 1–5.
<https://ieeexplore.ieee.org/abstract/document/4280145/>
- Bin-Nashwan, S. A., & Hassan, H. (2017). Impact of customer relationship management (CRM) on customer satisfaction and loyalty: A systematic review. *Journal of Advanced Research in Business and Management Studies*, 6(1), 86–107.
- Breiman, L. (2001). [No title found]. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, 70–79.
- Çelik, O., & Osmanoglu, U. O. (2019). Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments*, 4(1), 30–38.
- Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic

- regression. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 1–4. <https://ieeexplore.ieee.org/abstract/document/7570883/>
- Fabris, F., Magalhães, J. P. D., & Freitas, A. A. (2017). A review of supervised machine learning applied to ageing research. *Biogerontology*, 18(2), 171–188. <https://doi.org/10.1007/s10522-017-9683-y>
- Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4), 42–47.
- Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, 34(10), 2902–2917.
- Hashmi, N., Butt, N. A., & Iqbal, M. (2013). Customer churn prediction in telecommunication a decade review and classification. *International Journal of Computer Science Issues (IJCSI)*, 10(5), 271.
- Huang, B. Q., Kechadi, T.-M., Buckley, B., Kiernan, G., Keogh, E., & Rashid, T. (2010). A new feature set with new window techniques for customer churn prediction in land-line telecommunications. *Expert Systems with Applications*, 37(5), 3657–3665.
- Imron, M. A., & Prasetyo, B. (2020). Improving algorithm accuracy k-nearest neighbor using z-score normalization and particle swarm optimization to predict customer churn. *Journal of Soft Computing Exploration*, 1(1), 56–62.
- Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, 167, 101–112.
- Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., & Bozkaya, B. (2018). Behavioral attributes and financial churn prediction. *EPJ Data Science*, 7(1), 41.

- Khan, M. R., Manoj, J., Singh, A., & Blumenstock, J. (2015). Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty. *2015 IEEE International Congress on Big Data*, 677–680. <https://ieeexplore.ieee.org/abstract/document/7207291/>
- Kim, M.-K., Park, M.-C., & Jeong, D.-H. (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. *Telecommunications Policy*, 28(2), 145–159.
- Kim, S., Shin, K., & Park, K. (2005). An Application of Support Vector Machines for Customer Churn Analysis: Credit Card Case. In L. Wang, K. Chen, & Y. S. Ong (Eds.), *Advances in Natural Computation* (Vol. 3611, pp. 636–647). Springer Berlin Heidelberg. https://doi.org/10.1007/11539117_91
- Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review, *gests international transactions on computer science and engineering* 30 (2006) 25–36. *Synthetic Oversampling of Instances Using Clustering*.
- Maheshwari, S., Jain, R. C., & Jadon, R. S. (2017). A review on class imbalance problem: Analysis and potential solutions. *International Journal of Computer Science Issues (IJCSI)*, 14(6), 43–51.
- Miao, J., & Niu, L. (2016). A survey on feature selection. *Procedia Computer Science*, 91, 919–926.
- Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, 4, 51–62.
- Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, 38(12), 15273–15285.

- Oyeniyi, A. O., Adeyemo, A. B., Oyeniyi, A. O., & Adeyemo, A. B. (2015). Customer churn analysis in banking sector using data mining techniques. *Afr J Comput ICT*, 8(3), 165–174.
- Sayed, H., Abdel-Fattah, M. A., & Kholief, S. (2018). Predicting potential banking customer churn using apache spark ML and MLlib packages: A comparative study. *International Journal of Advanced Computer Science and Applications*, 9(11).
https://www.researchgate.net/profile/Manal-Abdel-Fattah-2/publication/329427276_Predicting_Potential_Banking_Customer_Churn_using_Apache_Spark_ML_and_MLlib_Packages_A_Comparative_Study/links/5c08086f4585157ac1aaf58e/Predicting-Potential-Banking-Customer-Churn-using-Apache-Spark-ML-and-MLlib-Packages-A-Comparative-Study.pdf
- Senanayake, D., Muthugama, L., Mendis, L., & Madushanka, T. (2015). Customer Churn Prediction: A Cognitive Approach. *Internation Journal of Computer, Electrical, Automation, Control and Information Engineering*, 9(3).
https://www.researchgate.net/profile/Tiroshan-Madushanka/publication/284464825_Customer_Churn_Prediction_A_Cognitive_Approach/links/5653cc5108aefe619b1975ae/Customer-Churn-Prediction-A-Cognitive-Approach.pdf
- Tsai, C.-F., & Lu, Y.-H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, 36(10), 12547–12553.
- Umayaparvathi, V., & Iyakutti, K. (2012). Applications of data mining techniques in telecom churn prediction. *International Journal of Computer Applications*, 42(20), 5–9.

- Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55, 1–9.
- Van den Poel, D., & Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, 157(1), 196–217.
- Xia, G., & Jin, W. (2008). Model of customer churn prediction on support vector machine. *Systems Engineering-Theory & Practice*, 28(1), 71–77.
- Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3), 5445–5449.
- Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17(5–6), 375–381. <https://doi.org/10.1080/713827180>
- Allain, H. (2020). *Improving productivity and reducing costs of mobile app development with Flutter and Backend-as-a-Service* [Master's Thesis]. <https://aaltodoc.aalto.fi/handle/123456789/97522>
- Code, V. S. (2019). Visual studio code. *Recuperado El Octubre De*. http://mentorthis.s3.amazonaws.com/upload/files/2022/06/55yIf7PjAppzmBhYxk1L_06_8b52883468172443a3960cb347a3a4f5_file.pdf
- Dinh, D., & Wang, Z. (2020). *Modern front-end web development: How libraries and frameworks transform everything*. <https://www.theseus.fi/handle/10024/342325>
- Grotov, K., Titov, S., Sotnikov, V., Golubev, Y., & Bryksin, T. (2022). A large-scale comparison of Python code in Jupyter notebooks and scripts. *Proceedings of the 19th*

International Conference on Mining Software Repositories, 353–364.

<https://doi.org/10.1145/3524842.3528447>

Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of *Scikit-learn* Package in Python Programming Language. *Journal of Educational and Behavioral Statistics*, 44(3), 348–361. <https://doi.org/10.3102/1076998619832248>

Relan, K. (2019). Beginning with Flask. In K. Relan, *Building REST APIs with Flask* (pp. 1–26). Apress. https://doi.org/10.1007/978-1-4842-5022-8_1

Robinson, D. (2017). The incredible growth of Python. *Stack Overflow—Sep*, 6.

Appendix A: Prediction Page

```
<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="utf-8">
<meta content="width=device-width, initial-scale=1.0" name="viewport">
<title>CUSTOMER CHURN PREDICTION</title>
<meta content="" name="description">
<meta content="" name="keywords">
<!-- Favicons -->
<link href="{ {url_for('static', filename = 'img/fav2.png') }}" rel="icon">
<link href="{ {url_for('static', filename = 'img/fav2.png') }}" rel="apple-touch-
icon">
<!-- Google Fonts -->
<link
href="https://fonts.googleapis.com/css?family=Open+Sans:300,300i,400,400i,
600,600i,700,700i|Raleway:300,300i,400,400i,500,500i,600,600i,700,700i|Po
ppins:300,300i,400,400i,500,500i,600,600i,700,700i"
rel="stylesheet">
<!-- Vendor CSS Files -->
<link href="{ {url_for('static', filename = 'vendor/aos/aos.css') }}"
rel="stylesheet">
<link href="{ {url_for('static', filename
='vendor/bootstrap/css/bootstrap.min.css') }}" rel="stylesheet">
<link href="{ {url_for('static', filename = 'vendor/bootstrap-icons/bootstrap-
icons.css') }}" rel="stylesheet">
<link href="{ {url_for('static', filename
='vendor/boxicons/css/boxicons.min.css') }}" rel="stylesheet">
```

```

<link                href="{{url_for('static',                filename
='vendor/glightbox/css/glightbox.min.css')}}" rel="stylesheet">

<link href="{{ {{url_for('static', filename ='vendor/remixicon/remixicon.css')}}"
rel="stylesheet">

<link    href="{{ {{url_for('static',    filename    ='vendor/swiper/swiper-
bundle.min.css')}}" rel="stylesheet">

<!-- Template Main CSS File -->

<link    rel="stylesheet"    type="text/css"    href="{{ {{    url_for('static',
filename='css/index.css') }} }}">

</head>

<body>

<!-- ===== Header ===== -->

<header id="header" class="fixed-top d-flex align-items-center">

<div class="container d-flex align-items-center">

<h1                class="logo                me-auto"><a
href="#hero">SOMTELCHURN<span>.</span></a></h1>

<!-- <a href="index.html" class="logo me-auto"></a> -->

<nav id="navbar" class="navbar order-last order-lg-0">

<ul>

<li><a class="nav-link scrollto active" href="#hero">Home</a></li>

<li><a class="nav-link scrollto" href="#about">About</a></li>

<li><a class="nav-link scrollto" href="#visualization">Visualization</a></li>

<li><a class="nav-link scrollto" href="#project">Project</a></li>

{% if 'username' in session %}

<li><a class="nav-link scrollto" href="#statistics">System Statistics</a></li>

{% else %}

{% endif %}

```

```

<li><a class="nav-link scrollto" href="#team">Team</a></li>

{% if 'username' in session %}<li class=""> <a class="nav-link text-primary
fw-bold" href="#">Welcome, {{ session['username'][:15] }}..!</a>

</li><li class=""><a class="nav-link text-danger fw-bold" href="{{
url_for('logout') }}"><i>Logout</i></a></li>

{% else %}

<li class="nav-item"><a class="nav-link text-primary scrollto fw-bold"
href="{{ url_for('login') }}">Login</a></li>

{% endif %}

</ul>

<i class="bi bi-list mobile-nav-toggle"></i></nav><!-- .navbar --><!-- <a
href="#about" class="get-started-btn scrollto">Get Started</a -->

</div></header><!-- End Header --><!-- ===== Hero Section ===== --
>

<section id="hero" class="d-flex align-items-center">

<div class="container" data-aos="zoom-out" data-aos-delay="100">

<div class="row"><div class="col-xl-6"><h1>Customer churn prediction
based on telecommunication Somalia</h1><h2>The telecommunication
companies in Mogadishu Somalia has challenges about keeping and the loyalty
of their customers to predict either leaving or not.</h2><a href="#about"
class="btn-get-started scrollto">Get Started</a></div></div></div>

</section><!-- End Hero --><main id="main"><!-- ===== Project Section
===== --><section id="project" class="project"><div class="container"
data-aos="fade-up"><div class="section-
title"><h2>Project</h2><p>Customer churn prediction based on
telecommunication somalia. </p></div><div class="row" data-aos="fade-up"
data-aos-delay="100"><div class="col-lg-12">

<form id="predictionForm" action="{{ url_for('predict') }}" method="POST"
role="form" class="project-form">

<div class="row">

<div class="col-lg-6 form-group">

```



```

<label for="" class="form-label mb-2">Senior Citizen</label>

<select name="SeniorCitizen" class="form-control" id="SeniorCitizen"
required>

<option value="1" {% if SeniorCitizen==1 %}selected{% endif
%}>1</option>

<option value="0" {% if SeniorCitizen==0 %}selected{% endif
%}>0</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Monthly Charges</label>

<input type="text" name="MonthlyCharges" class="form-control"
id="MonthlyCharges"

value="{{MonthlyCharges}}" step="any" placeholder="Monthly Charges"
required

oninput="validateInput(this)">

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Total Charges</label>

<input type="text" name="TotalCharges" class="form-control"
id="TotalCharges" value="{{TotalCharges}}"

step="any" placeholder="Total Charges" required
oninput="validateInput(this)">

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Gender</label>

<select name="Gender" class="form-control" id="Gender" required>

<option value="Female" {% if Gender=='Female' %}selected{% endif
%}>Female</option>

```

```

<option value="Male" {% if Gender=='Male' %}selected{% endif
%}>Male</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Partner</label>

<select name="Partner" class="form-control" id="Partner" required>

<option value="Yes" {% if Partner=='Yes' %}selected{% endif
%}>Yes</option>

<option value="No" {% if Partner=='No' %}selected{% endif
%}>No</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Dependents</label>

<select name="Dependents" class="form-control" id="Dependents" required>

<option value="Yes" {% if Dependents=='Yes' %}selected{% endif
%}>Yes</option>

<option value="No" {% if Dependents=='No' %}selected{% endif
%}>No</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Phone Service</label>

<select name="PhoneService" class="form-control" id="PhoneService"
required>

<option value="Yes" {% if PhoneService=='Yes' %}selected{% endif
%}>Yes</option>

```

```

<option value="No" {% if PhoneService=='No' %}selected{% endif
%}>No</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Multiple Lines</label>

<select name="MultipleLines" class="form-control" id="MultipleLines"
required>

<option value="Yes" {% if MultipleLines=='Yes' %}selected{% endif
%}>Yes</option>

<option value="No" {% if MultipleLines=='No' %}selected{% endif
%}>No</option>

<option value="No phone service" {% if MultipleLines=='No phone service'
%}selected{% endif %}>No
phone service</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Internet Service</label>

<select name="InternetService" class="form-control" id="InternetService"
required>

<option value="DSL" {% if InternetService=='DSL' %}selected{% endif
%}>DSL</option>

<option value="Fiber optic" {% if InternetService=='Fiber optic'
%}selected{% endif %}>Fiber optic

</option>

<option value="No" {% if InternetService=='No' %}selected{% endif
%}>No</option>

</select>

```

```

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Online Security</label>

<select name="OnlineSecurity" class="form-control" id="OnlineSecurity"
required>

<option value="Yes" {% if OnlineSecurity=='Yes' %}selected{% endif
%}>Yes</option>

<option value="No" {% if OnlineSecurity=='No' %}selected{% endif
%}>No</option>

<option value="No internet service" {% if OnlineSecurity=='No internet
service' %}selected{% endif
%}>No internet service</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Online Backup</label>

<select name="OnlineBackup" class="form-control" id="OnlineBackup"
required>

<option value="Yes" {% if OnlineBackup=='Yes' %}selected{% endif
%}>Yes</option>

<option value="No" {% if OnlineBackup=='No' %}selected{% endif
%}>No</option>

<option value="No internet service" {% if OnlineBackup=='No internet
service' %}selected{% endif %}>
No internet service</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Device Protection</label>

```

```

<select name="DeviceProtection" class="form-control" id="DeviceProtection"
required>

<option value="Yes" {% if DeviceProtection=='Yes' %}selected{% endif
%}>Yes</option>

<option value="No" {% if DeviceProtection=='No' %}selected{% endif
%}>No</option>

<option value="No internet service" {% if DeviceProtection=='No internet
service' %}selected{% endif
%}>No internet service</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Tech Support</label>

<select name="TechSupport" class="form-control" id="TechSupport"
required>

<option value="Yes" {% if TechSupport=='Yes' %}selected{% endif
%}>Yes</option>

<option value="No" {% if TechSupport=='No' %}selected{% endif
%}>No</option>

<option value="No internet service" {% if TechSupport=='No internet service'
%}selected{% endif %}>
No internet service</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Streaming TV</label>

<select name="StreamingTV" class="form-control" id="StreamingTV"
required>

```

```

<option value="Yes" {% if StreamingTV=='Yes' %}selected{% endif
%}>Yes</option>

<option value="No" {% if StreamingTV=='No' %}selected{% endif
%}>No</option>

<option value="No internet service" {% if StreamingTV=='No internet service'
%}selected{% endif %}>
No internet service</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Streaming Movies</label>

<select name="StreamingMovies" class="form-control"
id="StreamingMovies" required>

<option value="Yes" {% if StreamingMovies=='Yes' %}selected{% endif
%}>Yes</option>

<option value="No" {% if StreamingMovies=='No' %}selected{% endif
%}>No</option>

<option value="No internet service" {% if StreamingMovies=='No internet
service' %}selected{% endif
%}>No internet service</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Contract</label>

<select name="Contract" class="form-control" id="Contract" required>

<option value="Month-to-month" {% if Contract=='Month-to-month'
%}selected{% endif %}>Month-to-month

</option>

```

```

<option value="One year" {% if Contract=='One year' %}selected{% endif
%}>One year</option>

<option value="Two year" {% if Contract=='Two year' %}selected{% endif
%}>Two year</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Paperless Billing</label>

<select name="PaperlessBilling" class="form-control" id="PaperlessBilling"
required>

<option value="Yes" {% if PaperlessBilling=='Yes' %}selected{% endif
%}>Yes</option>

<option value="No" {% if PaperlessBilling=='No' %}selected{% endif
%}>No</option>

</select>

</div>

<div class="col-lg-6 form-group">

<label for="" class="form-label mb-2">Payment Method</label>

<select name="PaymentMethod" class="form-control" id="PaymentMethod"
required>

<option value="Electronic check" {% if PaymentMethod=='Electronic check'
%}selected{% endif %}>
Electronic check</option>

<option value="Mailed check" {% if PaymentMethod=='Mailed check'
%}selected{% endif %}>Mailed check

</option>

<option value="Credit card (automatic)" {% if PaymentMethod=='Credit card
(automatic)' %}selected{%
endif %}>Credit card (automatic)</option>

```

```

<option value="Bank transfer (automatic)" {% if PaymentMethod=='Bank
transfer (automatic)'
%}selected{% endif %}>Bank transfer (automatic)</option>
</select>
</div>
<div class="col-lg-6 form-group">
<label for="" class="form-label mb-2">Number of Months</label>
<input type="text" name="tenure" class="form-control" id="tenure"
value="{{tenure}}"
placeholder="Number of Month" required oninput="validateUserInput(this)">
</div>
</div>
<div class="d-flex justify-content-between">
<div class="text-start"><button type="submit">Predict</button></div>
</div>

<div class="alert alert-primary mt-3" id="outputDisplay" role="alert"
style="display: none;">
<div id="loadingSpinner" style="display: none;">
<div class="spinner-border text-primary" role="status">
<span class="visually-hidden">Loading...</span>
</div>
</div>
<span id="outputMessage"></span>
</div>
<!-- <div class="alert alert-primary" role="alert">
{{output2}}

```


</div> -->

</form>

</div>

</section>

<!-- End Project Section --> <!-- ===== System Statistics Section =====
-->

{% if 'username' in session %}

<section id="statistics" class="statistics">

<div class="container footer-top" data-aos="fade-up">

<div class="section-title"><h2>System Statistics </h2><p>summarize
system.</p></div>

<div class="row mt-1"><div class="col-12"><div class="table-responsive">

<table class="table table-bordered" id="statistics_table"><thead class="table-
dark"><tr><th scope="col">#</th><th scope="col">Gender</th><th
scope="col">Ph.service</th><th scope="col">I.service</th><th
scope="col">O.security</th><th scope="col">O.backup</th><th
scope="col">S.tv</th>

<th scope="col">S.movies</th><th scope="col">Contract</th><th
scope="col">P.method</th><th scope="col">M.charges</th><th
scope="col">T.charges</th><th scope="col">Num.months</th><th
scope="col">P.type</th></tr></thead><tbody></tbody></table></div>

<div class="d-flex justify-content-between ">

<div class="text-start"><button type="button"
onclick="downloadExcel()">Export</button></div></div></div></div>

<div class="mt-4 row counters"><div class="col-lg-4 col-6 text-center">

<span class="purecounter counter-total" data-purecounter-start="0" data-
purecounter-duration="1"><p>Total Entries</p>

```

</div><div class="col-lg-4 col-6 text-center"> <span class="purecounter
counter-churn" data-purecounter-start="0" data-purecounter-
duration="1"></span><p>Churn</p> </div><div class="col-lg-4 col-6 text-
center"> <span class="purecounter counter-not-churn" data-purecounter-
start="0" data-purecounter-duration="1"></span><p>Not Churn</p></div>
</div></div> </section>

{% else %}

{% endif %} <!-- End System Statistics Section --> </main>

<!-- Vendor JS Files -->

<script src="{ { url_for('static', filename
='vendor/purecounter/purecounter_vanilla.js') } }"></script>

<script src="{ { url_for('static', filename ='vendor/aos/aos.js') } }"></script>

<script src="{ { url_for('static', filename
='vendor/bootstrap/js/bootstrap.bundle.min.js') } }"></script>

<script src="{ { url_for('static', filename
='vendor/glightbox/js/glightbox.min.js') } }"></script>

<script src="{ { url_for('static', filename
='vendor/isotope-layout/isotope.pkgd.min.js') } }"></script>

<script src="{ { url_for('static', filename
='vendor/swiper/swiper-bundle.min.js') } }"></script>

<script src="{ { url_for('static', filename ='vendor/php-email-form/validate.js')
} }"></script>

<!-- Template Main JS File -->

<script src="{ { url_for('static', filename ='js/main.js') } }"></script>

<script src="{ { url_for('static', filename ='js/valid.js') } }"></script>

<script
src="https://ajax.googleapis.com/ajax/libs/jquery/3.5.1/jquery.min.js"></script
>

<script>

$(document).ready(function() {

```

```

setInterval( getStatistics, 5000);

$('#predictionForm').on('submit', function (e) {
e.preventDefault(); $('#outputDisplay').show();
$('#outputMessage').hide(); $('#loadingSpinner').show();
$.ajax({ url: $(this).attr('action'),
type: 'POST',
data: $(this).serialize(),
success: function (response) {
$('#outputMessage').text(response.data);
setTimeout(function () {
$('#loadingSpinner').hide();
$('#outputMessage').text(response.message).show();
getStatistics()
}, 2000); },
error: function () {
setTimeout(function () {
$('#loadingSpinner').hide();
$('#outputMessage').text('Error processing your request').show();
$('#outputDisplay').show();
}, 1000); }}));});
function getStatistics(){
$.ajax({
url: '/api/statistics',
type: 'GET',
success: function(response) {
console.log(response); // See what data is actually being returned

```

```

$('.counter-total').attr('data-purecounter-end', response.statistics.total_data || 0);
$('.counter-churn').attr('data-purecounter-end', response.statistics.total_churn || 0);
$('.counter-not-churn').attr('data-purecounter-end', response.statistics.total_not_churn || 0)

// Update table
if (response.tableData && response.tableData.length > 0) {
var tableRows = response.tableData.map(function(row) {
return `<tr>
<td>${row.id}</td><td>${row.gender}</td><td>${row.phone_service}</td>
<td>${row.internet_service}</td>          <td>${row.online_security}</td>
<td>${row.online_backup}</td><td>${row.streaming_tv}</td><td>${row.streaming_movies}</td><td>${row.contract}</td><td>${row.payment_method}</td><td>${row.monthly_charges}</td><td>${row.total_charges}</td><td>${row.number_of_months}</td><td>${row.predictedType}</td></tr>`;
}).join('');
'#statistics_table tbody').html(tableRows); }
else {
$('#statistics_table tbody').html('<tr><td colspan="4">No data available</td></tr>');}
new PureCounter();},
error: function() {console.error('Failed to fetch statistics');
$('#statistics_table tbody').html('<tr><td colspan="4">Error loading data</td></tr>');
}});} });
function downloadExcel() {
window.location.href = '/download-excel';}
</script></body></html>

```

Appendix B: SERVER

```
from flask import Flask, render_template, request, url_for, redirect, session,
jsonify

import re

from flask import send_file

import pandas as pd

import pickle

from connection import *

import sqlite3

app = Flask(__name__)

app.secret_key = "zxsdasdasdasdsd"

app.teardown_appcontext(close_db)

@app.before_request

def initialize():

    create_table()

    # Load your data
```

```

df_1          =          pd.read_csv("../Project/Telco-Customer-Churn-
Prediction/first_telc.csv")

@app.route("/")

def index():

    if 'username' in session:

        return render_template('index.html', query="")

    return redirect(url_for('login'))

@app.route('/login', methods=['GET', 'POST'])

def login():

    if request.method == 'POST':

        email = request.form['email']

        password = request.form['password']

        conn = get_db()

        cursor = conn.cursor()

        cursor.execute("SELECT * FROM users WHERE email = ? AND
password = ?", (email, password))

        user = cursor.fetchone()

```

```

print()

if user:

    session['username'] = user['name']

    session['email'] = user['email']

    return redirect(url_for('index'))

else:

    return render_template('login.html', error='Invalid email or password',
email=email, password=password)

    return render_template('login.html', email="", password=")

@app.route('/signup', methods=['GET', 'POST'])

def signup():

    if request.method == 'POST':

        name = request.form['name']

        email = request.form['email']

        password = request.form['password']

        if not is_valid_name(name) or not is_gmail_address(email) or not
is_strong_password(password):

```

```

        return render_template('register.html', error='Please ensure all fields are
valid.', name=name, email=email, password=password)

    conn = get_db()

    cursor = conn.cursor()

    cursor.execute("SELECT * FROM users WHERE email = ?", (email,))

    user = cursor.fetchone()

    if user:

        return render_template('register.html', error='Email already exists.',
name=name, email=email, password=password)

    cursor.execute("INSERT INTO users (name, email, password) VALUES
(?, ?, ?)", (name, email, password))

    conn.commit()

    return redirect(url_for('login'))

    return render_template('register.html', name="", email="", password=")

def is_valid_name(name):

    # Name should contain only alphabets and spaces

    return bool(re.match("^[a-zA-Z ]+$", name))

```



```

def is_gmail_address(email):

    # Check if the email ends with @gmail.com

    return email.endswith("@gmail.com")

def is_strong_password(password):

    # Check for minimum length of 6, include upper, lower and digits

    if len(password) < 6:

        return False

    if not re.search("[a-z]", password):

        return False

    if not re.search("[A-Z]", password):

        return False

    if not re.search("[0-9]", password):

        return False

    return True

@app.route('/logout')

def logout():

    session.pop('username', None)

```

```

return redirect(url_for('login'))

@app.route('/predict', methods=['GET', 'POST'])

def predict():

    if request.method == 'POST':

        if 'username' not in session:

            return redirect(url_for('login'))

        """
        SeniorCitizen    MonthlyCharges    TotalCharges    gend
er    Partner    Dependents    PhoneService    MultipleLines    I
nternetService    OnlineSecurity    OnlineBackup    DeviceProtectio
n    TechSupport    StreamingTV    StreamingMovies    Contract
PaperlessBilling    PaymentMethod    tenure    """

        inputQuery1 = request.form['SeniorCitizen']

        inputQuery2 = request.form['MonthlyCharges']

        inputQuery3 = request.form['TotalCharges']

        inputQuery4 = request.form['Gender']

        inputQuery5 = request.form['Partner']

        inputQuery6 = request.form['Dependents']

        inputQuery7 = request.form['PhoneService']

```

inputQuery8 = request.form['MultipleLines']

inputQuery9 = request.form['InternetService']

inputQuery10 = request.form['OnlineSecurity']

inputQuery11 = request.form['OnlineBackup']

inputQuery12 = request.form['DeviceProtection']

inputQuery13 = request.form['TechSupport']

inputQuery14 = request.form['StreamingTV']

inputQuery15 = request.form['StreamingMovies']

inputQuery16 = request.form['Contract']

inputQuery17 = request.form['PaperlessBilling']

inputQuery18 = request.form['PaymentMethod']

inputQuery19 = request.form['tenure']

data = [[inputQuery1, inputQuery2, inputQuery3, inputQuery4, inputQuery5,
inputQuery6, inputQuery7,

inputQuery8, inputQuery9, inputQuery10, inputQuery11,
inputQuery12, inputQuery13, inputQuery14,

```
inputQuery15, inputQuery16, inputQuery17, inputQuery18,  
inputQuery19]]
```

```
new_df = pd.DataFrame(data, columns = ['SeniorCitizen',  
'MonthlyCharges', 'TotalCharges', 'gender',  
  
                                     'Partner', 'Dependents', 'PhoneService',  
  
                                     'MultipleLines', 'InternetService',  
  
                                     'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',  
  
                                     'TechSupport',  
  
                                     'StreamingTV', 'StreamingMovies', 'Contract',  
  
                                     'PaperlessBilling',  
  
                                     'PaymentMethod', 'tenure'])
```

```
df_2 = pd.concat([df_1, new_df], ignore_index = True)
```

```
# Group the tenure in bins of 12 months
```

```
labels = ["{0} - {1}".format(i, i + 11) for i in range(1, 72, 12)]
```

```
df_2['tenure_group'] = pd.cut(df_2.tenure.astype(int), range(1, 80, 12),  
right=False, labels=labels)
```

```
#drop column tenure
```

```
df_2.drop(columns= ['tenure'], axis=1, inplace=True)
```

```

df_2.SeniorCitizen = pd.to_numeric(df_2.SeniorCitizen, errors='coerce')

df_2.MonthlyCharges = pd.to_numeric(df_2.MonthlyCharges,
errors='coerce')

df_2.TotalCharges = pd.to_numeric(df_2.TotalCharges, errors='coerce')

new_df__dummies = pd.get_dummies(df_2[['gender',
'SeniorCitizen','MonthlyCharges', 'TotalCharges', 'Partner', 'Dependents',
'PhoneService',

'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup',

'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies',

'Contract', 'PaperlessBilling', 'PaymentMethod','tenure_group']])

# Load model

model = pickle.load(open("../Project/Telco-Customer-Churn-
Prediction/modeldt.sav", "rb"))

# model = pickle.load(open("../Project/Telco-Customer-Churn-
Prediction/modellog.sav", "rb"))

# model = pickle.load(open("../Project/Telco-Customer-Churn-
Prediction/modelrf.sav", "rb"))

```

```

        # model = pickle.load(open("../Project/Telco-Customer-Churn-
Prediction/modelsvm.sav", "rb"))

    single = model.predict(new_df__dummies.tail(1))

    if single==1:

        try:

            InsertPredictionsData(inputQuery4,inputQuery5,inputQuery6,inp
utQuery7,inputQuery8,inputQuery9,inputQuery10,inputQuery11,inputQuery1
2,inputQuery13,inputQuery14,inputQuery15,inputQuery16,inputQuery17,inp
utQuery18,inputQuery2,inputQuery3,inputQuery19, "Churn")

            result = "This customer is churned!!."

            return jsonify(message=result)

        except Exception as e:

            return jsonify(message=str(e)), 500

    else:

        try:

            InsertPredictionsData(inputQuery4,inputQuery5,inputQuery6,input
Query7,inputQuery8,inputQuery9,inputQuery10,inputQuery11,inputQuery12

```

```
,inputQuery13,inputQuery14,inputQuery15,inputQuery16,inputQuery17,inputQuery18,inputQuery2,inputQuery3,inputQuery19, "Not churn")
```

```
    result = "This customer is not churn."
```

```
    return jsonify(message=result, data=data)
```

```
except Exception as e:
```

```
    return jsonify(message=str(e)), 500
```

```
elif request.method == 'GET':
```

```
    return render_template('index.html')
```

```
return 'Bad Request!', 400
```

```
@app.route('/api/statistics')
```

```
def get_statistics():
```

```
    data_table = SummarizePrediction("table")
```

```
    data_churn = SummarizePrediction("churn")
```

```
    data_not_churn = SummarizePrediction("not churn")
```

```
    table_data = [dict(row) for row in data_table] if data_table else []
```

```
    stats_data = {
```

```
        'total_data': len(data_table) if data_table else 0,
```

```

        'total_churn': len(data_churn) if data_churn else 0,

        'total_not_churn': len(data_not_churn) if data_not_churn else 0

    }    return jsonify({

        'statistics': stats_data,

        'tableData': table_data

    })

@app.route('/download-excel')

def download_excel():

    conn = get_db()

    cursor = conn.cursor()

    cursor.execute("SELECT * FROM predictions")

    data = cursor.fetchall()

    df = pd.DataFrame(data, columns=[column[0] for column in
cursor.description])

    excel_path = 'C:/Users/Zakar/Desktop/THESIS/Project/Telco-Customer-
Churn-Prediction/predictions.xlsx'

    df.to_excel(excel_path, index=False, engine='openpyxl')

```



```

        return send_file(excel_path, as_attachment=True,
download_name='Predictions.xlsx')

@app.errorhandler(404)

def page_not_found(e):

    return render_template('404.html'), 404

if __name__ == '__main__':

    with app.app_context():

        create_table()

        create_predictinTable()

    app.run(debug=True)

```

END