2023

Customer churn prediction tool using deep learning: a case of an ecommerce business operating in Kenya.

Narina, Papetua
School of Computing and Engineering Sciences
Strathmore University

Customer Churn Prediction Tool Using Deep Learning: A Case of an eCommerce Business Operating in Kenya

By

Papetua Narina

145033

Master of Science in Information Technology at Strathmore University

2023

Customer Churn Prediction Tool Using Deep Learning: A Case of an eCommerce Business Operating in Kenya

Papetua Narina

Submitted in partial fulfilment for the award of the Degree of
Master of Science in Information Technology at Strathmore University

School of Computing and Engineering Sciences

Strathmore University

Nairobi, Kenya

July 2023.

Declaration and Approval

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due. reference is made in the thesis itself.

Student's Name: **Papetua Narina**

Signed: _____ Date: _____28/05/2023_____

Approval

This thesis of Papetua Narina was reviewed and approved by the following:

Dr. Allan Omondi
School of Computing & Engineering Sciences,
Strathmore University

Dr. Vincent Omwenga
School of Computing & Engineering Sciences,
Strathmore University

Dr. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Abstract

The problem of customer churn poses significant challenges for businesses, particularly in the e-commerce sector. With a high level of competition and low customer retention rates, businesses face the risk of losing customers and experiencing a decline in revenue. Previous research in customer churn prediction exhibited limitations in terms of accuracy and the lack of consumer-facing tools for predictions. Moreover, existing studies were often conducted before the COVID-19 era, failing to capture the impact of recent changes in consumer behavior and market dynamics. This research aimed to develop an effective customer churn prediction model for e-commerce businesses in Kenya. The study aimed to address the limitations of existing research by leveraging advanced machine learning techniques and considering the specific challenges faced by businesses in the post-COVID-19 era. To achieve the objective, an agile software development methodology was adopted. This approach allowed for continuous iterations and refinements during the model development process. Customer dataset was obtained from Kaggle an online platform for sharing datasets. The dataset included customer demographic information, transaction history, and customer engagement metrics. The data was carefully pre-processed to handle missing values, outliers, and ensure data quality. The multilayer perceptron model (MLP), a powerful deep learning algorithm, was employed to train the customer churn prediction model. The dataset was split into training and testing sets, with an 80-20 ratio, to assess the model's performance. The results of the study indicated that the developed customer churn prediction model achieved high accuracy, with precision, recall, and F1 scores of 94%. This demonstrated the model's effectiveness in identifying potential churners and enabling businesses to take proactive measures for customer retention. The findings of this study have significant implications for e-commerce businesses, providing them with a valuable tool to predict customer churn and implement targeted retention strategies. By leveraging the power of advanced machine learning techniques, businesses could enhance customer satisfaction, optimize resource allocation, and drive sustainable growth in a highly competitive market.

Keywords: ***Customer Churn, Deep Learning, Multi-Layer perceptron, E-commerce, Big Data.***

Table of Contents

Declaration and Approval ii

Abstract iii

Table of Contents iv

List of Figures ix

List of Equations xi

Abbreviations / Acronyms xii

Acknowledgments xiii

Dedication xiv

Definition of Terms xv

Chapter 1: Introduction 1

1.1 Background 1

1.2 Problem Statement 3

1.3 Aim 4

1.4 Specific Objectives 4

1.5 Research Questions 4

1.6 Justification 4

1.8 Scope and Limitations 5

Chapter 2: Literature Review 7

2.1 Introduction 7

2.2 Theoretical Literature 7

2.2.1 Dissonance Theory 8

2.2.2 The Contrast Theory 8

2.2.3 The Expectancy Disconfirmation Paradigm 9

2.3 Empirical Literature 10

2.3.1 Customer Churn 10
2.3.3 Predicting Customer Churn 11
2.4 Factors that Lead to Customer Churn 13
2.4.1 Price Factor 13
2.4.2 Product Factor 14
2.4.3 Customer Factor 14
2.4.4 Business Factor 15
2.4.5 Service Factor 15
2.5 Models and Frameworks 16
2.5.1 Models 16
2.4.2 Frameworks 19
2.6 Architecture and Design 22
2.6.1 Big Data Architecture 22
2.6.2 Ensemble Architecture 23
2.6.3 Hybrid Firefly Architecture 24
2.7 Algorithms 27
2.7.1 Machine Learning 27
2.7.2 Random Forests 27
2.7.3 K-Means 29
2.7.4 Decision Tree 30
2.7 Gaps in Existing Systems 31
2.8 Conceptual Model 32
2.9 Conclusion 33
Chapter 3: Research Methodology 34
3.1 Introduction 34

3.2 Research Design and Philosophy 34
3.2.1 Research Design 34
3.2.2 Research Philosophy 35
3.3 Population and Sampling 37
3.3.1 Population 38
3.3.2 Sampling Size 38
3.4 Data Collection Method and Analysis 38
3.5 Research Quality and Reliability 39
3.6 System Development Methodology 40
3.6.1 Plan 41
3.6.2 Design 41
3.6.3 Develop 41
3.6.4 Test 42
3.6.5 Deploy 42
3.6.6 Continuous Iteration 42
3.7 Utilization and Dissemination of Research Results 42
3.8 Ethical Considerations/Issues 43
3.9 Conclusion 43
Chapter 4: systems analysis and design 44
4.1 Introduction 44
4.2 Requirement Specifications 44
4.2.1 Functional Requirements 45
4.2.2 Non-Functional Requirements 45
4.3 System Architecture 46
4.4 Diagrammatic Representation of the System 47

4.4.1 Use Case 48
4.4.2 Sequence Diagram 50
4.4.3 Class Diagram 51
4.4.4 Database Schema 52
4.5 Wireframes 52
4.5.1 Home Page 52
4.5.2 Register Page 53
4.5.3 Login Page 54
4.5.4 Upload Page 55
4.5.5 Customer Churn Profile Page 56
4.6 Conclusion 57
Chapter 5: System Testing and Implementation 59
5.1 Introduction 59
5.2 System Implementation 59
5.2.1 Loading Dataset 59
5.2.2 Data Preprocessing 62
5.2.3 Feature Selection 63
5.2.4 Model Training 64
5.2.5 Model Validation 68
5.3 Customer Churn Prediction Interface 68
5.4 System Testing 72
5.5 Testing Model Accuracy 74
5.4 Model Validation/Deployment 74
5.5 Conclusion 74
Chapter 6: Discussion 76

6.1 Investigating the Factors That Lead to Customer Churn in The E-Commerce Industry 76
6.2 Existing Models and Algorithms Used for Customer Churn Prediction 76
6.3 Customer Churn Prediction Tool in The E-Commerce Industry Using Deep Learning Techniques 78
6.4 To test the developed tool. 78
Chapter 7: Conclusion and Recommendation 80
7.1 Conclusion 80
7.2 Recommendations 82
7.3 Limitations of the Study 82
7.4 Future Work 82
References 83
Appendices 90
Appendix A: Summary of Literature Review 90
Appendix B: Project Gantt Chart 92
Appendix C: Dataset 93
Appendix D: Data Analysis 94
Appendix E: Consent Form 95
Appendix F: Budget 96
Appendix G: Ethical Review 97
Appendix H: Similarity Report 98

List of Figures

Figure 2.1 Representation of a Neural Network Neural Network 17
Figure 2.2TensorFlow Symbolic Computational Graph of a Feed Forward Neural Network 20
Figure 2.3 Five-Step Life Cycle of Deep Learning Model in Keras 21
Figure 2.4 Big Data Architecture 23
Figure 2.5 Ensemble Architecture 24
Figure 2.6 Hybrid Firefly Architecture 25
Figure 2.7 Random Forest Algorithm. 27
Figure 2.8 3D Clusters of K-means algorithm 28
Figure 2.9 Decision Tree Algorithm 29
Figure 2.10: Conceptual Model 31
Figure 3.1: Research Onion 34
Figure 3.2 Agile Methodology 39
Figure 4.1 System Architecture 42
Figure 4.2 Use Case Diagram. 43
Figure 4.3 Sequence Diagram. 45
Figure 4.4 Class Diagram. 46
Figure 4.4 Database Schema. 47
Figure 4.7 Home Page Wireframe 48
Figure 4.8 Register Page Wireframe 48
Figure 4.6 Login Page Wireframe 49
Figure 4.6 Upload Page Wireframe 50
Figure 4.6 Churn Profile Page Wireframe 51
Figure 5.1 Loading Dataset 53
Figure 5.2 Cleaning Data 54
Figure 5.3 Cleaning Dataset 54
Figure 5.4 Data Preprocessing 55
Figure 5.6 Model Training 55
Figure 5.7 Model Training 55
Figure 5.8 Model Training 56
Figure 5.9 Model Validation 56

Figure 5.10 Login Interface 57
Figure 5.11 Registration Interface 57
Figure 5.12 Dashboard 58
Figure 5.13 Prediction Interface 59
Figure 5.14 Prediction History 59
Figure 5.15 Accuracy Score. 60

List of Equations

Equation 2.1 SVM Formula 15

Abbreviations / Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Program Interface
B2C	Business to Consumer
CNN	Convolution Neural Network
DL	Deep Learning
ML	Machine Learning
SVM	Support Vector Machine

Acknowledgments

I must register my gratitude to Professor Ismail Ateya. His guidance, patience, and insights helped me think through the paper and get relevant research materials. Please accept my gratitude and sincere appreciation, and to Dr. Allan Omondi, my supervisor, for his unwavering support.

Dedication

I dedicate this work to my Parents. Thank you for your unwavering support and belief in me and the facilitation to pursue my dreams. To my friends, Annette, Elijah and Wayne, thank you for holding my hand and encouraging me when all seemed bleak.

Definition of Terms.

Business To Customer (B2C)	B2C is a transaction between a company and an individual who serves as the end customer (Heaslip, 2022).
Customer Relationship Management (CRM)	CRM is a set of practices, strategies, and technologies businesses use to manage and analyze customer interactions and data throughout the customer lifecycle (Chai et al., 2020).
Customer	A customer is an individual or business that purchases another company's goods or services (Kenton, 2021).
Customer Churn	Customer churn is the percentage of customers that stopped using a company's product or service during a specific time frame (Swetha Amaresan, 2018).
Partial Churners	These are customers who haven't purchased in three months.

Chapter 1: Introduction

1.1 Background

Loyal customers are essential in improving business performance and can boost an enterprise's core competitiveness. With herd mentality, loyal customers can help enterprises reduce the cost of publicity and negotiation and attract more new customers, lowering customer development costs and increasing the opportunities and time for enterprises to obtain basic profits (Zhao et al., 2021). They can increase the chance and time for businesses to earn basic profits, assist enterprises in earning premium income, consolidate the market position, reduce market risks, and raise entry barriers for other companies.

Customer churn is a critical issue frequently linked to the business's current cycle. When a company is in the development stage of its life cycle, deals grow exponentially, and the number of new clients far outnumbers the number of churners (Kriti, 2019). On the other hand, organizations in a mature life cycle place a premium on reducing customer churn. The primary causes of customer churn are classified as either accidental or intentional. Accidental churn occurs when conditions change, preventing customers from using services in the future, such as financial conditions that make benefits unreasonably expensive for the client. Intentional churn occurs when customers switch to another organization that provides comparable services, such as better ideas from competitors, more developed services, and a lower cost for a similar product or service.

Many businesses concentrate on acquiring new customers while ignoring the need to retain existing customers and increase their consumption potential, e-commerce being no exception. Reichheld and Sasser (2014) discovered that the longer a company's business relationship with its customers lasts the more profits it will make from its existing customers. The net present value of customers in the business environment increases by 25% to 95% for every 5% increase

in customer retention rate (Reichheld & Sasser, 2014). According to Jones and Sasser (2014) research, when an enterprise's customer churn rate falls by 5%, the enterprise's average profit rate rises by 25%-85%. As a result, the practical significance of customer churn prediction is that it will benefit businesses financially.

For starters, loyal customers have a higher retention rate than new customers, and the likelihood of competitive marketing activities is lower. Additionally, because the enterprise knows the preferences of existing customers, the cost of providing services is lower. Secondly, churned customers may refer other customers in their social network to competitors, whereas loyal customers will bring in more new customers. Thirdly, customer churn results in missed cross-selling and up-selling opportunities, resulting in a decrease in profits. Predicting customer churn behavior, analyzing the root causes of customer churn, identifying the links that need to be improved in the operation and management process, regaining churned customers, and establishing a stronger customer relationship have all become strategic priorities for businesses in the e-commerce industry.

While defection rates are a good predictor of profit swings, they do more than show where profits are going. They also direct managers' attention to the reasons customers are leaving. Companies cannot keep customers captive, so the only way to keep them is to outperform the competition consistently. Companies that solicit feedback from defecting customers can identify and strengthen the weaknesses that matter before profits begin to dwindle. Defection analysis is thus a guide that assists businesses in managing continuous improvement.

This study presents a novel way of predicting customer churn using deep learning techniques for e-commerce businesses. This will allow e-commerce companies to identify the section of customers likely to switch to competitors. Along with the model, the study developed a consumer-facing tool for businesses to utilize for profit maximization giving them access to tools for customer retention. Reducing defections in half will more than double the average company's growth rate.

1.2 Problem Statement

Customer churn prediction is a critical aspect that businesses need to address, particularly in the rapidly growing and highly competitive e-commerce industry. With the rise of technology and the increasing number of players in the market, customer retention has become a significant challenge. In the context of Kenya's e-commerce sector, where multiple major players exist, it has been observed that the customer retention rate is as low as 4.6% (Olumide, 2019), highlighting the need for effective strategies to mitigate customer churn.

The evolving landscape of e-commerce, facilitated by the Internet and digital transactions, has empowered consumers with flexibility, convenience, and a wide range of choices. As a result, the e-commerce market in Kenya is expected to grow substantially, reaching a value of KES 70-120 billion in the medium term and KES 400 billion in the long term (Gachenge, 2020). However, the cost of acquiring new customers is significantly higher compared to retaining existing customers who are at risk of churning. Understanding the true cost of losing a customer is crucial for businesses to make informed decisions about investments aimed at customer retention.

Traditional accounting systems often focus on current period costs and revenues, overlooking the expected cash flows over a customer's lifetime. However, it has been observed across industries that customers who stay with a company for longer periods generate more profits over time if they receive satisfactory service. The longer a company can retain a customer, the greater the potential for financial gains. Therefore, accurately predicting customer churn is vital for the survival and success of businesses in the e-commerce industry. By identifying potential churn customers in advance, companies can take proactive measures to mitigate the risks of losing customers to competitors.

In light of these challenges, the development of a customer churn prediction tool becomes essential. Such a tool would enable businesses to forecast and anticipate customer churn, allowing them to implement targeted strategies to retain customers. By leveraging machine

learning techniques, businesses can gain valuable insights into customer behavior, identify churn indicators, and take appropriate actions to minimize churn rates. The customer churn prediction tool will be a valuable asset for e-commerce businesses, providing them with a competitive advantage and enabling them to build long-term customer relationships, maximize profits, and thrive in a highly competitive market.

1.3 Aim

To develop a customer churn prediction tool in e-commerce using deep learning techniques in machine learning.

1.4 Specific Objectives

i.) To determine the factors that leads to customer churn in the e-commerce industry.

ii.) To determine existing models and algorithms used for customer churn prediction.

iii.) To develop a customer churn prediction tool in the ecommerce industry using deep learning techniques.

iv.) To test and evaluate the developed tool.

1.5 Research Questions

i.) What are the factors that lead to customer churn in e-commerce?

ii.) What are the existing models and algorithms used for customer churn prediction?

iii.) How can a customer churn prediction tool be developed?

iv.) How can the developed model be tested and evaluated?

1.6 Justification

Customer prediction models play a crucial role in solving customer retention issues and helping businesses build stronger relationships with their users. In various industries,

businesses have recognized the importance of reducing customer churn and have taken proactive measures to retain their existing customer base. By understanding the hierarchical dynamics of user behavior and identifying the underlying causes of churn, companies can implement targeted strategies to maximize customer retention. Rather than solely focusing on acquiring new customers, firms have started prioritizing the potential of their existing customer base. This shift in mindset allows businesses to capitalize on the untapped potential of their current customers.

To achieve the best customer retention outcomes, it is essential to transform the customer's status from being unknown to known. This involves predicting the customer's future decisions, which can be a highly complex task. However, by leveraging churn prediction models, businesses gain valuable insights into early warning signs and patterns that indicate a customer's likelihood of churning. By accurately forecasting customer churn, e-commerce players and other businesses can take proactive measures to prevent churn from occurring in the first place. This empowers companies to implement targeted strategies, personalized offers, and tailored interventions to retain customers before they decide to leave.

The implementation of customer prediction models not only helps lower customer churn rates but also extends the average duration of customer relationships. As customers stay loyal and engaged for longer periods, businesses can reap the benefits of increased customer lifetime value and profitability. Customer prediction models offer businesses the opportunity to foresee and anticipate customer churn, enabling them to implement effective strategies for retention. By leveraging these models, companies can reduce churn rates, extend customer relationships, and ultimately boost their overall profitability.

1.8 Scope and Limitations

This study developed a customer churn prediction tool to help e-commerce companies forecast customer churn. The tool will come in handy for these companies when planning retention strategies since the focus will narrow down to the high-risk customers who are likely to churn.

This study will also help managers use customer data to analyse customer behaviours. Customer churn is considered a lost profit. Hence this will be of great importance as far as companies' profits are concerned. For e-commerce companies, the significance is even greater since consumers can easily compare products or services and switch from one vendor to the next. However, this study was only limited to e-commerce companies hence other industries such as Telecom, Banking, and Media were out of this scope.

Chapter 2: Literature Review

2.1 Introduction

Customer churn prediction has emerged as a critical issue in the e-commerce industry. To address this issue, e-commerce companies must identify these customers before they leave. As a result, creating a unique classifier that predicts future churns is critical. This classifier must be able to recognize users who are likely to churn in the near future so that the company can respond quickly with the appropriate action. Understanding the factors that contribute to customer churn is critical in any business. Different market domains share some common characteristics, such as the price offered to the customer, the benefits provided, the number of years the customer has been associated with the organization, and so on.

To better grasp the idea of customer churn prediction, this chapter analyses a wide range of empirical and theoretical literature. To further comprehend the approaches taken to this problem, the chapter also examines the various algorithms utilized in customer churn prediction. After determining where the existing literature falls short, a conceptual model of the proposed system is developed.

2.2 Theoretical Literature

In the e-commerce world, client dissatisfaction is the primary driver of customer churn (Al Kurdi et al., 2020). The marketing and consumer behaviour literature agrees that customer satisfaction is a relative concept that is always evaluated in comparison to gold standard (Olander, 1977). Over the course of its development, multiple rival theories have been proposed, each based on a different set of criteria for explaining consumer happiness. Some of these theories include the Expectancy-Disconfirmation Paradigm (EDP), the Value-Precept Theory, the Attribution Theory, the Equity Theory, the Comparison Level Theory, the Evaluation Congruity Theory, the Person-Situation Fit model, the Performance-Importance model, the Dissonance and Contrast Theory, and the Contrast Theory.

2.2.1 Dissonance Theory

A person who expected a high-value product but received a low-value one would notice the discrepancy and experience cognitive dissonance, according to the Dissonance Theory (Yuksel, 2008). Unconfirmed expectations, in other words, produce cognitive dissonance or psychological distress. Dissonance, according to this notion, causes demands to lessen it, which could be accomplished by altering the perceived discrepancy. Because detecting disconfirmation is regarded to be psychologically painful, this hypothesis maintains that post-exposure ratings are mostly a function of expectation level. As a result, customers are said to distort expectation-discrepant performance in order to perceptually match their past expectation level.

This idea has received limited scientific support, in part because it is uncertain whether consumers would engage in such discrepancy corrections as suggested by the model in every consumption context. If the Dissonance Theory is right, organizations should seek to greatly exceed product performance in order to earn a higher product evaluation (Yi, 1990). The veracity of this assumption, however, is brought into question. Raising expectations far above product performance and then failing to satisfy these expectations may backfire since tiny variations may be overlooked, whereas significant disparities may result in a negative review. This proposition disregards the concept of "tolerance level." The tolerance level implies that consumers are willing to accept an acceptable range of performance around a given estimate.

This theory is applicable in e-commerce since customer happiness has been established as the most important element in determining client defection (Capraro et al., 2003). Customer satisfaction is directly tied to the quality of a company's product or service.

2.2.2 The Contrast Theory

When product performance falls short of consumer expectations, the mismatch between expectation and outcome encourages the consumer to exaggerate the gap, according to this theory. The Contrast theory states that a consumer who obtains a less valuable product

magnifies the difference between the received and expected product. Product performance that falls short of expectations, according to this hypothesis, will be evaluated lower than it is.

Poor performance would be worse than just poor if the Contrast Theory were applied to consumption, while good performance would be better than a good rating would imply (Oliver, 1997). According to dissonance theory, the opposite effects occur, and perceived performance is drawn to the initial expectation level, whether lower or higher than the consumer's expectations. The performance of the products or services in the relevant industries influences customer churn. Customers are more likely to defect to a competitor selling the same product with higher performance if a product falls short of their expectations. As a result, this theory serves as the foundation for examining the variables for predicting customer attrition.

2.2.3 The Expectancy Disconfirmation Paradigm

The Expectancy-Disconfirmation Paradigm (EDP) was proposed as the most promising theoretical framework for assessing customer satisfaction based on the shortcomings of the preceding early theories of consumer satisfaction (Zhang et al., 2021). According to the model, consumers purchase goods and services with pre-purchase expectations about the expected performance. The level of expectation becomes a standard against which the product is judged. After using the product or service, the results are compared to expectations. Confirmation occurs if the outcome matches the expectation. When there is a mismatch between expectations and outcomes, disconfirmation occurs. As a result of a positive or negative difference between expectations and perceptions, a customer is either satisfied or dissatisfied.

Despite its widespread popularity, however, the EDP is not without flaws. The main criticisms leveled at this approach center on the use of expectations as a comparison standard in measuring customer satisfaction, the dynamic nature of expectations and the timing of their measurement, the meaning of expectations to respondents, the use of difference scores in assessing satisfaction, and the EDP's reliability and validity in predicting customer satisfaction. This theory is critical when analyzing customer churn as customer expectations drive their

future purchase behaviors, and they may easily defect to competitors when these expectations are unmet.

2.3 Empirical Literature

Predicting customer churn is a critical issue for businesses. If a company can accurately predict who will leave, it can target those customers with a retention-focused campaign, which is far less expensive than targeting new customers (Matuszelański & Kopczewska, 2022). Technically, churn prediction is a typical classification task because the variable to predict is binary (churn or no churn). However, such binary prediction is less useful for subsequent retention campaign efforts. It is also critical that the machine learning model can predict the likelihood of a customer leaving and assist in creating a ranking of customers from most to least likely to churn. This churn likelihood ranking is hugely appealing to businesses.

2.3.1 Customer Churn

Customer churn refers to targeted customers who have decided to switch from one service provider, product, or company to another in the market. According to the literature, there are three types of customer churn (Amin et al., 2017). Churn can be either voluntary (active churn) or involuntary (passive churn). While both types of churn result in customer and revenue loss, their underlying causes and prevention strategies differ.

2.3.1.1 Active Churn

When a customer actively chooses to cancel their subscription, this is referred to as active churn. Business owners are primarily concerned with this type of churn because these customers make deliberate decisions to leave your company. There are several causes of active churn. First, a customer's experience with the product or service did not live up to their expectations or solve the problems they expected it to. Furthermore, a customer may have had a negative experience with your product or service, prompting them to look for alternatives. Again, superior services provided by competitors may attract their attention with a more

appealing solution that meets their needs or budget. Finally, customers may close their doors or go out of business and no longer require your services.

2.3.1.2 Passive Churn

Passive churn occurs when a company discontinues a service provided to a customer due to non-payment. Subscriptions are canceled when a customer's payment attempt fails without their knowledge.

2.3.3 Predicting Customer Churn

Several empirical studies have been conducted to predict customer churn in the e-commerce industry. Xiahou and Harada (2022) proposed a loss prediction model that combined k-means customer segmentation with support vector machine (SVM) prediction. The method categorizes customers into three groups and identifies the core customer groups. The support vector machine and logistic regression were compared to predict customer churn. The results showed that each prediction index improved significantly after customer segmentation, proving that k-means clustering segmentation is required. The SVM prediction was more accurate than the logistic regression prediction. However, this study had several limitations. Firstly, the result of several segmentations could significantly impact the model's prediction performance. In addition, the K-means algorithm was the only choice of this study; hence the model did not have convincing results. Lastly, the study only used a small number of predictive variables, which limits the promotion of results because a lot of shopping information is presented on B2C websites, and some of it may be ignored.

Matuszelaski and Kopczewska (2022) study used a transactions dataset from the Brazilian-based Olist e-commerce retail company to build and test a comprehensive churn model for e-commerce. The study was unique because customers were not bound by contract, their loyalty rate was meager, and churn was predicted without a long purchase history but by using only the first transaction. These challenges were met by enriching the transactional database (including typical behavioral data on the purchase value, shipping cost, product categories, and

several items in the first purchase) by analyzing customer location (zip code) and socio-economic environment and perception features provided by the customer as score and text reviews by the customer. This research has some limitations. The first is technical—only two algorithms were tested. Furthermore, the transactions under consideration occurred between 2016 and 2018, making them pre-COVID-19. One can expect that during the COVID-19 pandemic, e-commerce was driven by different factors and customer behavior than before and after the pandemic.

Wu and Meng (2016) presented a model for predicting e-commerce customer churn based on improved synthetic minority oversampling technique (SMOTE) and AdaBoost. First, the churn data is processed with improved SMOTE, which combines oversampling and under sampling methods to address the imbalance problem, and then the AdaBoost algorithm is integrated to predict. Finally, an empirical study on a B2C E-commerce platform demonstrated that this model outperformed mature customer churn prediction algorithms in terms of efficiency and accuracy by 6%.

Pondel et al (2021) research sought to create a deep learning model for predicting customer churn in e-commerce. The experiment was run on e-commerce data, with 75% of buyers being one-time customers. The prediction based on this business specificity (many one-time customers and very few repeat customers) is challenging and, by definition, must be somewhat inaccurate. From another angle, correct prediction and subsequent actions that resulted in higher customer retention were very appealing for overall business performance. Predictions with 74% accuracy, 78% precision, and 68% recall were very promising in this case but still have a long way to giving reliable customer churn predictions.

Researchers and academicians have developed several churn predictions models in the telecom sector. For instance, the primary contribution of Saheed and Hambali's (2021) work was the development of a churn prediction model that allows telecom carriers to forecast which customers are likely to churn. Machine learning (ML) approaches such as the Support Vector

Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest (RF), and Naive Bayes were used to create the model in this work (NB). This paper proposed a novel feature selection technique that combines the Information Gain and Ranker methods. The accuracy, precision, F-measure standard measures and 10-fold cross-validation were used to assess the model's performance. When feature selection is considered, the accuracy was 95.02%, and the accuracy was 92.92% without feature selection. When the results were compared to the existing methods, their models outperformed them in precision and F-measure.

The study conducted by Shabankareh et al (2021) is based on stacking data mining algorithms. According to implementation results, stacking methods can help improve customer churn identification results in various organizations. The data from the telecommunications industry was analyzed in this study to provide an effective churn detection solution. In this study, the authors decided to analyze data from one major telecommunications company to gain a better view and understanding of customer churn and, as a result, develop a better and more effective churn-prediction method.

Jain et al (2020) research aimed at predicting customer churn ahead of time so that proper customer retention steps could be taken using exploratory data analysis and to make customized offers for the targets. Their implementation for churn prediction consisted of a comparative study of four algorithmic models, namely logistic regression, random forest, SVM, and XGBoost, on three domains: banking, telecommunications, and information technology.

2.4 Factors that Lead to Customer Churn

2.4.1 Price Factor

Customers tend to buy lower-priced products or services when product quality and service are homogenized. Customers expect enterprises to deliver value and improve customer satisfaction, so enterprises should provide products or services that meet or even exceed their expectations. Customers may be willing to buy products and services continuously because of

corporate behavior and emotional reliance on the enterprise's delivered value. Price promotion is an effective win-back strategy for price-sensitive customers (Zhao et al., 2021).

E-commerce user penetration has been close to 50 percent in Kenya and is expected to reach 60 percent by 2027. On the one hand, many low-end users regard shopping online as a rigid demand in daily life and are significantly price sensitive; on the other hand, the online shopping experience between crucial players such as Jumia and Kilimall is shrinking, and service homogenization is severe. The price of products heavily influences users' consumption behaviors.

2.4.2 Product Factor

Customer churn caused by product factors occurs when there are flaws in product design or when the actual needs of customers and the market are not fully considered when designing products, which harms customer consumption. According to the life cycle value theory, each customer's future profit potential is unequal. In general, the closer the time between customer purchases, the higher the purchase frequency, the higher the monetary value they pay, the more likely they are to be interested in subsequent transactions, and the lower the likelihood of churn (Zhao et al., 2021).

These customers are more likely to refer other customers, allowing businesses to increase their market share and profits. Enterprises will prioritize marketing to groups with higher product dependence and life cycle value and invest more resources. However, the churned customers' previous consumption experience and behaviors determine whether they are willing to return to the previous business.

2.4.3 Customer Factor

According to Verhoef (2003), user characteristics such as consumption level and personal income can affect the churn rate. Customer value is reflected in user characteristics, which can

be a critical indicator for evaluating customer contribution. Customers are divided into valuable, mid-value, low-value, no-value, and below-zero customers to identify them better. According to Haucap (2015), high-income customers are more likely to sign service contracts and prefer bundled sales of convergence businesses and services.

Customers with greater purchasing power and income are less likely to churn. Customer value is directly proportional to market share and closely relates to customer loyalty. Companies that place a high value on their customers typically have lower operating costs. The greater the customer value, the greater the commitment, the greater the customer stability, and the lower the likelihood of churn. Furthermore, as the level of customer income rises, so will the performance of win-back agents (Zhao et al., 2021).

2.4.4 Business Factor

Convergence business is a common method of customer retention, and product mix synergy can be used to create more value for customers. Customers expect to be able to purchase all necessary products and services from the same e-commerce business. Businesses can save money on promotion and marketing. Bundling can reduce expenditure and psychological costs, and convergence business is a crucial business type that e-commerce players have evaluated in recent years; the convergence of electronics, clothes, automotive parts, groceries, and other products not only allows users to enjoy more additional services, but it also increases the user churn cost and threshold (Zhao et al., 2021).

2.4.5 Service Factor

According to Reinartz and Kumar (2003), one of the significant factors influencing customer churn is the perception of service quality. According to customer relationship management theory, the higher the level of customer satisfaction, the more difficult it is to reduce customer churn. Customer relationship management can help an enterprise maintain a better customer relationship, promote that relationship, and reduce customer churn. Service is the primary product of businesses. Customers will be disappointed if there is a gap between perceived and

expected service quality. Some customers will express their dissatisfaction with enterprises through complaints.

Customer churn is typically caused by an enterprise's failure to meet their expectations or dissatisfaction with using products or services, other than a reduction in their demands. Competitors may go to great lengths to attract customers by providing better services. As a result, if a company wants to increase customer loyalty, it must improve customer satisfaction and be committed to delivering more value to its customers.

2.5 Models and Frameworks

2.5.1 Models

Churn prediction has received a lot of attention from both industry and academicians. There has been extensive research in this field, with models and frameworks developed and applied to solve customer churn problems in contractual and non-contractual settings.

2.4.1.1 Neural Networks Model

The Neural Networks Model is used to develop non-linear functionality. Because of its comparable data processing structure, the model can learn. Due to the biological brain, these techniques produce successful results when applied to various problems such as classification, control, and prediction. Because of its likelihood prediction, the model differs from the classification model and the decision tree (Ahmad Naz et al., 2018). There are several neural network techniques, each with advantages and disadvantages. The researcher believes that neural networks are superior to decision trees and regression analysis models for churn prediction. The main limitation of this model is that it requires a lot of computational power. This is caused by computation done through backpropagation on each node and the extensive nature of the dataset needed for training.

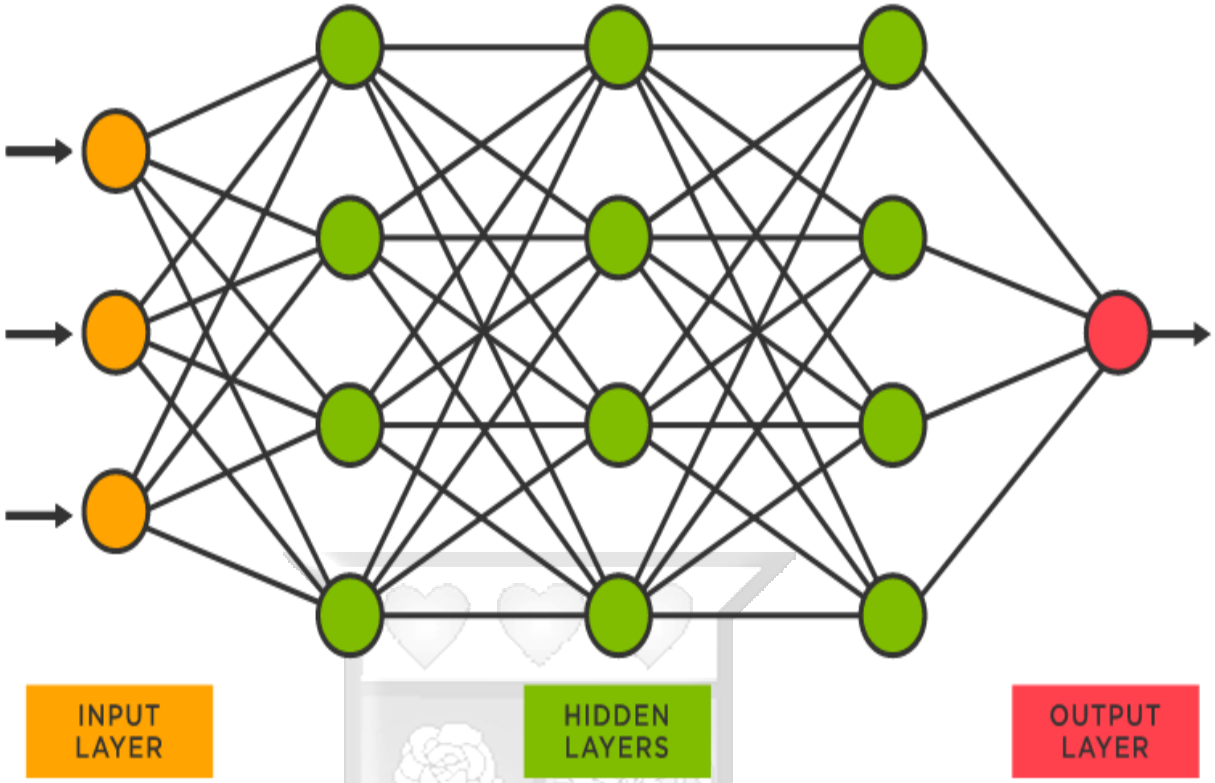


Figure 2.1 Representation of a Neural Network Neural Network

(Kulkarni et al., 2019)).

2.4.1.2 Support Vector Machine (SVM) Model

The SVM classifier deals with linear permutations of a subset of the training set by finding a maximum edge over an energized plane. With the help of the most critical part of vectors, nonlinearly divisible input features, the SVM plots the data into a high dimensional features space close to infinite and then categorizes the data by the most increased scope hyper-plane (Ahmad Naz et al., 2018).

$$f(x) = sgn\left(\sum_i^M \quad y_i \alpha_i \Phi\left(X_i, X\right) + \delta\right)$$ Equation 2.1 SVM Formula

Where:

M=The number of samples in training dataset

Xi=Shows vector support when ai > 0.

I=Shows a core function.

X=Unidentified sample feature vector.

d = doorstep.

(ai) is a parameter resulting from a curved quadratic programming problem concerning linear constraint. This technique shows that Polynomial kernel & Gaussian radial basis functions (RBF) are frequently used in favor of kernel functions. The (d) is another parameter resulting from picking any i where ai> 0 and the Karush–Kuhn– Tucker condition. The major drawback of using SVM is its inability to perform well when the data has more overlapping noise, that is, target classes.

2.4.1.3 Decision Tree Model

The decision tree is the most prominent predictive model used for the classification of upcoming trials. The decision tree comprises two steps: tree construction and tree pruning. The training set data is recursively partitioned per the values of the attributes during tree building (Ahmad Naz et al., 2018). This process is repeated until no partition has identical values left. Due to noisy data, some values may be removed from the data during this process. The branches with the highest estimated error rate are chosen and removed in pruning. Tree pruning is the process of predicting accuracy and reducing the complexity of a decision tree. In their customer churn prediction research, Wai-Ho Au et al (2003) applied decision trees on a database of 100,000 records provided by a carrier in Malaysia. The model was robust in this classification task based on the contractual setting of the telecom carrier.

2.4.1.4 Convolutional Neural Network Model

CNN is a deep learning model for processing data with a grid pattern, such as images, inspired by animal visual cortex organization (Yamashita et al., 2018) and designed to learn spatial hierarchies of features automatically and adaptively, from low- to high-level patterns. CNN is a mathematical construct comprised of three layers: convolution, pooling, and fully connected. The first two layers, convolution and pooling, extract features, and the third, a fully connected

layer, map the extracted features into the final output, such as classification. A convolution layer is essential in CNN, which comprises a stack of mathematical operations such as convolution and linear operation.

Tariq et al (2021) used a 2-D convolutional neural network, a deep learning technique, in their study to predict customer churn. The proposed CNN model is a layered architecture with two phases: data load and preprocessing layer and 2-D CNN layer. The model achieved a high level of accuracy, precision, and recall.

2.4.2 Frameworks

Machine learning frameworks have been developed to make it easy to write and deploy machine learning models. The powerful frameworks for machine learning are TensorFlow, PyTorch, Keras, and ScikLearn.

2.4.2.1 TensorFlow

TensorFlow is Google's open-source machine learning and deep learning framework that is easy to use and adapt to build the current mainstream deep learning model (Yu et al., 2019). TensorFlow employs a declarative programming paradigm. This allows researchers to focus on the symbolic definition of what needs to be computed rather than how exactly and in what order these computations are to be performed, as is the case in imperative programming. TensorFlow's architecture is founded on graphs. This graph aggregates and explains all of the series computations performed during training as shown in Figure 2.1. This abstract representation of the model can be optimized for numerical stability and performance, and individual parts can be translated to be executed by a computer processor or graphics chip (Rampasek & Goldenberg, 2016). Perhaps most importantly, the symbolic representation enables automatic differentiation, which provides a convenient way to optimize many functions.

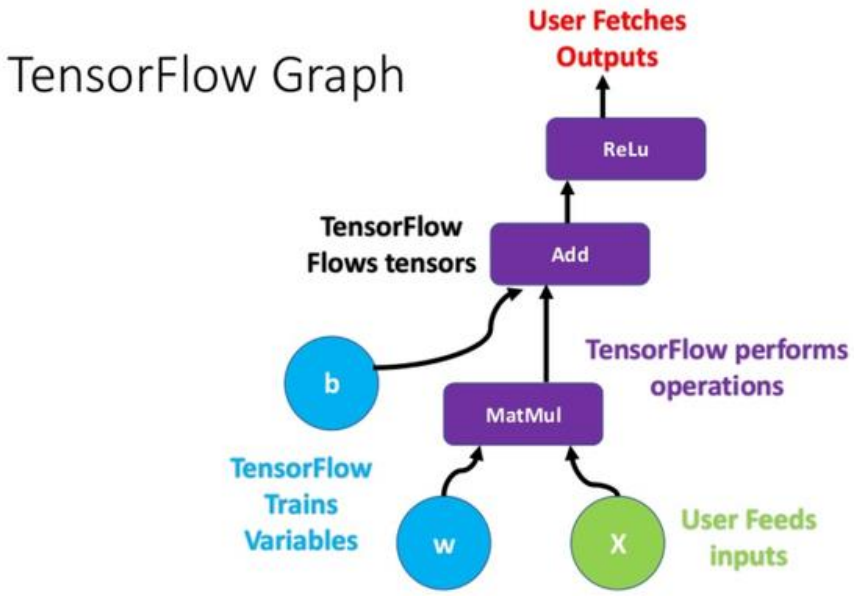


Figure 2.2 TensorFlow Graph (Sharma, 2022).

2.4.2.2 PyTorch

PyTorch is a free and open-source machine learning (ML) framework built on Python and the Torch library (Yasar, 2021). Torch is an open-source ML library written in the Lua scripting language used to create deep neural networks. It is one of the most popular platforms for deep learning research. The framework is designed to shorten the time between research prototyping and deployment. The PyTorch framework supports over 200 different mathematical operations. PyTorch's popularity is growing, making creating artificial neural network models easier. Data scientists primarily use PyTorch for research and artificial intelligence (AI) applications. PyTorch is distributed under a BSD license that has been modified. Figure 2.2 shows the architecture of the PyTorch Framework.

2.4.2.3 Keras

Keras is a Python-based machine-learning framework. Because each line of code creates one layer of a network, it is easier to build complete solutions and read. This framework appears to have the most advanced algorithm optimizers, normalization routines, and activation functions (Erickson et al., 2017). Although Keras supports both Theano and TensorFlow

backends, the assumptions for the input data dimension differ, necessitating careful design for the code to work with both backends. The framework is well documented, and examples addressing many problems are provided. There are also pre-trained models of commonly used architectures for transfer learning implementation. Figure 2.2 shows a 5-step lifecycle of the deep learning model in Keras.

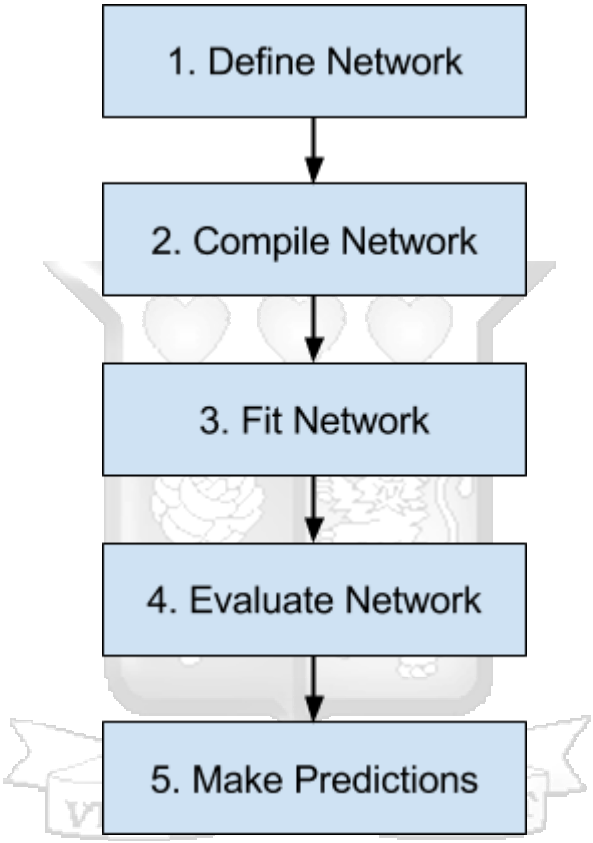


Figure 2.3 Five-Step Life Cycle of Deep Learning Model in Keras
(Brownlee, 2016).

2.4.2.4 Scikit-Learn

Scikit-learn is a Python module incorporating a diverse set of cutting-edge machine-learning algorithms for medium-scale supervised and unsupervised problems. Using a general-purpose, high-level language, this framework aims to bring machine learning to non-specialists. The emphasis is on usability, performance, documentation, and API consistency. It has few

dependencies and is distributed under a simplified BSD license, making it suitable for use in both academic and commercial settings.

2.6 Architecture and Design

Customer churn prediction has seen the number of architectures and designs adopted increase rapidly.

2.6.1 Big Data Architecture

Ahmad et al (2019) created an architecture for customer churn prediction that included Hadoop Distributed File System HDFS2 to store data, Spark execution engine to process data, Yarn to manage resources, Zeppelin as the development user interface, Ambari to monitor the system, Ranger to secure the system, and Flume System and Scoop tool to acquire data from outside the SYTL-BD framework into HDFS. They collected data for nine months using hardware resources that included 12 nodes with 32 Gigabyte RAM, 10 Terabyte storage capacity, and 16 cores processor for each node.

The dataset was used to extract churn predictive model features. Figure 2.3 depicts the various stages of the data life cycle. Because it performs the processing on RAM, the Spark engine was used in most of the model's phases, such as data processing, feature engineering, training, and testing the model. There are numerous additional benefits. One of these benefits is that this engine includes a wide range of libraries for implementing all stages of the machine learning lifecycle. The only limitation of the architecture is that it does not perform well in settings that are not contractual.

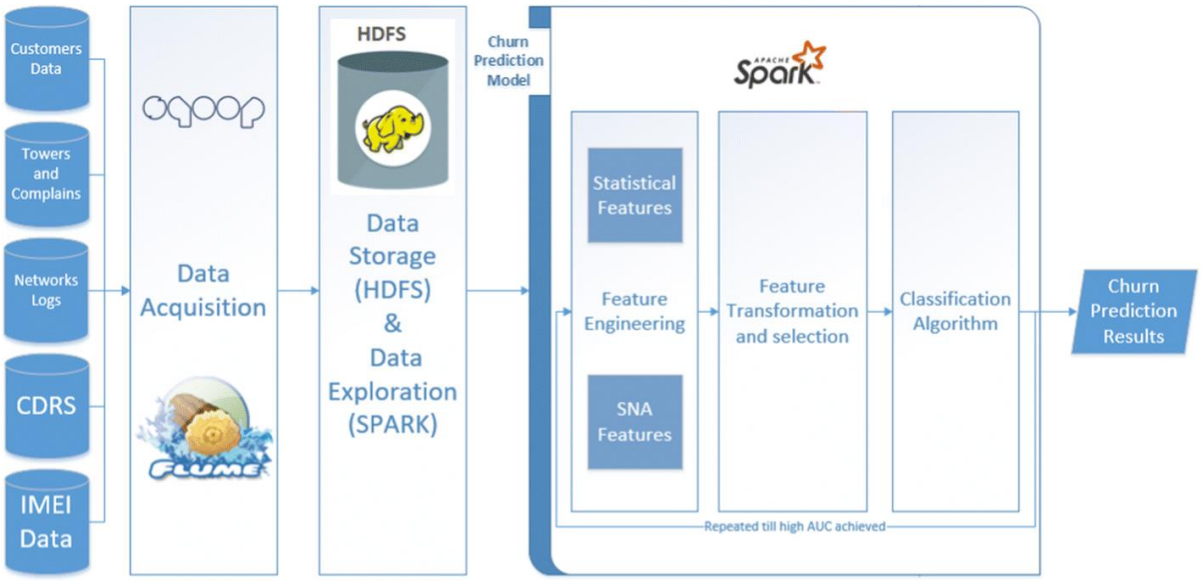


Figure 2.4 Big Data Architecture *(Ahmad et al., 2019)*

2.6.2 Ensemble Architecture

In the telecom sector, Kulkarni et al (2019) architecture examined data from churned customers and their attributes before the churn using ensemble classifiers. The dataset includes customers' demographic information, total charges, and the type of service they receive from the company. It comprises churn data from over a thousand consumers divided across 21 Kaggle parameters. They attempted to predict existing customers' reactions by fitting statistical models that link the predictors to the response. There are some three types of ensemble learning,

- **bagging**, that often considers homogeneous weak learners, learns them independently from each other in parallel and combines them following some kind of deterministic averaging process.
- **boosting**, that often considers homogeneous weak learners, learns them sequentially in a very adaptative way (a base model depends on the previous ones) and combines them following a deterministic strategy.

- **stacking**, that often considers heterogeneous weak learners, learns them in parallel and combines them by training a meta-model to output a prediction based on the different weak models' predictions.

This method is known as supervised learning. Figure 2.4 shows the architecture of the system.

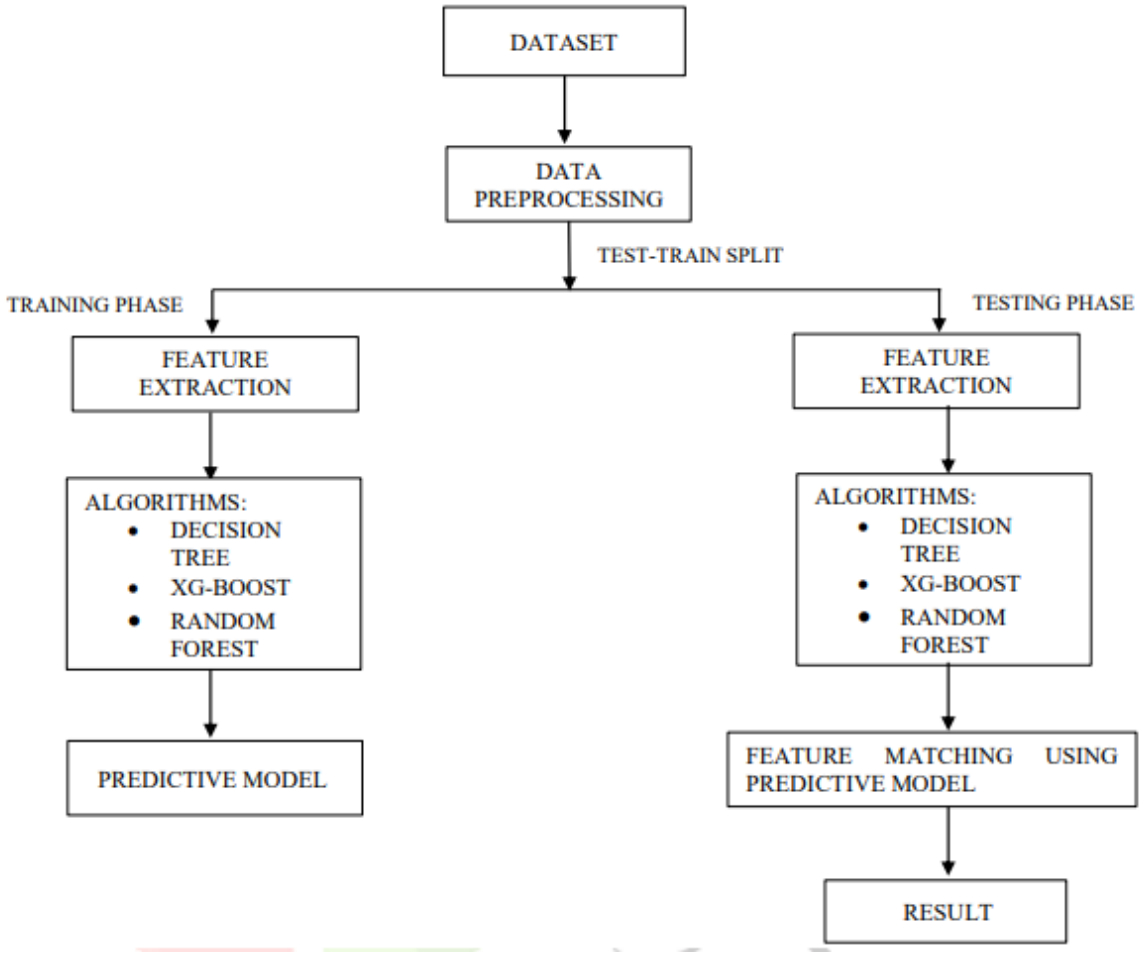


Figure 2.5 Ensemble Architecture (Kulkarni et al., 2019)

2.6.3 Hybrid Firefly Architecture

Ahmed & Maheswari (2017) proposed a hybrid firefly architecture to address the massive computational requirements caused by comparisons. Figure 2.5 depicts the operation of the hybrid firefly algorithm. The classification process begins with the creation of the search space.

The initial firefly population is generated and distributed across the search space. The distribution of fireflies occurs at random. Each firefly's position is recorded, and the initial intensity of the fireflies (Intensity) is identified based on their distance from the test data.

The Simulated Annealing module uses firefly intensities and test data to find the best solution for the test data. Firefly 0 is assigned to the test data, while the remaining fireflies are given to the training set. This architecture does not include any analysis of imbalance levels or data scarcity. Incorporating Game theory into the decision-making process will also help improve the accuracy, which is currently low, and in identifying churn.

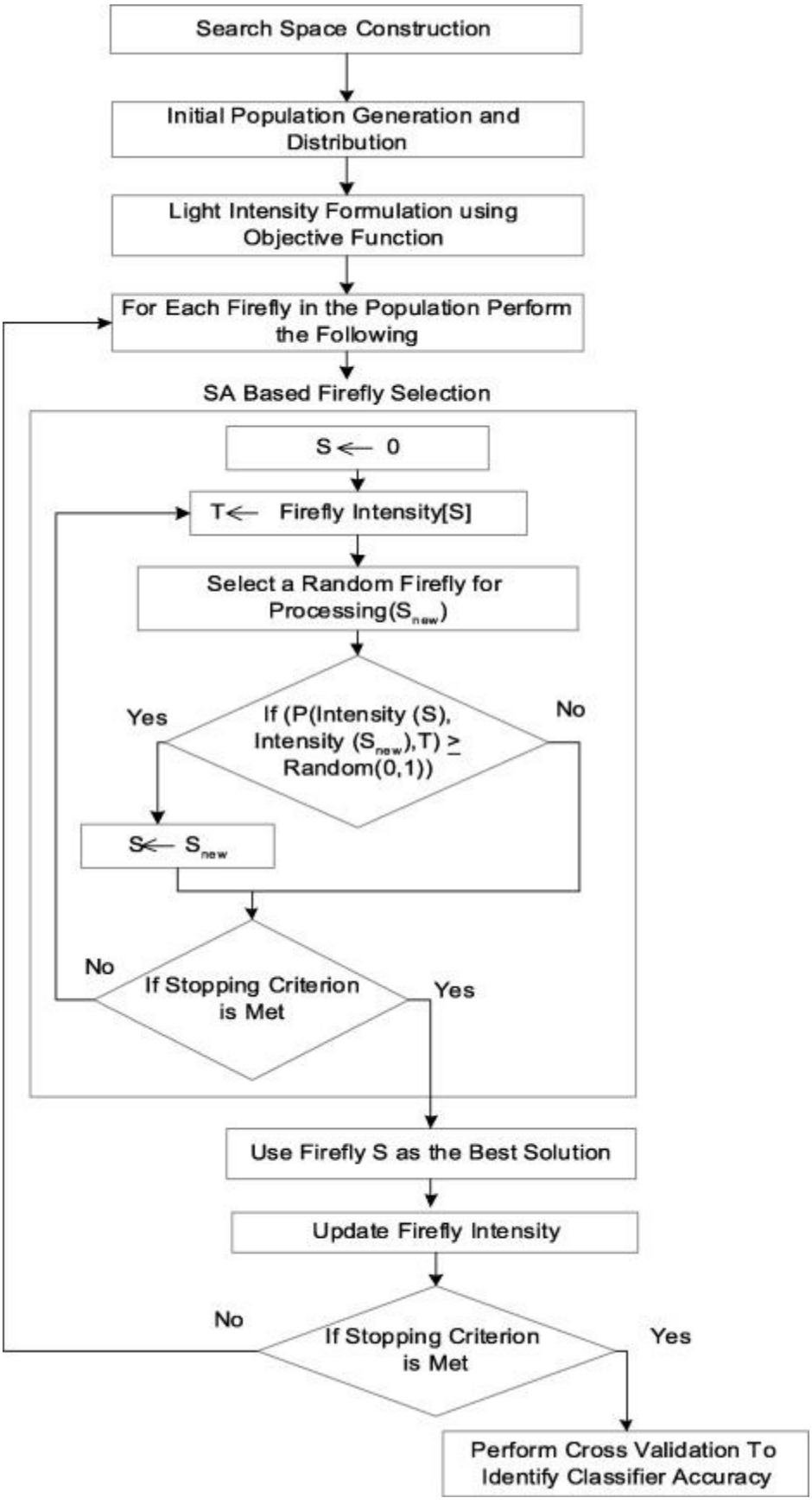

Search Space Construction
Initial Population Generation and Distribution
Light Intensity Formulation using Objective Function
For Each Firefly in the Population Perform the Following
SA Based Firefly Selection
S ← 0
T ← Firefly Intensity[S]
Select a Random Firefly for Processing(S_new)
If (P(Intensity (S), Intensity (S_new),T) ≥ Random(0,1))
Yes
No
S ← S_new
If Stopping Criterion is Met
No
Yes
Use Firefly S as the Best Solution
Update Firefly Intensity
If Stopping Criterion is Met
No
Yes
Perform Cross Validation To Identify Classifier Accuracy

Figure 2.6 Hybrid Firefly Architecture (Ahmed & Maheswari, 2017)

2.7 Algorithms

2.7.1 Machine Learning

Machine learning (ML) is an umbrella term for various algorithms that make intelligent predictions based on a data set. These data sets are frequently large, containing millions of unique data points. Recent advances in machine learning appear to have achieved a human level of semantic understanding and information extraction, as well as the ability to detect abstract patterns more accurately than human experts (Nichols et al., 2018).

There are two types of machine learning: supervised learning and unsupervised learning. Supervised learning entails teaching the model with input data that already has the correct output (Nichols et al., 2018). In image classification tasks, supervised learning is more widely used. Unsupervised learning is the process by which a model trains itself on data. Typically, this will entail tasks such as cluster detection or pattern recognition.

2.7.2 Random Forests

Random forests are a tree predictor combination in which each tree is dependent on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). As the number of trees in a forest grows, the generalization error approaches a limit. The generalization error of a forest of tree classifiers is determined by the strength of the trees in the forest and their correlation.

Individual decision trees are easily interpretable, but this interpretability is lost in random forests due to the aggregated nature of the decision trees. On the other hand, random forests are frequently much better at prediction tasks. When compared to decision trees, the random forest algorithm estimates the error rate more accurately. More specifically, as the number of trees increases, the error rate is mathematically proven always to converge (Schonlau & Zou,

2020). In their study, (Xie et al., 2009) demonstrated the implementation of improved balanced random forests (IBRF) to churn prediction. They examined the efficacy of the standard random forests approach in predicting customer churn, integrating sampling techniques and cost-sensitive learning to achieve a better performance than the majority of existing algorithms. The nature of IBRF is that the best features are learned iteratively by modifying the class distribution and increasing penalties for misclassifying the minority class. The method was applied to a real-world bank customer churn data set. In comparison to other algorithms, such as artificial neural networks, decision trees, and class-weighted core support vector machines (CWC-SVM), it significantly improved prediction accuracy. Figure 2.5 shows the structure of random forests.

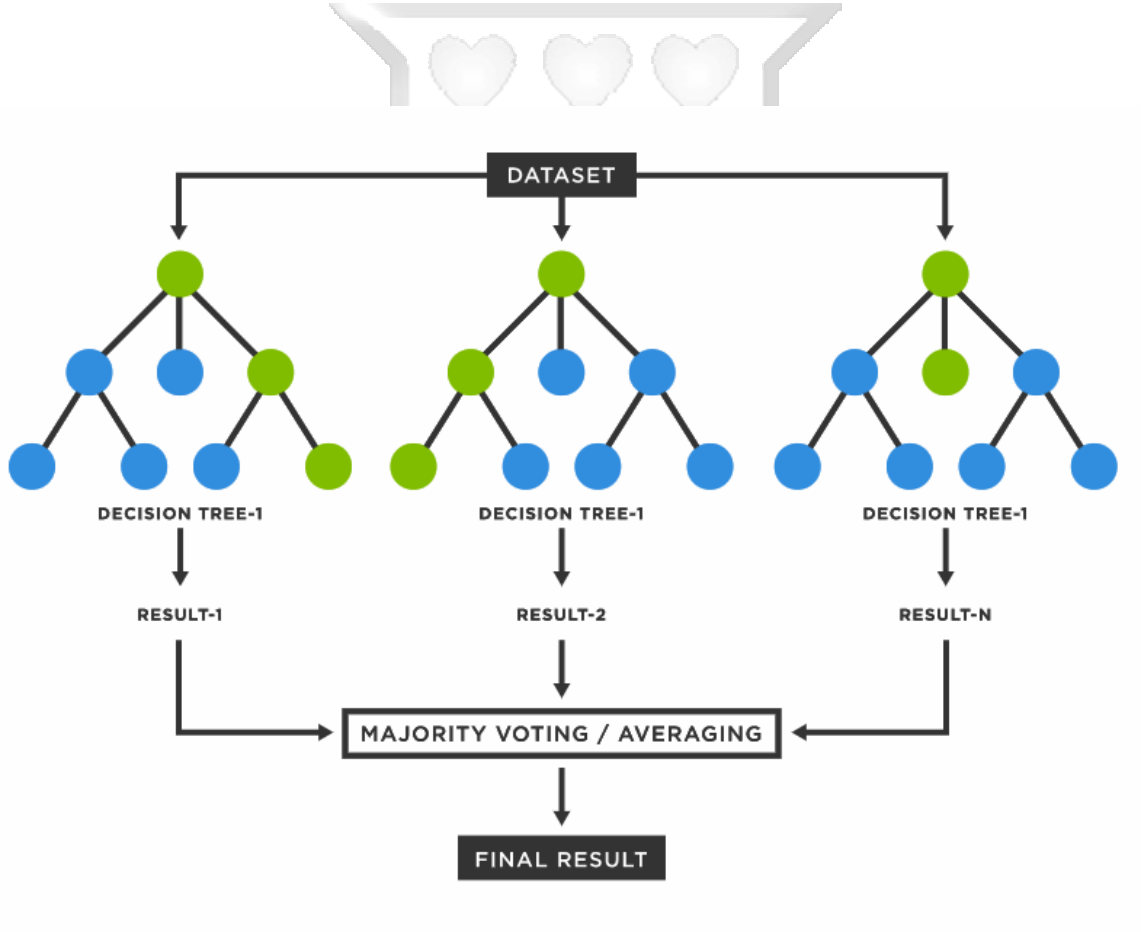


Figure 2.7 Random Forest Algorithm (Schonlau & Zou, 2020).

2.7.3 K-Means

The k-means algorithm is an iterative algorithm that attempts to partition the dataset into K distinct non-overlapping subgroups (clusters), with each data point belonging to only one of these groups. It tries to keep intra-cluster data points as similar as possible while keeping clusters as different (far) as possible. It assigns data points to clusters so that the sum of the squared distances between the data points and the cluster's centroid (the arithmetic mean of all the data points in that cluster) is as small as possible. The lower the variation within clusters, the more homogeneous (similar) the data points within the same cluster are (Ahmed et al., 2020).

The use of the K-means algorithm for predicting customer attrition was investigated by Xiahou and Harada (2022). They evaluated the predictive power of the Support Vector Machine (SVM) and Linear Regression (LR) models using customer behavior data from a B2C e-commerce business. In order to evaluate the prediction performance of the two models, the k-means algorithm was first used for clustering subdivision to classify customers into three categories, and then predictions were made for these three customer types. It was determined the accuracy, recall, precision, and AUC. The results of the experiment demonstrate that the consumer segmentation of each prediction index has significantly improved. Consequently, they demonstrated the necessity of k-means segmentation clustering. Figure 2.7 shows an example of K-means clustering using K as four.

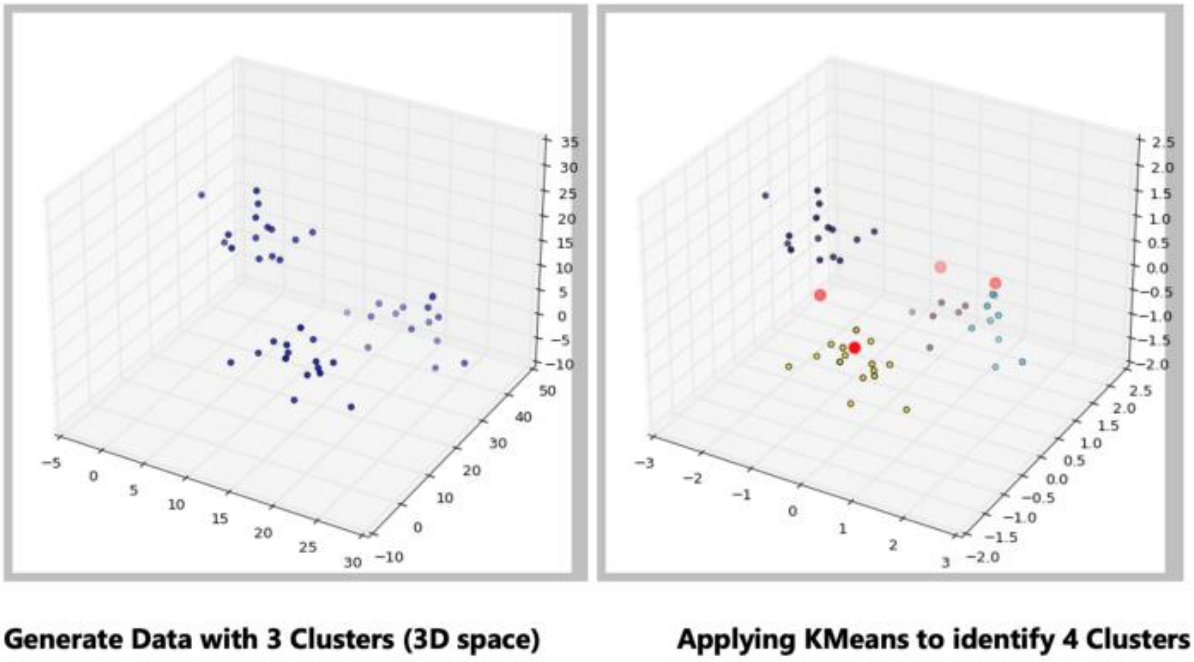


Figure 2.8 3D Clusters of K-means algorithm (Ahmed et al., 2020)

2.7.4 Decision Tree

A decision tree is a recursive partition of the instance space that expresses a classifier. The decision tree is made up of nodes that form a rooted tree, which means it is a directed tree with no incoming edges (Krzysztof Grabczewski, 2016). Every other node has only one incoming advantage. An internal or test node has outgoing edges. All other nodes are known as leaves (terminal or decision nodes). Each internal node in a decision tree divides the instance space into two or more sub-spaces based on a discrete function of the input attribute values.

In the most basic and standard case, each test considers a single attribute, and the instance space is partitioned based on the attribute's value. The condition in the case of numeric attributes refers to a range. Each leaf is assigned to a class that represents the best target value. Alternatively, the leaf could contain a probability vector indicating the likelihood of the target attribute having a specific value. Instances are classified by navigating them from the tree's root to a leaf based on the results of the tests along the way.

(Ramadhanti et al., 2023) utilized data mining techniques with decision tree algorithms to predict customer attrition in an Indonesian telecommunications company. The optimal decision tree model had criterion information gain parameters of minimal gain = 0.01 and maximum depth = 6. This decision tree model had an accuracy of 78.28% and a customer retention rate of 19.6%. According to their model, the analyzed company's customers tend to leave voluntarily. Important determinants of customer churn include contract type, number of monthly downloads, tenure, customer satisfaction value, and add-on services. In this corporation, the type of contract has the greatest influence on customer churn. Figure 2.8 depicts the structure of a decision tree (Ampadu, 2021).

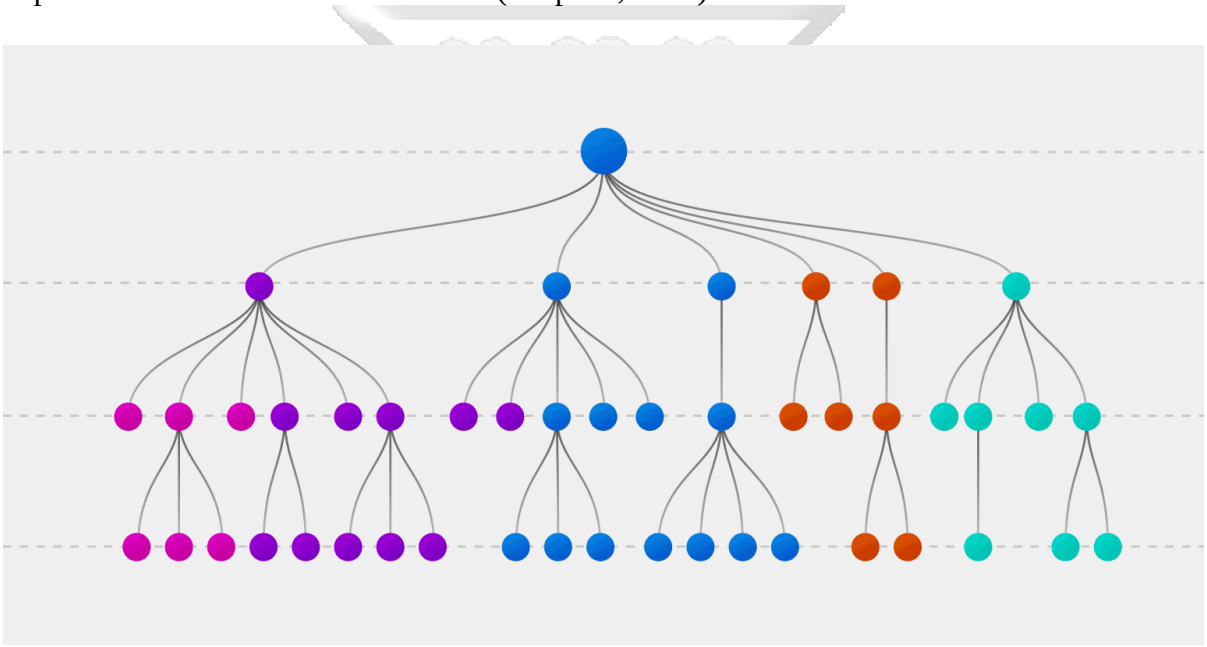


Figure 2.9 Decision Tree Algorithm (Ampadu, 2021)

2.7 Gaps in Existing Systems

Several studies on customer churn prediction have been conducted. Still, most of these studies have been undertaken in contractual settings, where churn is defined as a client resigning from using a company's services by canceling their subscription or breaching the contract. This method of determining churn differs from that used in business, such as e-commerce, where the customer is not required to notify the company of their intention to resign.

One issue that arises in the non-contractual setting is the definition of churn. Because there is no precise moment when a customer discontinues using the company's services. Customers who do not make any new purchases from a retail store for three months are classified as "partial churners." In other approaches, "churners" are all customers with a lower-than-average frequency of purchases because these customers have been shown to provide little value to the company.

Empirical studies conducted on e-commerce have shown massive potential in the sector. However, most traditional churn prediction models cannot fully achieve the business goal of maximizing customer retention, specifically how to minimize customer churn through the commodity recommendation system. Furthermore, most developed models are challenging to deploy and run regularly to understand changes in customer behaviour and allow relationship managers to act accordingly. Finally, most studies did not collect additional customer data, such as reviews, ratings, and so on, to better understand why a specific group of customers may be leaving.

2.8 Conceptual Model

Data from Kaggle is preprocessed, which includes cleaning and deleting extraneous columns. After that, the preprocessed data is divided into train and test data. Train data is used to train deep learning algorithms to generate the model. After that, the model is validated using test

data to ensure accuracy, precision, and recall. Figure 2.9 displays the proposed system's conceptual model.

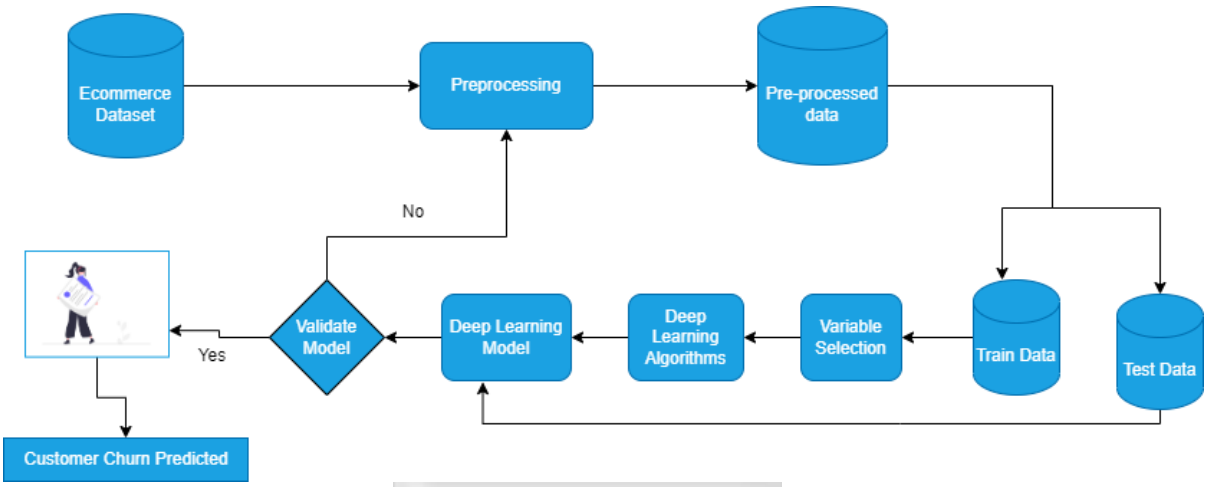


Figure 2.10: Conceptual Model

2.9 Conclusion

By examining a variety of algorithms such as Random Forests, K-means, and Decision Trees, a deeper understanding of the approaches used in customer churn prediction was gained. The examination of these algorithms shed light on their strengths and limitations in addressing the issue of customer churn. Moreover, this chapter has identified several key factors that influence customer churn, including business factors, price factors, product factors, service factors, and customer factors. Understanding these factors is crucial in developing an effective customer churn prediction system as they provide valuable insights into the reasons behind customer churn. Building upon the insights gained from the literature review and algorithm analysis, a conceptual model of the proposed customer churn prediction system was developed. This model served as a foundation for the subsequent steps in designing and implementing an accurate and reliable system for predicting customer churn.

Chapter 3: Research Methodology

3.1 Introduction

The research methodology is the path researchers must take to conduct their research (Sileyew, 2019). It demonstrates how researchers formulate their problems and objectives and present their findings based on the data collected during the study period. This chapter illustrates how the research outcome were obtained per the study's objectives. As a result, this chapter discusses the research methods used during the research process. It includes the study's methodology, from the research design to the dissemination of the results. Both qualitative and quantitative research methods were considered to meet the study's objectives. Because data was obtained from all aspects of the data source during the study period, the study employed these mixed strategies. As a result, this chapter aims to satisfy the researcher's research plan and target.

3.2 Research Design and Philosophy

The term "research philosophy" refers to a belief about how knowledge about a phenomenon should be obtained, processed, and applied. In other words, the study is guided by the core belief system. According to (Saunders et al., 2012), there are ten philosophies: Positivism, realism, Interpretivism, objectivism, subjectivism, pragmatism, functionalism, interpretive, radical humanist, and radical structuralist, as shown in figure3.1, in the research onion. We examine two of these philosophies: positivism and Interpretivism.

3.2.1 Research Design

The term "research design" refers to the overall strategy chosen by the researcher to integrate various components of the study coherently and logically, ensuring that the research problem is effectively addressed (Cooper & Schindler, 2014). The research design serves as the data collection, measurement, and analysis blueprint. The research design phase is critical because it focuses on the researcher's approaches to achieving goals (Gachenge, 2020). Descriptive, exploratory, and explanatory research designs are the most common. A descriptive research

design is a scientific method for describing the characteristics or behaviors of the Population under study.

On the other hand, exploratory research design refers to unstructured research conducted to gain background information about the research problem and does not usually include a sampling design, research objectives, or a questionnaire. In contrast, the explanatory research design is a research design that can be used in determining causality or studying to find out whether one variable or more explains the causes or the effect of one variable or more (Tobi & Kampen, 2017). This study employed a descriptive research design, a scientific method for describing the characteristics or behaviors of the Population under investigation. The descriptive design was used in this study because it is desirable when the researcher wishes to project findings to a larger population being investigated from a representative sample, thus justifying its selection.

3.2.2 Research Philosophy

This study adopted positivism as the research-preferred philosophy. Positivism, as a philosophy, holds that only "factual" information received by observation (the senses), including measurement, is reliable. In positivist investigations, the researcher's participation is confined to data gathering and objective interpretation. In other words, the researcher is an impartial analyst who separates herself from personal ideals while investigating. The study outcomes in these sorts of investigations are frequently apparent and quantitative (Žukauskas et al., 2018).

Positivism is based on quantitative observations that result in statistical analysis. It has been the dominating research method in business and management and STEM disciplines for decades. It has been stated that "as a philosophy, positivism is consistent with the empiricist idea that knowledge is derived from human experience." It takes an atomistic, ontological view of the universe, seeing it as a collection of distinct, observable elements and events that interact in observable, defined, and common ways" (Balmer & Murcott, 2017).

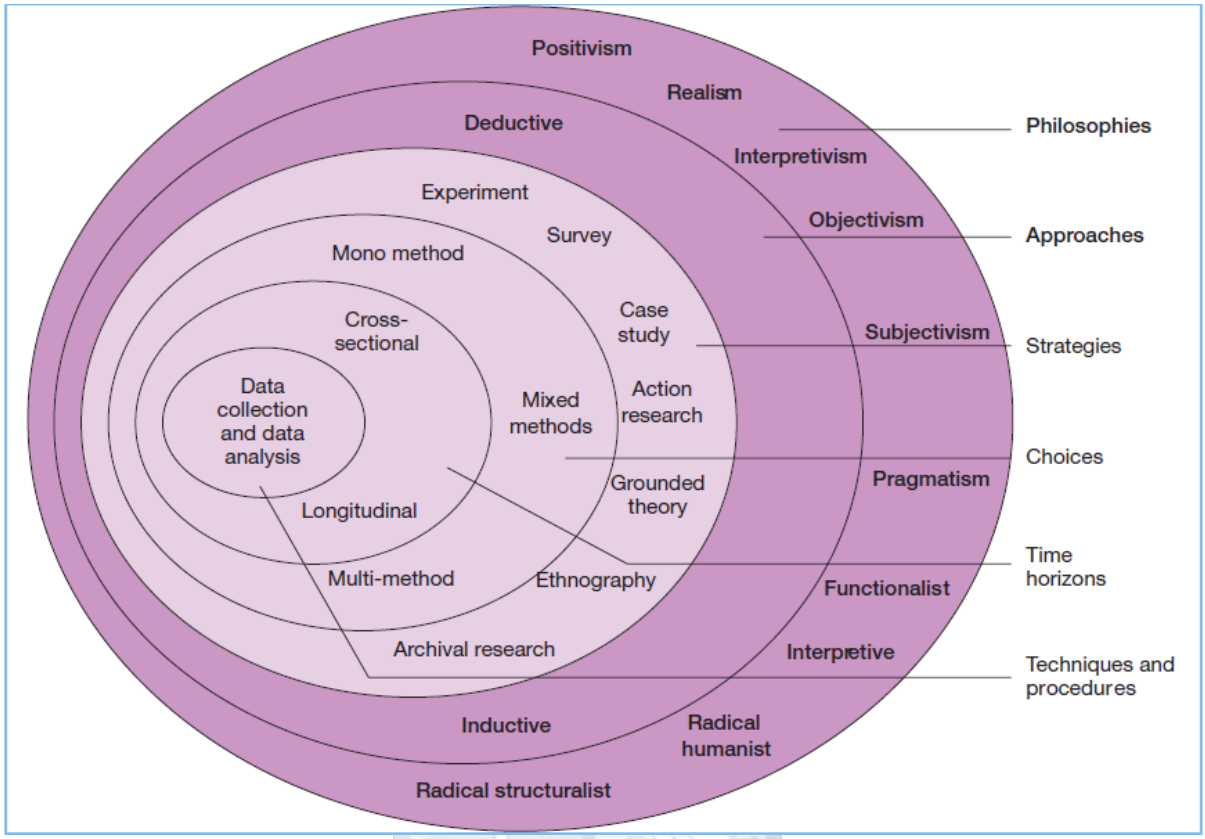


Figure 3.1: Research Onion

(Saunders et al., *Research methods for business students* 2012)

Positivism usually takes on a deductive approach to the research. Positive thinking promotes scientific discoveries and practitioners' approaches to academic learning (Balmer & Murcott, 2017). The philosophical foundations of the positivism paradigm are:

i.) Ontology

It refers to the nature of reality. The positivist paradigm is predicated on believing that a single concrete reality can be comprehended, identified, and quantified. This enables explanation and prediction in a causal relationship framework to function organically, as Causal conclusions are based on the following:

(1) time.

Precedence (i.e., X must first occur for X to produce Y).

(2) association.

(i.e., precede Y in time), X and Y are linked)

(3) the absence of co-founders

That is, no factors other than the discovered factors impact the result.

Within the space identified, X is the lone cause of Y (Park et al., 2020).

ii.) Epistemology

This refers to the nature of technology. Positivists argue that knowledge can and must be generated objectively without regard for the values of the researchers or participants. When sufficiently established, knowledge is truth—that is, it is definite, coherent with reality, and accurate. Absolute isolation between the research participant and the researcher is required to create truth effectively.

Positivists use dualism and objectivity to attain this distinction. In other words, positivist thinking holds that participants and researchers may be distinguished (dualism) (Park et al., 2020).

3.3 Population and Sampling

In this section, datasets and their sample sizes were examined, and results analysed. This study used 80% of the data for training, while 20% was used for testing. There is no hard and fast rule stating that the sample size for machine learning training and testing must be 80% and 20%, respectively. This is a common ratio, but the exact proportions can vary depending on the size and characteristics of the dataset, as well as the analysis's goals.

The advantage of using an 80/20 split is that it provides a large enough sample for training the model while still leaving enough data for testing and evaluating its performance. This can help ensure that the model isn't overfitting to the training data, which can lead to poor performance on new, previously unseen data.

3.3.1 Population

This study, as it focuses on e-commerce data was keen on historical customer data including, but not limited to, demographics, behavioural and revenue and subscription data. Subscription data included the subscription date, plan or pricing tier and the monthly recurring revenue at the individual customer level. For analysis and forecasting, website-purchasing customers' consumption information was selected. The dataset contains customer consumption records and historical platform behavior interactions.

3.3.2 Sampling Size

A sample size is a subset of the Population that has been chosen. The sample size should be big enough to offer the needed trust in the data and the researcher, according to (Saunders et al., 2019). The larger the sample size, the more precision is sought and the higher the confidence level. The researcher should be able to calculate the likely response rate or the percentage of cases in the sample that will react and adjust the sample size accordingly.

This study utilized Kaggle's data on e-commerce. Kaggle is an online platform that allows users to locate datasets for machine learning. Kaggle contains data that has been collected from multiple ecommerce sites, making the data set rich both in size and complexity and also in that customer patterns from all over the world are looked in to.

3.4 Data Collection Method and Analysis

Data collection is collecting information from all relevant sources to find answers to the research problem, test the hypothesis, and evaluate the outcomes (Cooper & Schindler, 2014).

3.4.1 Data Collection

Secondary data was used in this study from Kaggle. Secondary data was the preferred data collection method because of the confidentiality of actual data from e-commerce players. In addition, the data was freely available for use; hence it saved time otherwise spent on collecting primary data. The data used for this study belonged to an e-commerce website.

3.3.2 Data Analysis

In this research study, data analysis included cleaning the data to remove any missing values and inspecting and transforming the data into a format that could be analyzed to gain insight into the study's objectives. This study employed both qualitative and quantitative data analysis techniques. Because there was insufficient time and data resources to conduct extensive research, the study relied on a deductive approach for qualitative data analysis.

3.5 Research Quality and Reliability

In qualitative research, reliability relates to the consistency of replies to various data set coders. It can be improved by taking thorough field notes, recording them, and transcribing digital data. The fundamental approach to qualitative research rigor and quality is systematic, self-aware data collection, interpretation, and communication. Validity in qualitative research, however, be described differently than in quantitative research since response categories are not organized and therefore are not mutually exclusive; the open-ended nature of qualitative inquiry does not lend itself well to probability theory or statistical inference to a broader population (Guest et al. 2020).

Data saturation is the most popular qualitative research method for determining sample sizes. Saturation has been defined as a state within a study where no additional data can be found to develop different properties or themes by the researcher (Glaser & Strauss, 2017). In broad terms, saturation is used in qualitative exploration as a criterion for discontinuing data collection and analysis. Morse pointed out more than 20 years ago that satiety is an essential factor in severity. Saturation is therefore present in all high-quality qualitative exploration (Saunders et al., 2017).

Since most qualitative research uses non-probabilistic, targeted sampling consistent with the nature and goals of qualitative research, asking questions to saturation is recognized as a good

practice (Guest et al., 2020). As part of research reliability confirmation, the researcher also provided the supervisor with a sample of the interview transcripts to enable him to carry out independent data analysis to test the stability of responses to multiple coders of data sets.

3.6 System Development Methodology

Agile software development was the preferred software development methodology for this study. This is because Agile methodology recognizes that software development is a collaborative process and needs to be treated as such (Hoda et al., 2018). The creators of the doctrine argued that these projects and software are unlike any other products or services purchased by companies. Firms do not need to go into truck factories to work with manufacturers on development. However, one cannot adopt the same approach when faced with software. All types of software need some customization, and the larger the organization, the more customization is necessary (Hoda et al., 2018). The nature of software development necessitates close cooperation between the developers and their customers.

The methodology rests on several key factors. The first is openness. The methodology requires opening effective communication channels from the beginning (Al-Saqqa et al., 2020). Secondly, the teams also have to work on common goals. This approach does not just assume that both the client and developer have the same plans. Instead, it insists on formalizing these long-term and short-term goals by a team that includes both parties. Finally, the method depends on continuous development and does not include ceremonies or ostentatious handing-over ceremonies (Al-Saqqa et al., 2020). By the time the project is over, the client should effectively control everything.

A vital advantage of this approach is that it looks at the problem holistically. Instead of considering the product or service as something that a company makes and then sells to another company, it creates a collaborative paradigm (Al-Saqqa et al., 2020). Another notable advantage is that the approach is highly flexible, and one can apply their twist to any problem. While Agile management has some complex and non-negotiable rules, it is not an all-

encompassing doctrine (Al-Saqqa et al., 2020). Figure 3.2 shows the process in the agile methodology.

Figure 3.2 Agile Methodology

3.6.1 Plan

During the planning stage of the research thesis, the researcher, in collaboration with the supervisor, defined the research objectives. The scope of the research was carefully determined, taking into consideration the challenges faced by e-commerce companies in Kenya regarding customer churn. Information regarding the industry and its specific challenges was gathered through document reviews and literature research.

3.6.2 Design

In the design stage, user stories were created to capture the requirements of the customer churn prediction tool. These user stories were prioritized based on their importance and feasibility. The researcher collaborated with the supervisor to create a backlog of tasks, organizing them in a logical order.

3.6.3 Develop

To develop the customer churn prediction tool, the researcher utilized secondary data from Kaggle, a platform hosting dataset relevant to the e-commerce domain. This data was carefully preprocessed to ensure its quality and integrity. Deep learning models, employing frameworks such as TensorFlow or PyTorch, were implemented. Techniques like recurrent neural networks (RNNs) or convolutional neural networks (CNNs) were employed to design accurate

prediction models. The researcher worked iteratively to refine and fine-tune these models, incorporating feedback and insights from the supervisor.

3.6.4 Test

The test stage involved rigorous evaluation of the developed solution. Continuous integration and automated testing were employed to ensure the reliability and stability of the customer churn prediction tool. The researcher conducted thorough testing using historical churn data to assess the performance and generalizability of the models. Adjustments were made as needed to enhance the accuracy and effectiveness of the models.

3.6.5 Deploy

Upon successful testing, the churn prediction tool was deployed for practical use. The researcher and supervisor collaborated to assess its suitability and compatibility with the e-commerce industry in Kenya, ensuring it met the desired objectives.

3.6.6 Continuous Iteration

Throughout the study, the researcher and supervisor engaged in regular discussions and interactions to review progress, address challenges, and make necessary iterations. Feedback from the supervisor was carefully considered and incorporated into subsequent stages of the research to ensure the research thesis achieved its objectives.

3.7 Utilization and Dissemination of Research Results

In terms of application, the research findings are significant in developing a customer churn prediction model that meets the practical needs of businesses and is helpful for enterprise customer relationship management. According to the importance of consumer characteristics, companies can gain insight into the causes of customer churn. Based on the results of the churn model, this research can also help B2C e-commerce marketing managers optimize enterprise marketing strategies and retain customers through commodity promotion activities.

Research results will be published on Strathmore's University digital repository for access to researchers interested in customer churn prediction. This research will open more possibilities for researchers and academicians to improve how churn prediction for non-contractual settings. E-commerce players in Kenya, such as Jumia, can use these research findings to predict the likelihood of churn and stay ahead of the competition.

3.8 Ethical Considerations/Issues

The university's Code of Ethics guided the research process. The researcher always referred to the code whenever necessary. The researcher stored all data in a protected account for privacy and security reasons.

3.9 Conclusion

The Agile methodology was applied to the study, with the researcher and supervisor forming the project team. Secondary data from Kaggle was utilized for model development, and interactions primarily occurred between the researcher and supervisor rather than directly engaging with e-commerce companies. The stages of planning, design, development, testing, deployment, and continuous iteration were followed to successfully create a B2C customer churn prediction tool using deep learning for an e-commerce company operating in Kenya.

Chapter 4: systems analysis and design

4.1 Introduction

This chapter explains the system's overall architecture and detailed design while considering various requirements. UML diagrams were used to describe the overall architecture of the system, provide detailed descriptions of the system's components, and illustrate the interaction between the user and the system's components. To accomplish this, various design diagrams were created, including a representation of the system architecture, a use case diagram followed by detailed use case descriptions, a sequence diagram, a context diagram, a partial domain diagram, and a database design schema.

4.2 Requirement Specifications

The goal of this study was to create a model to help e-commerce businesses predict customer churn. Based on this goal, this section outlines the various requirements that the proposed system must meet. The system requirements for the customer churn prediction tool were obtained through a document review approach. This approach involved a thorough examination and analysis of relevant documents, such as business reports, customer feedback, historical data, and existing literature related to customer churn prediction.

During the document review process, various sources of information were considered to gather insights into the specific requirements of the tool. Business reports provided valuable information about the organization's goals, objectives, and key performance indicators (KPIs) related to customer retention. Customer feedback, such as complaints, surveys, and testimonials, offered valuable insights into the factors that contribute to customer churn and the areas that need improvement.

In addition to internal documents, existing literature on customer churn prediction was thoroughly reviewed. This involved studying research papers, industry publications, and case studies related to customer churn prediction models and techniques. The insights derived from the literature review were used to inform and validate the system requirements. By adopting a

document review approach, the researcher was able to gather a comprehensive understanding of the requirements for the customer churn prediction tool. This method ensured that the system requirements were grounded in real-world data, customer feedback, and industry best practices, ultimately leading to the development of an effective and accurate tool for predicting customer churn.

4.2.1 Functional Requirements

The functional requirements describe the system's behaviour as it relates to its functionality. The system should be able to:

i). Allow users to register.
ii). Allow users to login.
iii). Allow users to upload customer data.
iv). Predict customer churn.
v). Rank customers in based on the risk categories (High Risk, Medium Risk, Low Risk).
vi). Allow users to view their profiles.

4.2.2 Non-Functional Requirements

Non-functional requirements define the software's quality. These requirements define the system's general characteristics, behaviour, and features that affect the user's experience. They provide a better user experience while lowering costs. Non-functional requirements ensure that the software system complies with all applicable laws and regulations. Non-functional requirements have no effect on the system's functionality, but they do affect how it performs. At least some of the non-functional requirements must be met for a well-performing product.

i). Performance Requirement

The system should have a fast response to user's queries.

ii). Compatibility Requirement

The system should be compatible with all the existing browsers both old versions and newer versions.

iii). Security Requirement

The system should be secure from external attacks such as hackers.

iv). Usability Requirement

The system should be easy to use and operate for even those with basic knowledge of how to use systems.

v). Reliability Requirement

The system should provide reliable results to the users.

4.3 System Architecture

System architecture is a conceptual model that defines the structure, behaviour, and more views of a system. It details all the components and subcomponents of the system. Figure 4.1 shows the architecture of Customer Prediction Tool. The architecture consists of the model, API gateway and the web facing consumer application. The user is required to upload customer data in the system. The system will parse the data to the model via an API gateway and the after the prediction is complete, the results are returned to the user. The customer data will then be stored in the database for later access and analysis.

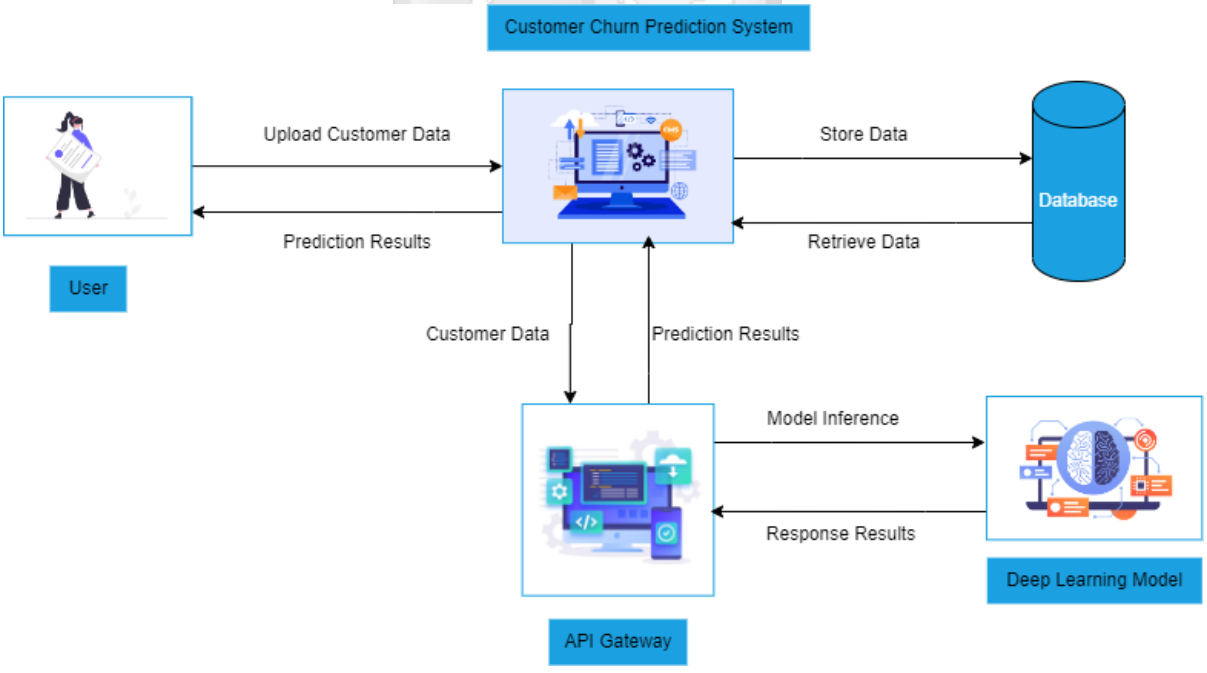


Figure 4.1 System Architecture

4.4 Diagrammatic Representation of the System

Object-Oriented Analysis and Design (OOAD) methodology was employed in design to ensure a systematic and efficient approach to designing the system. This methodology facilitated the creation of comprehensive and robust designs by leveraging the principles of object-oriented programming and modeling. The initial step involved defining the use cases, which were carefully identified and documented to capture the system's intended functionalities and interactions with its users. Use cases served as the foundation for understanding user requirements and forming a clear understanding of the system's behavior in various scenarios.

Following the identification of use cases, sequence diagrams were created to depict the dynamic interactions between the system components. These diagrams provided a visual representation of the flow of control and the sequence of messages exchanged among objects during different use cases. By visualizing these interactions, potential bottlenecks and inefficiencies were identified early on, enabling timely optimizations and improvements to be made. Sequence diagrams, class diagrams were designed to illustrate the static structure of the system. These diagrams showcased the various classes, their attributes, methods, and associations, highlighting the relationships and dependencies between objects. Class diagrams facilitated a deeper understanding of the system's architecture and aided in designing cohesive and modular components.

To ensure efficient data management, a well-defined database schema was crafted. The schema depicted the structure and organization of the system's data, including tables, fields, relationships, and constraints. By designing the database schema upfront, data integrity, consistency, and performance were addressed from the early stages, allowing for effective data storage and retrieval. Moreover, wireframes were created to outline the visual representation and user interface of the system. These visual prototypes showcased the arrangement of elements, navigation flows, and overall user experience. By employing wireframes, the design

team and stakeholders gained a clear understanding of the system's look and feel, enabling valuable feedback and iterative improvements.

4.4.1 Use Case

A use case diagram is a type of behavioural UML diagram that is commonly used to analyze various systems. They make it possible to visualize the various types of roles in a system as well as how those roles interact with the system. Figure 4.1 shows the Use Case Diagram for the customer churn prediction tool.

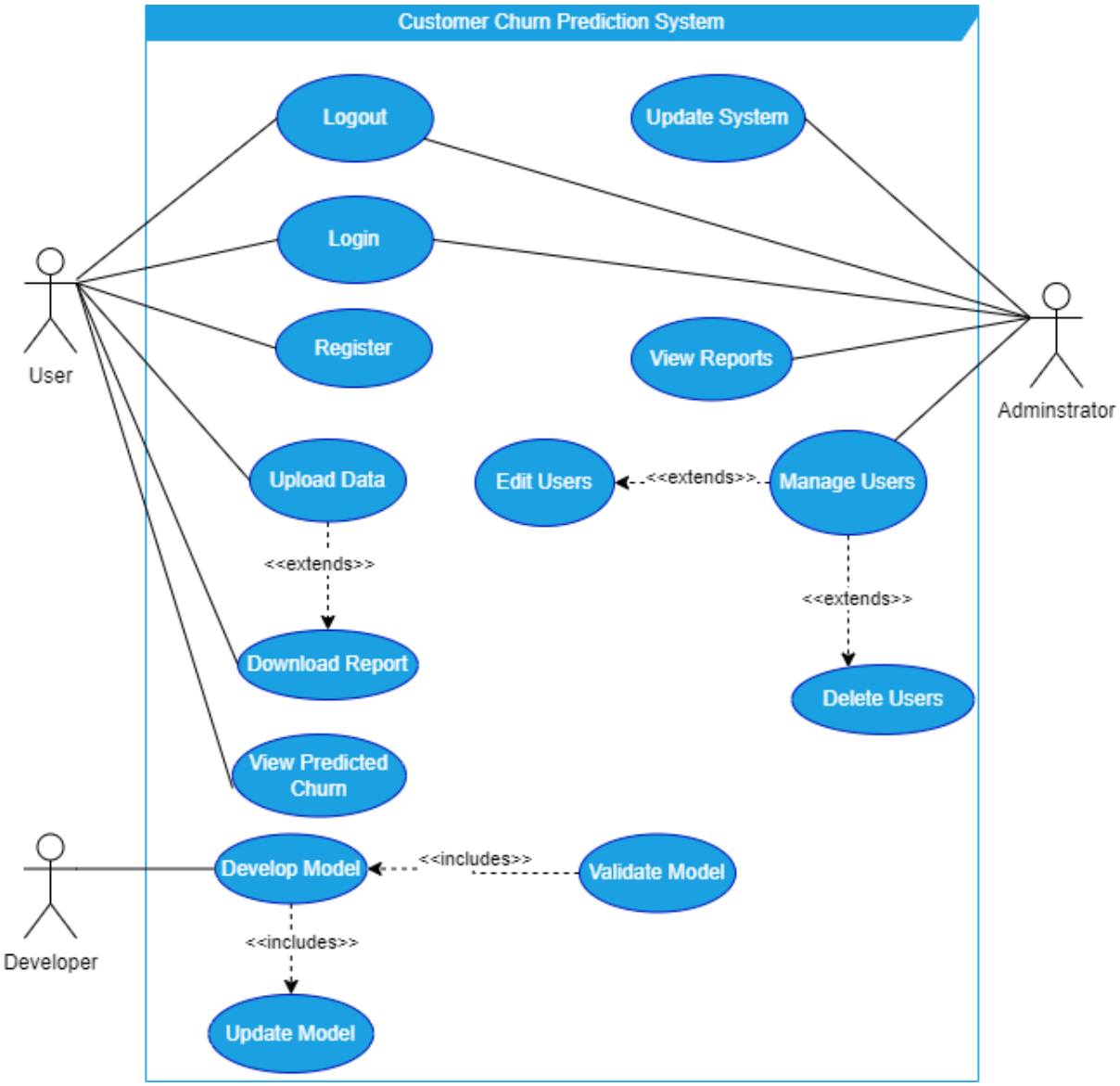


Figure 4.2 Use Case Diagram.

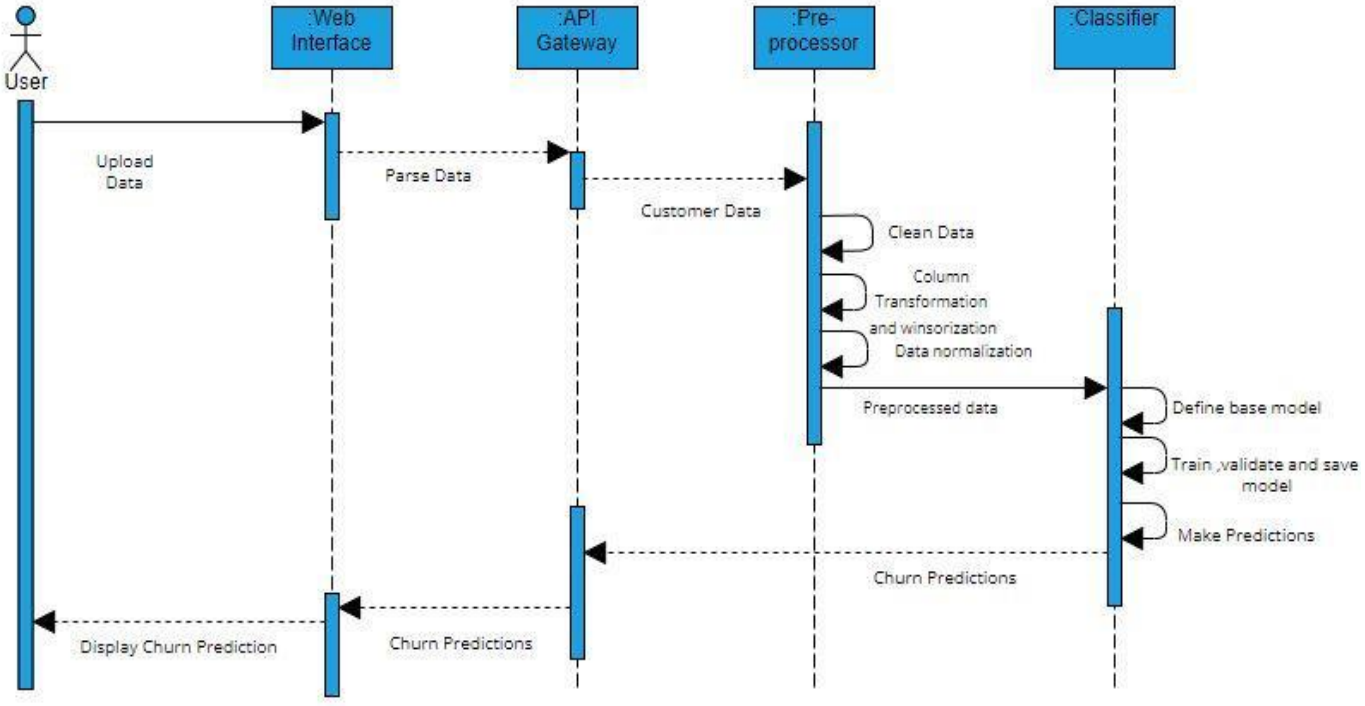
4.4.1.1 Detailed Use Case Diagram

The detailed explanation of the various use cases of the system are explained in Table 4.1 below.

Use Case	Preconditions	Main Success Scenario	Post Conditions
Registration	None	i). The fills in the registration form. ii). The system saves user details to the database.	None
Login	User is Registered	-The user is logged in two the system	None
Upload Data	The user is logged in to the system.	i). The user uploads customer data in csv format. ii). The system displays customer churn profiles to the user.	None
Download Report	-The user is logged in to the system. -The user has uploaded customer data in the system.	i). The user downloads customer churn report	-Customer churn report is downloaded
View Predicted Churn	The user must be logged in	i). Customer views predicted churn from uploaded data.	The user logs out of the system.
Manage Users	Admin must be logged in	i). Administrator views list of registered users in the system ii). The administrator edits /deletes a user from the system. iii). System updates the changes.	None
Update Model	Developer should understand the existing model	i. There will be improved performance. ii. Adaptability to new data iii. Increased efficiency	Changes Documentation and checking model stability

4.4.2 Sequence Diagram

A sequence diagram is a type of diagram that shows the interactions between objects or components in a system and the order in which these interactions take place. It is typically used to model the dynamic behaviour of a system and can help to understand and analyze the requirements of a system. Figure 4.3 depicts the sequence diagram of the proposed system.



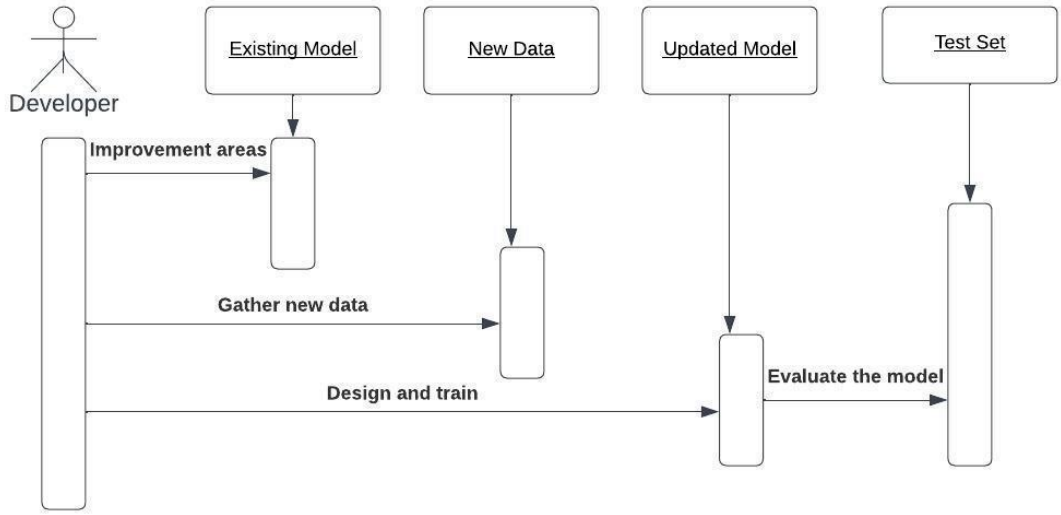


Figure 4.3 Sequence Diagram.

4.4.3 Class Diagram

A class diagram is a type of structural diagram that depicts the structure of a system or software application by showing the classes, their attributes, methods, and the relationships between them. The customer churn prediction tool consists of several classes as shown in Figure 4.4

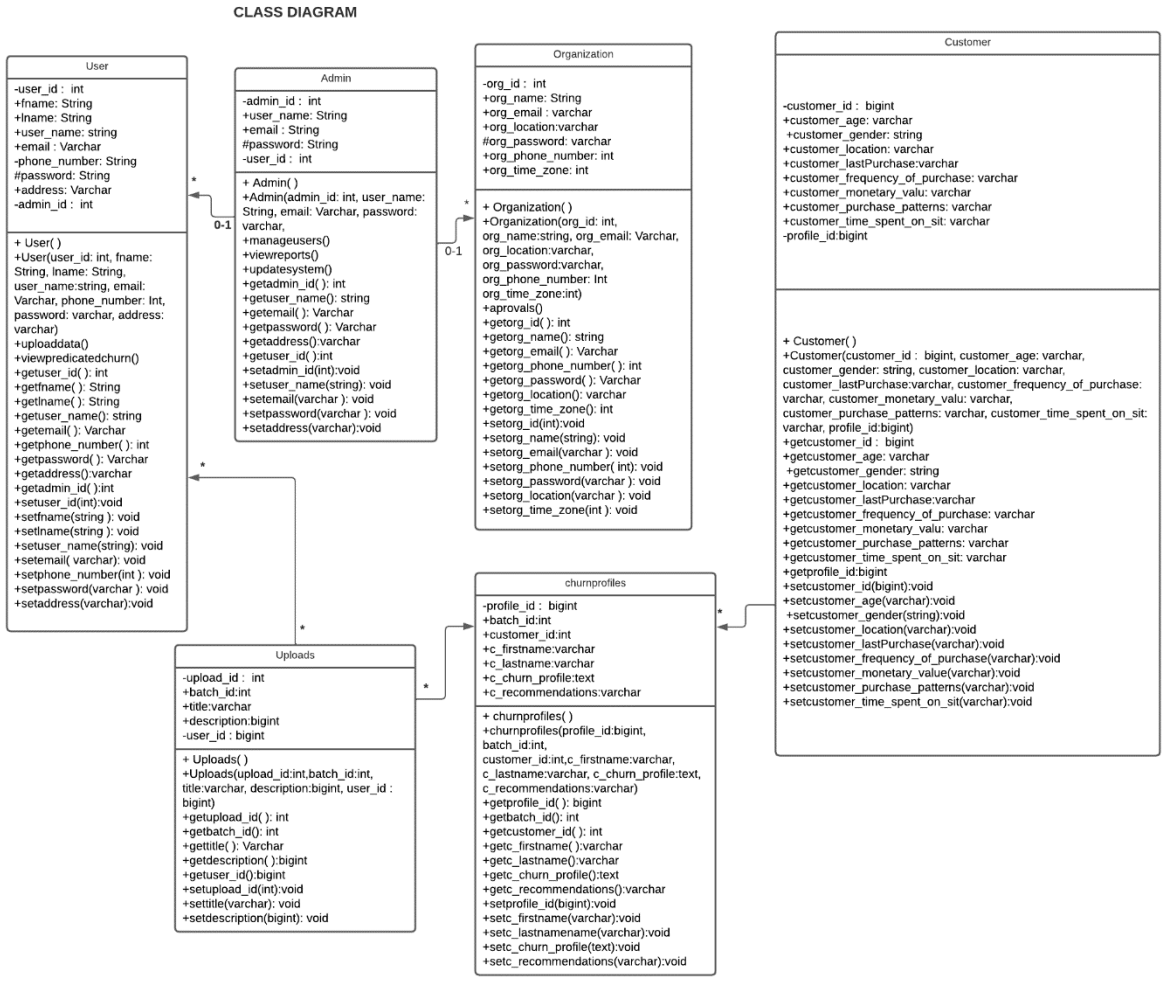


Figure 4.4 Class Diagram.

4.4.4 Database Schema

The database schema for a customer churn prediction tool is designed to store and manage the various data points required for the prediction of customer churn. The schema is composed of several tables that store information related to customers, user accounts, machine learning models, and predictions as illustrated in Figure 4.5.

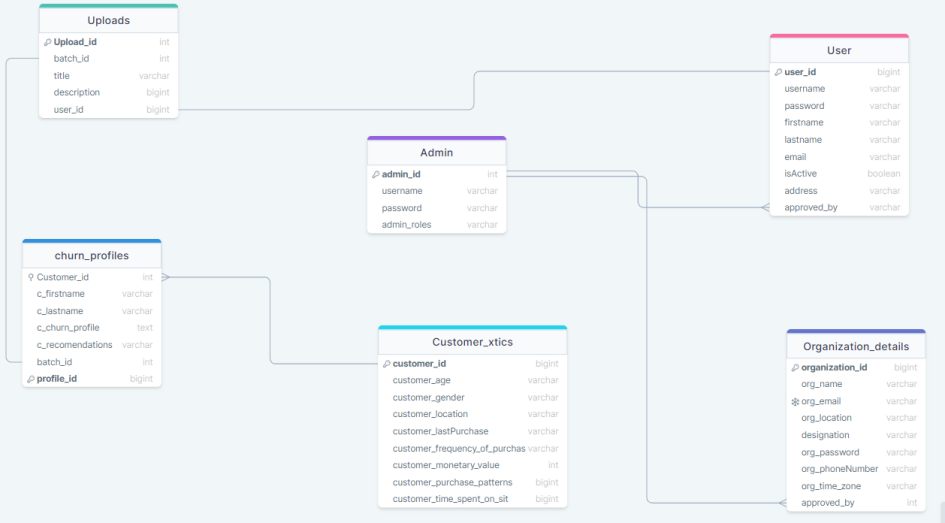


Figure 4.5 Database Schema.

4.5 Wireframes

4.5.1 Home Page

The home page of the customer churn prediction tool provides a brief overview of the tool's features and benefits. It includes a call-to-action button for companies to register and start using the tool. The page is designed with a clean and simple layout that highlights the tool's most important features and benefits. This is depicted in Figure 4.6.

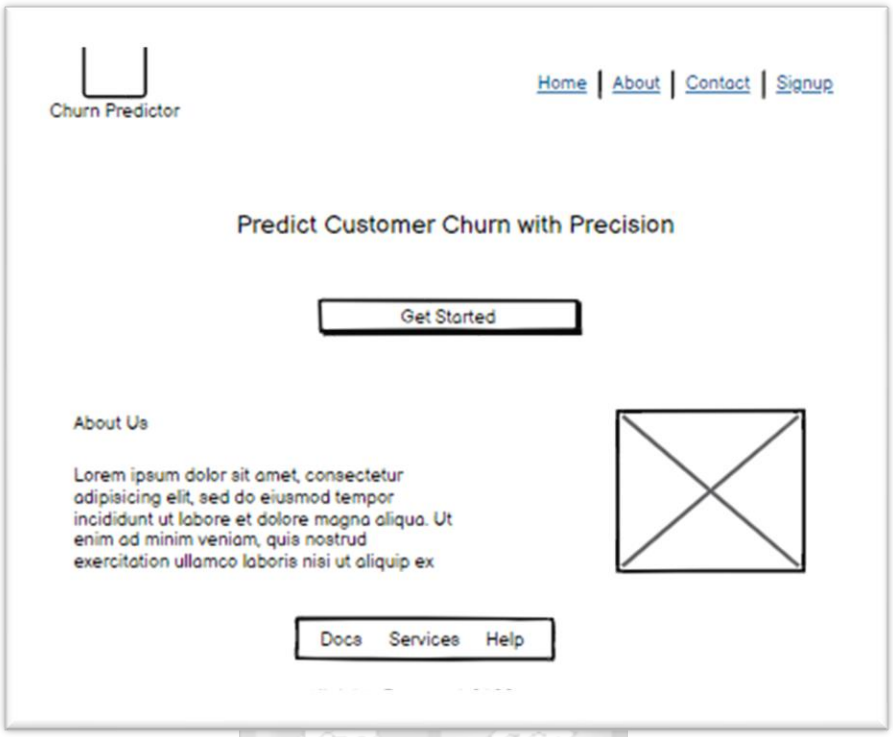


Figure 4.6 Home Page Wireframe.

4.5.2 Register Page

The register allows companies to sign up for the tool. It requires companies to enter their basic information, such as their name, email address, and company name. The register page is designed with a clean and simple layout that allows companies to quickly and easily sign up for the tool.

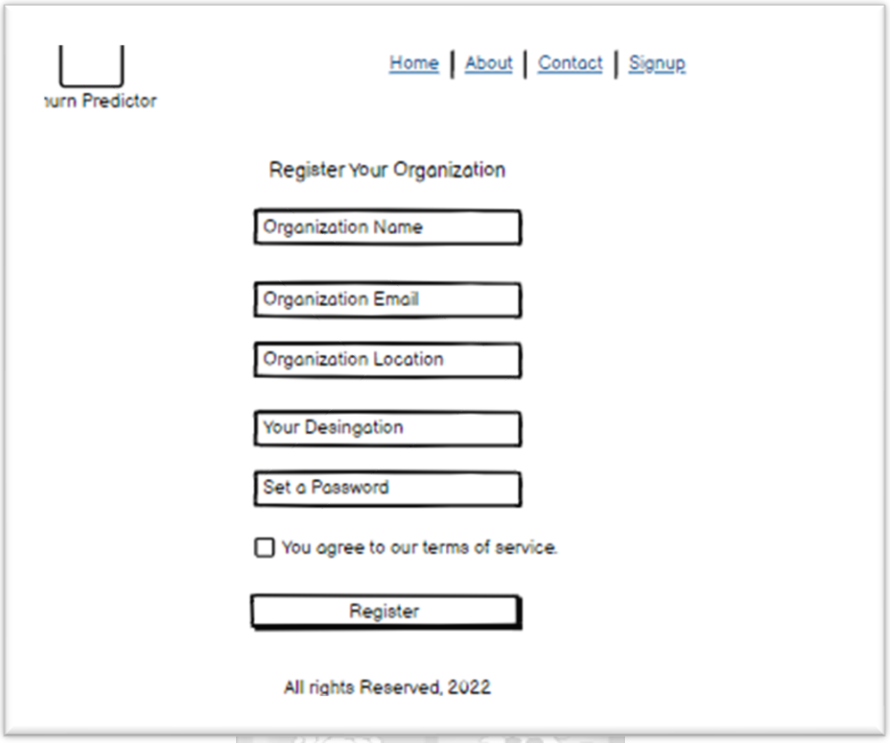


Figure 4.7 Register Page Wireframe

4.5.3 Login Page

The login page is the first page that companies see after they have registered for the tool. The page requires companies to enter their login credentials to access the tool. The login page is designed with a clean and simple layout that allows companies to access the tool quickly and easily.

Figure 4.8 Login Page Wireframe

4.5.4 Upload Page

The upload customer data page is the page where companies can upload their customer data into the tool. It is designed with a clean and simple layout that allows companies to upload their customer data quickly and easily.

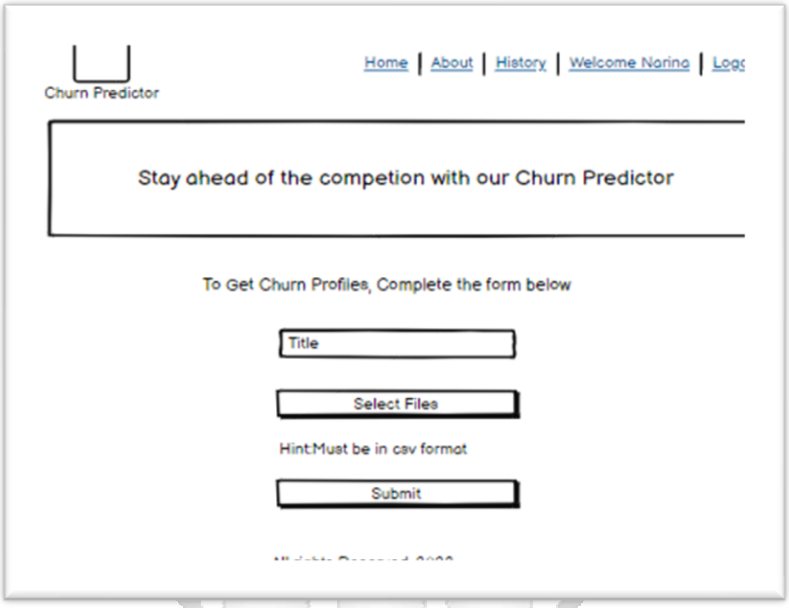


Figure 4.9 Upload Page Wireframe.

4.5.5 Customer Churn Profile Page

The customer churn profile page is the wireframe where companies can view their customers' churn profiles. It provides a detailed overview of each customer's churn profile, including churn probability.

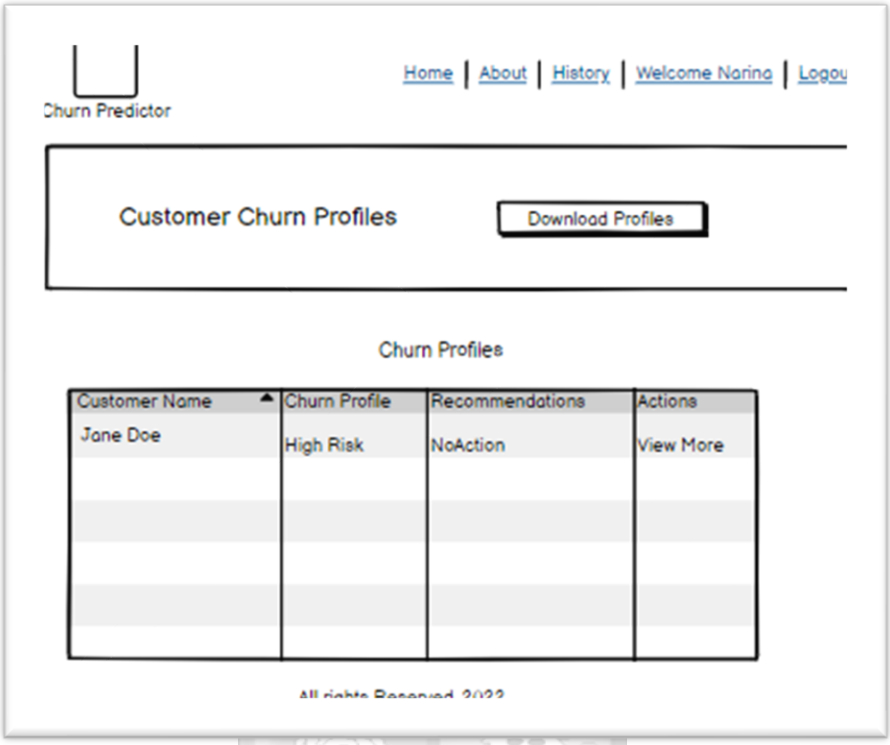


Figure 4.10 Churn Profile Page Wireframe

4.6 Conclusion

Through the use of UML diagrams, a clear depiction of the system's architecture was presented, accompanied by detailed descriptions of its components and the interactions between the user and these components. The design phase involved the creation of various design diagrams, each serving a specific purpose. The system architecture diagram provided a high-level representation of the system's structure, offering insights into the key components and their relationships. The use case diagram, followed by detailed use case descriptions, outlined the different functionalities and actions that the system supports, enabling a thorough understanding of its intended behavior.

Additionally, the sequence diagram provided a dynamic view of the system, illustrating the sequence of actions and interactions between the user and the system's components during specific scenarios. The context diagram offered a broader perspective by showcasing the

system's interactions with external entities and systems. The partial domain diagram facilitated a deeper understanding of the system's specific domain, identifying key concepts and their relationships. Finally, the database design schema provided a blueprint for organizing and structuring the system's data, ensuring efficient storage and retrieval of information.

Chapter 5: System Testing and Implementation

5.1 Introduction

This chapter describes how customer churn prediction model was developed. The chapter also covers how the model was tested and validated. The process of obtaining customer datasets from Kaggle was the first step in the process of implementing the algorithm. The second step involved building the model using python, Jupiter notebook and deep learning algorithms. The final section of this chapter describes the use of the model in predicting customer churn when a user inputs features.

5.2 System Implementation

The implementation of the customer churn model using deep learning consisted of numerous crucial processes, including data collection and preprocessing, model selection and architecture design, training and validation, deployment, and monitoring. To reliably predict customer churn using deep learning, it was necessary to collect and clean important data before designing and training a deep neural network capable of accurately classifying consumers as churning or non-churning.

5.2.1 Loading Dataset

This research data was obtained from Kaggle, a well-known online platform for data science and machine learning. The dataset was specifically obtained from the following link: [Olist Brazilian E-Commerce Dataset (https://www.kaggle.com/datasets/olistbr/brazilian-e-commerce) is available for download. Kaggle offers a variety of datasets contributed by the data science community, making it a valuable resource for data science researchers, practitioners, and enthusiasts.

The Brazilian E-Commerce dataset from Olist provides exhaustive data regarding online sales transactions in Brazil, including customer demographics, product specifics, order status, and seller performance. Researchers and data analysts can gain insights into the Brazilian e-

commerce market and investigate various aspects of consumer behavior, seller dynamics, and market trends by utilizing this dataset.

The following figure shows data read from a drive. It also shows the information in the specific excel where def data_info(df) function prints the first 5 rows of the dataset, shape, columns, datatypes, and number of unique values.

```
# read the dataset
path = "/content/drive/MyDrive/Customer Churn/customer-churn/"
df = pd.read_excel(path+'ECommerceDataset.xlsx', sheet_name='E Comm')

# information regarding the dataset
def data_info(df):
    """
    This function prints the first 5 rows of the dataset, the shape of the dataset,
    """
    print("First 5 rows:")
    print(df.head())
    print("Shape: ")
    print(df.shape)
    print("Columns:")
    print(df.columns)
    print("Data types: ")
    print(df.dtypes)
    print("Number of unique values: ")
    print(df.nunique())

  # call the function
data_info(df)
```

Figure 5.1 Loading Dataset

5.2.2 Cleaning Data

The following figure checks for the missing values in the dataset. (def missing _vales(df):) function prints the number of missing values in each column.

The def impute_missing_values function imputes the missing values in the dataset and calls the function.

```
# check for missing values
def missing_values(df):
    """
    This function prints the number of missing values in each column.
    """
    print("Missing values: ")
    for col in df.columns:
        if df[col].isnull().sum() > 0:
            print(col, ":", df[col].isnull().sum())

# call the function
missing_values(df)
```

Figure 5.2 Cleaning Data

```
# Impute missing values
def impute_missing_values(df):
    """
    This function imputes the missing values in the dataset.
    """
    for col in df.columns:
        if df[col].dtype == 'int64' or df[col].dtype == 'float64':
            df[col].fillna(df[col].mean(), inplace=True)
        else:
            df[col].fillna(df[col].mode()[0], inplace=True)
    return df

# call the function
df = impute_missing_values(df)
```

Figure 5.3 Cleaning Dataset

5.2.2 Data Preprocessing

In the data preprocessing phase of developing a customer churn prediction model using deep learning, several techniques were applied to ensure the data was properly formatted, free from inconsistencies, and optimized for training. The following techniques were employed:

First, data normalization was performed to rescale numerical features to a standardized range, typically between 0 and 1 or -1 and 1. This step prevented features with larger magnitudes from dominating the training process and ensured equal contribution from all features. Next, categorical encoding was used to transform categorical variables into numerical representations suitable for deep learning models. One-hot encoding or label encoding techniques were employed to achieve this conversion. One-hot encoding converted each category into a binary vector, while label encoding assigned a unique numeric label to each category.

To address class imbalance, which often occurs in customer churn prediction models, a thorough check was conducted. Techniques such as oversampling or undersampling were applied to balance the dataset, ensuring that the model was not biased towards the majority class. Another important step was handling outliers. Outliers, defined as data points significantly deviating from the norm, were identified using statistical methods such as z-score or interquartile range. These outliers were then managed by either removing them, replacing them with appropriate values (e.g., mean or median), or transforming them to minimize their impact on the model.

Furthermore, specific transformations were applied to certain columns to improve data quality. For example, feature scaling, such as standardization, was performed on numerical columns to achieve a mean of 0 and a standard deviation of 1. Skewed features were transformed using logarithmic transformations to obtain a more symmetric distribution. By applying these preprocessing techniques, the input data for the customer churn prediction model was prepared in a suitable format for deep learning. The data was standardized, categorical variables were

represented numerically, class imbalances were addressed, outliers were managed, and column transformations were performed to optimize the model's training process and enhance its predictive capabilities.

```
from sklearn.preprocessing import MinMaxScaler

# normalize the numerical columns
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# save y_train and y_test for mlp
y_train_mlp = y_train.copy()
y_test_mlp = y_test.copy()

# one hot encode the categorical columns # for deep learning
y_train = to_categorical(y_train)
y_test = to_categorical(y_test)
```

Figure 5.4 Data Preprocessing

5.2.3 Feature Selection

In order to predict customer churn, a predictive model was developed using various features. These features were carefully selected to provide valuable insights into customer behavior and preferences. Features Used in Customer Churn Prediction Model:

i). Tenure
ii). WarehouseToHome
iii). CashbackAmount
iv). NumberOfDeviceRegistered
v). SatisfactionScore
vi). NumberOfAddress
vii). Complain

viii). OrderAmountHikeFromlastYear

ix). CouponUsed

x). DaySinceLastOrder.

The chosen features provide valuable insights into different aspects of customer behavior and preferences. For instance, the tenure feature captures the duration of the customer's association with the company, helping assess the impact of loyalty on churn likelihood. The warehouse-to-home distance feature considers the logistical aspects of product delivery, which can influence customer satisfaction and subsequent churn decisions. Cashback amount reflects the incentives provided to customers and investigates its effect on churn behavior.

Furthermore, the number of registered devices reveals the extent of customer engagement and usage across multiple devices. Satisfaction score, obtained through surveys or feedback mechanisms, offers a quantifiable measure of customer satisfaction. The number of addresses indicates customer diversity and potential changes in location. Complaint history identifies customers who have expressed dissatisfaction, potentially indicating an increased likelihood of churn.

Additionally, the order amount hike from the previous year measures changes in customer spending patterns and their impact on churn propensity. Coupon usage highlights customers who are price-sensitive and may be more inclined to switch to competitors with better offers. Lastly, the day since the last order reflects customer engagement and purchase frequency, providing insights into potential churn risks.

5.2.4 Model Training

In the thesis, deep learning techniques were utilized to train a classification model using the Multilayer Perceptron (MLP) algorithm. The code snippet showcases the implementation of this approach. Here's an explanation of how deep learning was used: The code begins by

importing the necessary library, `sklearn.neural_network`, which includes the `MLPClassifier` class for building Multilayer Perceptron models. A function called `train_model` is defined, which serves the purpose of training the deep learning model. This function takes two inputs: `X_train`, representing the training data features, and `y_train_mlp`, which corresponds to the target labels associated with the training data.

Within the `train_model` function, an instance of the `MLPClassifier` is created, configuring the model to have three hidden layers, each consisting of 100 neurons. These hidden layers form the core of the deep learning model, enabling it to learn complex patterns and relationships within the data. The `max_iter` parameter is set to 1000, determining the maximum number of iterations (epochs) the model will undergo during training. This iterative process allows the model to refine its internal parameters, such as weights and biases, to minimize the difference between predicted and actual labels.

To train the model, the `model.fit(X_train, y_train_mlp)` line is executed. This step initiates the training process, where the MLP model adjusts its internal parameters based on the provided training data. Through forward and backward propagation, the model iteratively updates its weights and biases to improve its predictive accuracy. Finally, the trained MLP model is returned from the `train_model` function, ready for further evaluation and prediction tasks.

In summary, the code snippet demonstrates how deep learning techniques, specifically the Multilayer Perceptron algorithm, were employed in the thesis. By utilizing the MLPClassifier from the scikit-learn library, a deep learning model was trained on the provided training data to learn complex patterns and relationships, paving the way for accurate classification predictions. Figure 5.6 depicts the MLPClassifier.

```python
from sklearn.neural_network import MLPClassifier
# train the model using Multilayer Perceptron
def train_model(X_train, y_train_mlp):
    """
    This function trains the base model using Multilayer Perceptron.
    """
    model = MLPClassifier(hidden_layer_sizes=(100, 100, 100), max_iter=1000)
    model.fit(X_train, y_train_mlp)
    return model
```

Figure 5.6 Model Training

A sequential model was also used as shown in 5.7. The code includes a function called `train_deep_model`, which is responsible for training the base model using deep learning techniques. This function takes three inputs: `X_train` (the training data features), `y_train` (the corresponding target labels for the training data), and an optional parameter `epochs` which specifies the number of training epochs (default set to 3). Within the function, a `Sequential` model is initialized. The `Sequential` model is a linear stack of layers used in deep learning architectures.

The model is built by adding layers one by one using the `model.add()` function. The first layer added is a dense layer with 64 neurons, specifying the `input_shape` as the number of features in the training data. The activation function used in this layer is ReLU (Rectified Linear Unit), which introduces non-linearity into the model. Subsequently, additional dense layers are added to the model. These layers contain 32, 16, and 8 neurons, respectively, with ReLU activation functions. The addition of multiple dense layers allows the model to learn increasingly complex representations of the data.

The final dense layer added has 2 neurons and uses a sigmoid activation function. This layer is responsible for the binary classification output, where each neuron represents one of the two possible classes. The sigmoid activation function ensures that the output values

are in the range of [0, 1], providing probabilities for each class. The model is then compiled using the `model.compile()` function. The loss function chosen is binary cross-entropy, suitable for binary classification tasks. The optimizer used is Adam, a popular optimization algorithm in deep learning. Additionally, evaluation metrics such as accuracy, precision, recall, and F1 score are specified.

The training process is initiated by calling `model.fit()`, passing the training data (`X_train` and `y_train`). The `epochs` parameter determines the number of times the model will iterate over the entire training dataset during training. The `batch_size` parameter determines the number of samples processed before the model's internal parameters are updated. The `verbose` parameter is set to 1, indicating that training progress will be displayed.

```
# train a Deeplearning model
def traind_deep_model(X_train, y_train, epochs=3):
    """
    This function trains the base model using Deep Learning MLP.
    """
    model = Sequential()
    # print(model)
    model.add(Dense(64, input_shape=(X_train.shape[1],), activation='relu'))
    # print(model)
    model.add(Dense(32, activation='relu'))
    model.add(Dense(16, activation='relu'))
    # print(model)
    model.add(Dense(8, activation='relu'))
    # print(model)
    model.add(Dense(2, activation='sigmoid'))
    # model.add(Dense(y_train.shape, activation='softmax'))
    model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy', precision, recall, f1])
    model.fit(X_train, y_train, epochs=epochs, batch_size=32, verbose=0)
    return model
```

Figure 5.7 Model Training

```
# call the function
# base model
model = train_model(X_train, y_train_mlp)

# deep learning model
model_deep = traind_deep_model(X_train, y_train, epochs=2)
```

Figure 5.8 Model Training

5.2.5 Model Validation

```
# evaluate deep learning model
def evaluate_deep_model(y_test, y_pred):
    """
    This function evaluates the model using accuracy score, confusion matrix, and classification report.
    """
    # evaluate the model on the testing data
    loss, accuracy, f1_score, precision, recall = model_deep.evaluate(X_test, y_test)
    print("Test Loss: ", loss)
    print("Test Accuracy: ", accuracy)
    print("Test F1-Score: ", f1_score)
    print("Test precision: ", precision)
    print("Test recall: ", recall)
```

Figure 5.9 Model Validation.

5.3 Customer Churn Prediction Interface

Users can access the customer churn prediction model using the login/register portal. To access the prediction model, users will be able to create an account or sign in with their existing credentials. Users will be required to enter their username and password or establish a new account using their email address and other pertinent information. Figures 5.10 and 5.11 show the login and registration interface respectively.

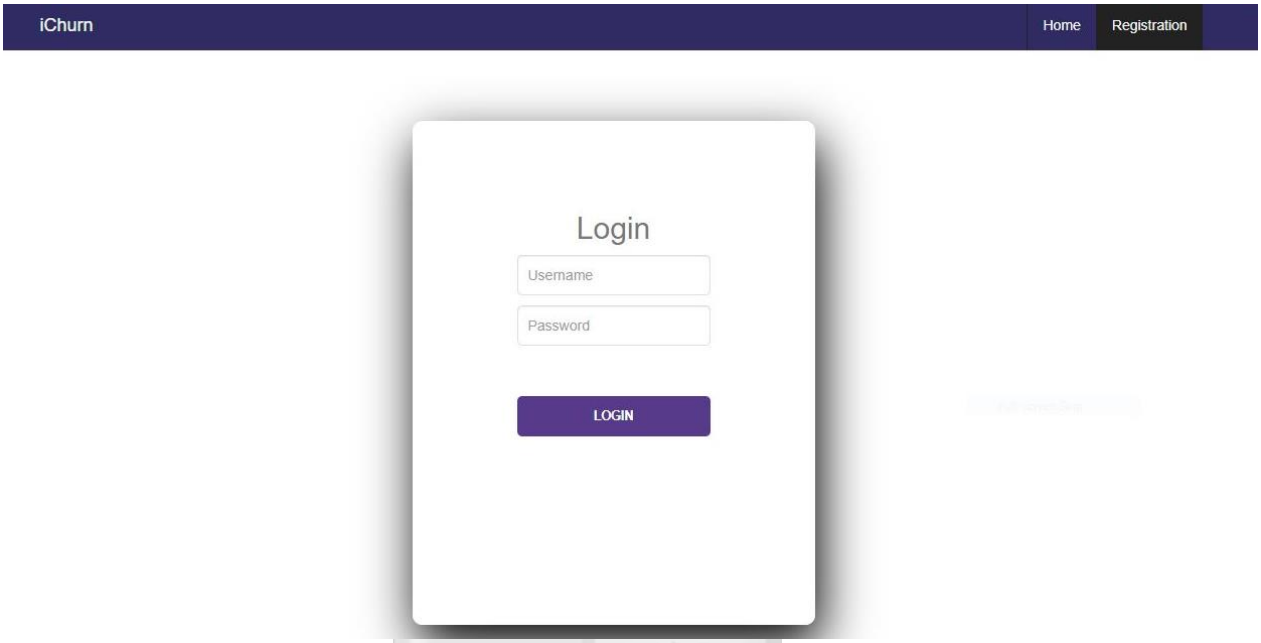


Figure 5.10 Login Interface

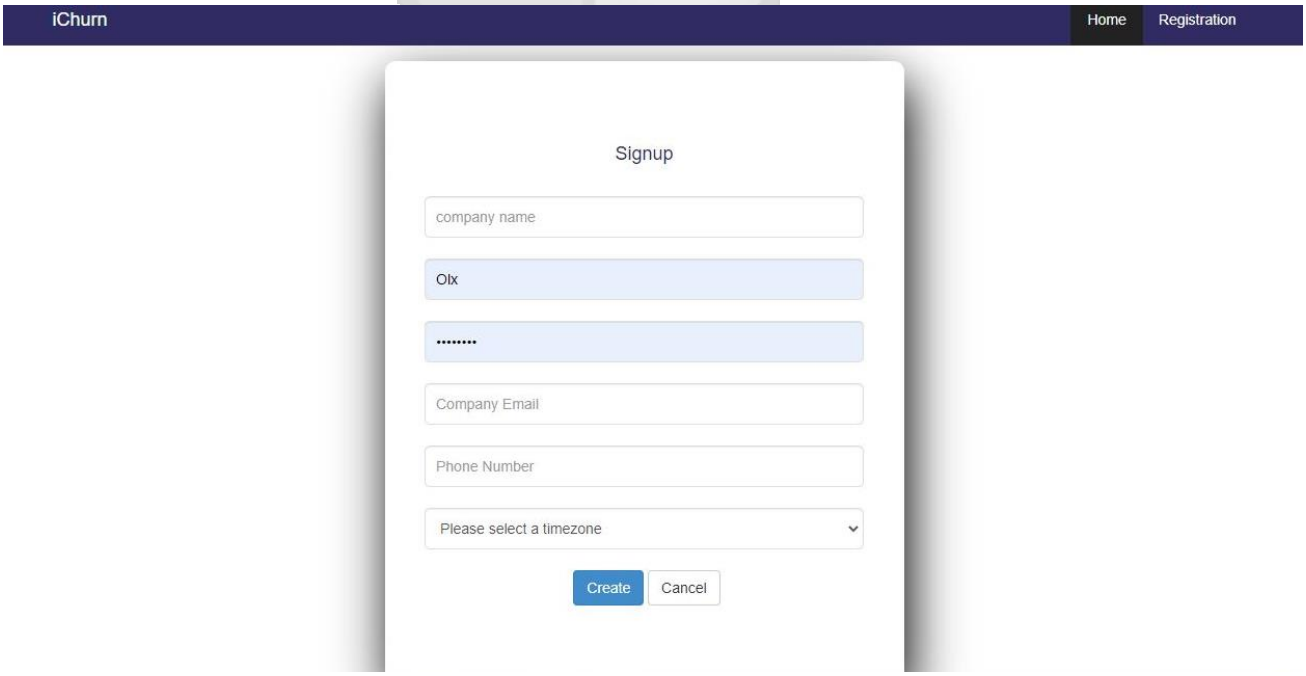


Figure 5.11 Registration Interface

After logging in, users are taken to the dashboard screen. Users will be able to see a summary of their account information, including their history of predictions and any pertinent reports, on this interface. Users will also be able to alter their account details and use the prediction interface as shown in Figure 5.12.

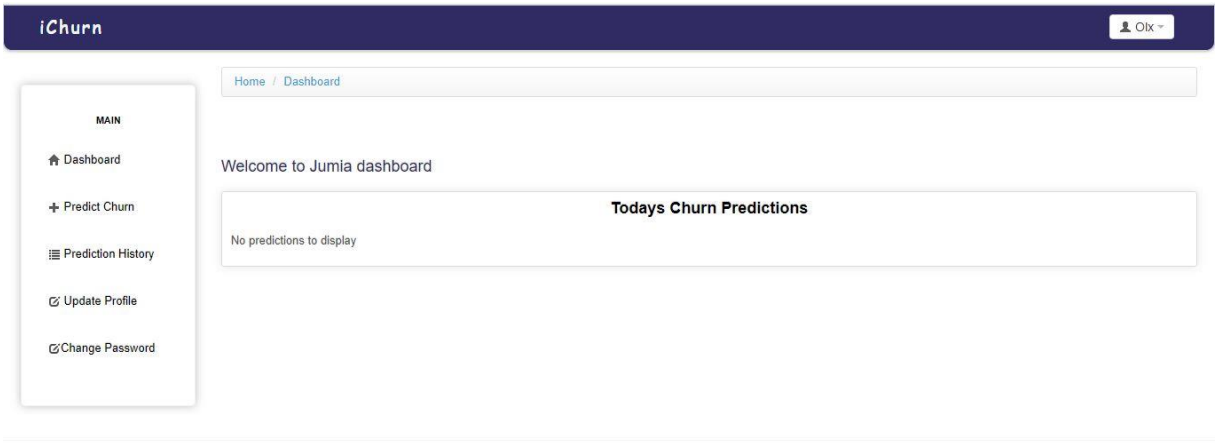


Figure 5.12 Dashboard

prediction interface is where users can input customer data and generate a churn prediction. This interface will require users to input data such as purchase history, website activity, and customer demographics. Once the data has been input, the prediction model will generate a probability of churn for that customer as depicted in Figure 5.13.

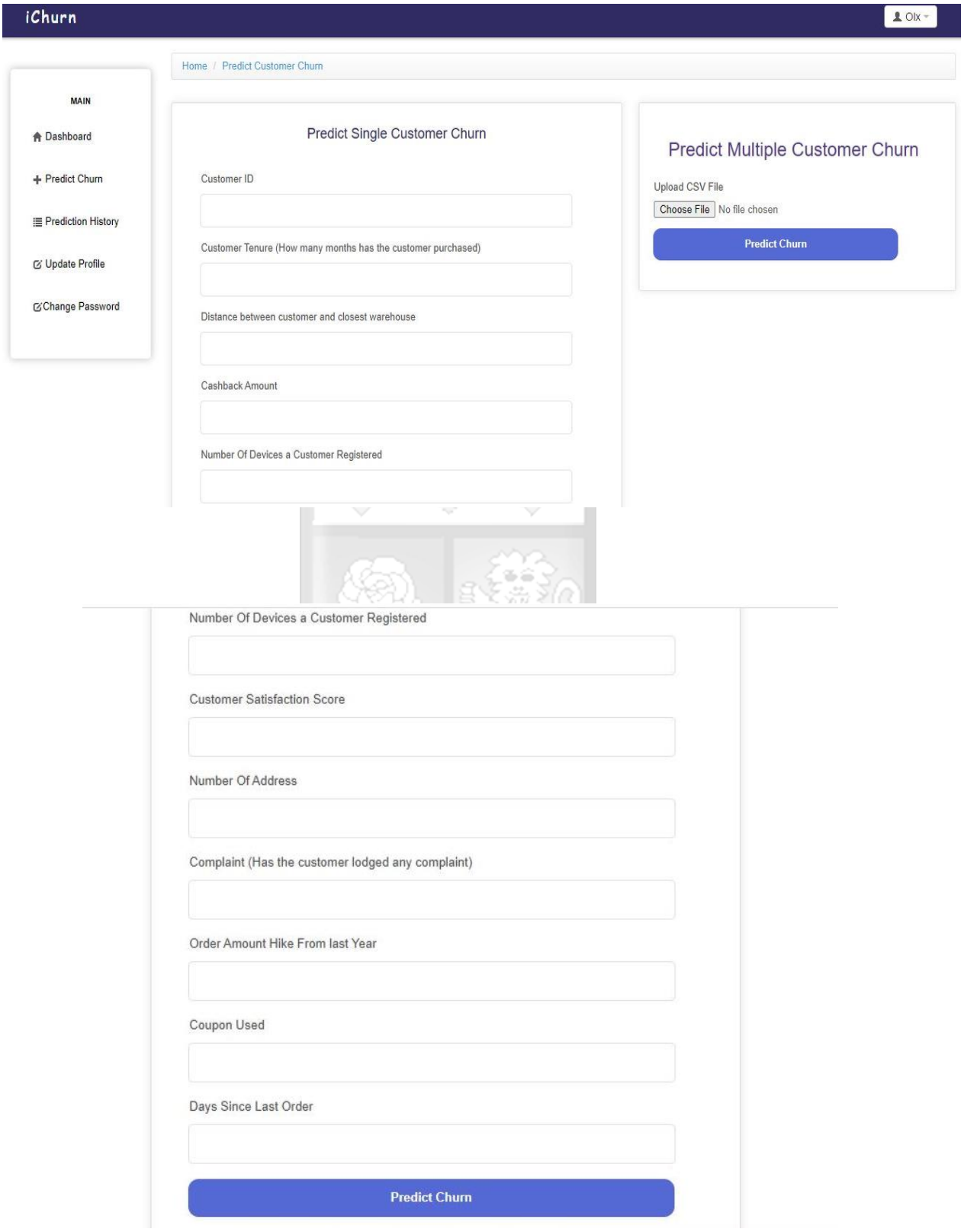


Figure 5.13 Prediction Interface

The prediction history interface shows a list of all the user's past predictions. This interface shows how likely it is that a customer will leave, when the prediction was made, and any other relevant information. The prediction history can also be filtered by a range of dates or other relevant criteria.

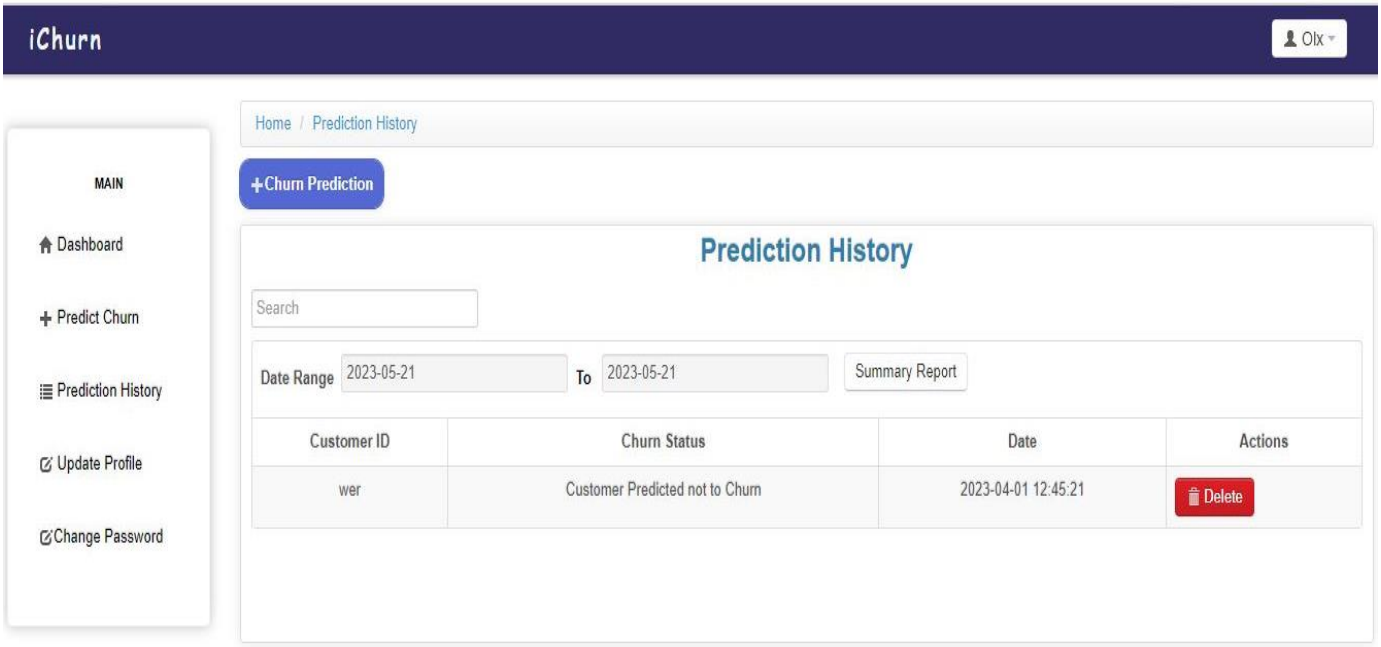


Figure 5.14 Prediction History

5.4 System Testing

The effectiveness of the model was measured by subjecting it to a classification task that mimicked real-world scenarios. This was accomplished by running an evaluation on fresh data, at which point the model successful. The following was how some tests were conducted and the test results:

i). Allow users to register: To test the user registration feature, the following steps were followed:

ii). Manually test the registration form fields and ensure that all required fields are present and functional.

iii). Test the form submission process to ensure that the user data is being correctly validated and stored in the database.

iv). Use automated testing tools to verify that the user registration process is functioning correctly.

Test Results: After conducting these tests, it was possible to see new user accounts in the database with all the required information.

i). Allow users to login: To test the user login feature, the following steps were followed:

ii). Manually test the login form and ensure that all required fields are present and functional.

iii). Test the login process to ensure that the user data is being correctly validated and authenticated against the stored credentials in the database.

iv). Use automated testing tools to verify that the login process is functioning correctly.

Test Results: After conducting these tests, it was possible to see that the user is redirected to a protected dashboard page after successful login.

i). Allow users to upload customer data: To test the feature of allowing users to upload customer data.

- Test the file upload functionality to ensure that the tool is able to accept the file types and sizes specified in the requirements.
- Manually check that the uploaded file is being validated for correctness and completeness, including checking for formatting errors and data type inconsistencies.
- Verify that the uploaded data is being stored correctly in the database.

Test Results: After conducting these tests, the upload functionality worked correctly, it was possible to see the uploaded customer data in the database, ready to be used for churn prediction analysis.

5.5 Testing Model Accuracy

The accuracy of the system in correctly predicting customer churns was looked at, and below are the results per the incidents looked at versus what is the actual state of events. After model testing, the accuracy stood at 94% as show below.

```
# call the function
# base model
evaluate_model(y_test_mlp, y_pred)

Accuracy Score:  0.9448275862068966
```

Figure 5.15 Accuracy Score.

5.4 Model Validation/Deployment

The system was used to validate a user entry of unknown and unseen customer profiles. The results proved that the model could provide high accuracy and precision scores in predicting customer churn.

5.5 Conclusion

The implementation process began by obtaining customer datasets from Kaggle, which served as the foundation for implementing the algorithm. Next, the model was built using Python and Jupiter Notebook, utilizing deep learning algorithms to enhance its predictive capabilities. The developed model was then subjected to rigorous testing and validation to assess its identification accuracy. This involved evaluating the model's performance against known customer churn cases, comparing its predictions with the actual outcomes. Through this testing process, the model's effectiveness in accurately predicting customer churn was assessed and refined.

The final section of this chapter focused on the practical application of the developed model. It highlighted how the model could be utilized to predict customer churn when provided with

relevant features. This practical aspect of the model demonstrates its potential to be integrated into real-world business environments, enabling companies to proactively address customer churn and improve retention rates.

Chapter 6: Discussion

6.1 Investigating the Factors That Lead to Customer Churn in The E-Commerce Industry

This objective was achieved during literature review stage. It was discovered that minimal research had been conducted on the customer churn prediction in ecommerce sector. Majority of the research focused on prediction in the telecom and banking sectors where clients have a contract with the respective businesses. The research discovered that five factors were majorly the cause of customer churn in ecommerce sectors. First, consumers are more likely to opt for the cheaper option when there is no discernible difference between products and services. In addition, the consumer is likely to churn when the designed product does not meet their needs. Furthermore, the churn rate can be affected by user characteristics like consumption level and personal income. Business factor and Service factor also play a bigger role in influencing a customer decision to leave a business.

6.2 Existing Models and Algorithms Used for Customer Churn Prediction

This objective was achieved by reviewing several studies that were undertaken on customer churn prediction in ecommerce churn prediction. In their customer churn prediction research, Wai-Ho Au et al (2003) applied decision trees on a database of 100,000 records provided by a carrier in Malaysia. The model was robust in this classification task based on the contractual setting of the telecom carrier. Xiahou and Harada (2022) loss prediction model combined k-means customer segmentation with support vector machine (SVM) prediction. The method categorizes customers into three groups and identifies the core customer groups. They compared support vector machine and logistic regression and discovered that SVM prediction was more accurate than the logistic regression prediction. However, this study had several limitations. The study only used the K-means algorithm hence the model did not have convincing results. In addition, the study only used a small number of predictive variables, which limits the promotion of results because a lot of shopping information is presented on B2C websites, and some of it may be ignored.

Pondel et al (2021) research created a deep learning model for predicting customer churn in e-commerce. The experiment was run on e-commerce data, with 75% of buyers being one-time customers. The prediction based on this business specificity. Predictions with 74% accuracy, 78% precision, and 68% recall were very promising in this case but still have a long way to giving reliable customer churn predictions. The accuracy and precision levels obtained in this study are much lesser compared to the one achieved using deep learning algorithms.

The developed churn prediction model outperformed all the reviewed models by achieving a top accuracy of 94% in the customer churn prediction. Customer churn is a critical business problem, and predicting customer churn is essential for retaining customers and improving customer satisfaction. SVM, Decision Trees, Random Forest, and Logistic Regression were the commonly used machine learning models in customer churn prediction by researchers. However, these models have shown weaknesses when dealing with large and complex datasets, which is common in customer churn prediction. In contrast, deep learning models used in this study has shown superiority in handling these types of data and have several features that make them particularly suitable for customer churn prediction.

One significant advantage of deep learning models is their ability to learn hierarchical representations of data which came in handy in customer churn prediction since customer behavior is complex and multifaceted. Deep learning models can learn the underlying patterns that drive customer churn, even if these patterns are not immediately obvious to humans. SVM, Decision Trees, Random Forest, and Logistic Regression models do not have this ability, and they are limited to only learning patterns that can be easily recognized and measured.

Deep learning models can also capture complex interactions between features. In customer churn prediction, it is not just the individual features that are important, but also the

interactions between them. Deep learning models can learn these complex interactions, which can lead to better predictions. SVM, Decision Trees, Random Forest, and Logistic Regression models cannot handle these complex interactions and are limited in their ability to capture non-linear relationships between variables. Another advantage of deep learning models is their ability to handle large and complex datasets. Customer churn prediction involves analyzing various customer attributes, such as demographics, usage patterns, and purchase history, which can result in a large and complex dataset. Traditional machine learning models like SVM, Decision Trees, Random Forest, and Logistic Regression models can struggle to handle such datasets. In contrast, deep learning models can handle this complexity and find patterns that are difficult to identify with traditional models.

6.3 Customer Churn Prediction Tool in The E-Commerce Industry Using Deep Learning Techniques

This objective was successfully achieved by developing a powerful tool utilizing deep learning techniques to predict client attrition within the e-commerce sector. To identify significant patterns and indicators of churn, the tool underwent training using extensive datasets encompassing consumer behavior and purchase history. Employing deep learning methodologies, the study incorporated artificial neural networks designed to mimic the information processing capabilities of the human brain. These networks, consisting of interconnected nodes or neurons, collaborated to analyze and classify the data effectively. To create an accurate prediction model for customer attrition, comprehensive data was collected and compiled, encompassing customer demographics, purchasing patterns, website usage, and customer feedback. Through the integration of these elements, the deep learning tool emerged as a reliable solution for forecasting and addressing customer attrition within the e-commerce domain.

6.4 To test the developed tool.

The objective of evaluating the model's real-world performance was successfully achieved through the utilization of a distinct test set. The analysis of the test results revealed compelling

metrics, indicating the deep learning model's accuracy in predicting customer churn. With a precision rate of 94%, a recall rate of 93%, and an impressive F1 score of 94%, the model demonstrated its ability to effectively identify potential churners, validating its reliability and effectiveness in practical scenarios.

Chapter 7: Conclusion and Recommendation

7.1 Conclusion

Customer churn is a significant issue in the B2C ecommerce industry, where competition is fierce and client retention is essential to a company's survival. Predicting client attrition is therefore critical for ecommerce businesses seeking to retain consumers and increase their bottom line. In this study, several machine learning methods, such as decision trees, SVM, neural networks were investigated and compared the results to the deep learning algorithms, for predicting customer churn in the B2C ecommerce industry. In predicting customer attrition, the deep learning model attained the greatest accuracy of 94%, beating previous machine learning methods. The deep learning algorithm is a robust and adaptable model that can learn from intricate patterns and interrelationships in vast datasets. It has been utilized extensively in picture and speech recognition, natural language processing, and other applications of artificial intelligence.

In the B2C ecommerce market, the remarkable accuracy demonstrated by the deep learning algorithm in predicting client attrition has significant consequences for ecommerce companies. Initially, it enables them to identify clients who are likely to churn in advance, allowing them to build effective retention efforts to prevent their departure. These retention techniques may include tailored promotions, discounts, customer loyalty programs, and enhanced customer service. By keeping customers, ecommerce businesses may increase customer loyalty, boost customer lifetime value, and decrease client acquisition expenses.

Second, the deep learning algorithm can assist ecommerce businesses in identifying the primary contributors to client turnover. They may include product quality, pricing, shipping delays, user experience, customer service, and other variables that affect consumer loyalty and happiness. By recognizing these elements, ecommerce organizations may concentrate their investments and resources towards enhancing them, hence enhancing the entire customer experience, and decreasing turnover. Lastly, the algorithm for deep learning can assist ecommerce businesses in segmenting their consumer base and developing retention tactics for

each segment. By segmenting clients based on their demographics, behavior, and preferences, ecommerce organizations may adjust their retention efforts to each segment's individual needs and preferences, hence enhancing the efficacy of these strategies.

This study concludes with significant insights into the usage of deep learning algorithms for predicting customer churn in the B2C ecommerce business. By effectively estimating client turnover, ecommerce businesses may maximize customer retention, customer loyalty, and ultimately their bottom line. We believe that our study will stimulate additional research in this area and assist e-commerce businesses in developing efficient customer retention strategies.

The customer churn prediction tool in the e-commerce industry that uses deep learning techniques is a valuable application of artificial intelligence and machine learning. By using the power of deep learning algorithms, e-commerce businesses can analyze massive amounts of data and find patterns that are predictive of client turnover. This enables businesses to take proactive actions to retain clients, enhance customer happiness, and eventually expand their businesses. Much work is required in data collecting, cleaning, and labelling, as well as model training and optimization, for the development and implementation of such a tool. Yet, there are various advantages to having an accurate client churn forecast tool. With the technology, businesses can customise their marketing tactics, identify clients at danger of churn, and create retention campaigns with a specific focus. In addition, this technology can provide insights into the underlying reasons that influence customer behavior and assist businesses in making informed decisions regarding how to enhance their products and services.

It is expected that as AI and machine learning technologies continue to progress, customer churn prediction systems will become more accurate and sophisticated. Businesses that adopt these technologies and employ them to increase client retention will enjoy a competitive advantage in the e-commerce sector. Hence, additional research and development is required to enhance the accuracy and effectiveness of customer churn prediction systems and enable e-commerce businesses to provide superior customer service.

7.2 Recommendations

Based on the findings of our study on customer churn, we recommend the following:

i). Creating a model for predicting customer turnover that satisfies the practical needs of businesses is beneficial for enterprise customer relationship management. Companies can gain insight into the causes of client turnover based on the significance of consumption variables.

7.3 Limitations of the Study

The results of this study contain limitations as well. The data for this study were gathered from a Brazilian ecommerce website operating in a B2C context; thus, the selection of data is limited. Ideally, multiple data sets should verify the research conclusions. The findings of customer segmentation have a significant impact on the predictive accuracy of the model. In addition, the study employed a small number of predictive variables, limiting the dissemination of our findings because B2C websites display a wealth of purchasing information, some of which may be disregarded.

7.4 Future Work

This study demonstrates the significance of customer churn prediction in the B2C ecommerce market, as well as the potential of deep learning algorithms to enhance the accuracy and efficacy of customer churn prediction. Nonetheless, there are limitations to the study that must be addressed in future research. This research was based on a single dataset; hence the findings may not be applicable to different industries such as Telecom or banking. To prove its efficacy, further research should evaluate the performance of the deep learning system using multiple datasets and scenarios. Second, the focus of this study was on forecasting customer attrition but did not investigate the efficacy of retention tactics in reducing customer churn. Future research should investigate the efficacy of various retention tactics in minimizing customer churn and boosting customer loyalty.

References

Ahmad Naz, N., Shoaib, U., & Shahzad Sarfraz, M. (2018). A Review on Customer Churn Prediction Data Mining Modeling Techniques. *Indian Journal of Science and Technology*, *11*(27), 1–7. https://doi.org/10.17485/ijst/2018/v11i27/121478

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0191-6

Ahmed, A. A. Q., & Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. *Egyptian Informatics Journal*, *18*(3), 215–220. https://doi.org/10.1016/j.eij.2017.02.002

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, *9*(8), 1295. https://doi.org/10.3390/electronics9081295

Al Kurdi, B., Alshurideh, M., & Alnaser, A. (2020). The impact of employee satisfaction on customer satisfaction: Theoretical and empirical underpinning. *Management Science Letters*, *10*(15), 3561–3570. https://doi.org/10.5267/j.msl.2020.6.038

Alsaqqa, S., Sawalha, S., & Abdel-Nabi, H. (2020). Agile Software Development: Methodologies and Trends. *International Journal of Interactive Mobile Technologies*, *14*(11), 246–270. https://doi.org/10.3991/ijim.v14i11.13269

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing*, *237*, 242–254. https://doi.org/10.1016/j.neucom.2016.12.009

Ampadu, H. (2021, May 14). *Decision Trees*. Ai-Pool.com; AI Pool Inc. https://ai-pool.com/a/s/decision-trees

Balmer, A. S., & Murcott, A. (2017). *The craft of writing in sociology : developing the argument in undergraduate essays and dissertations*. Manchester University Press.

Bramer, M. (2013). Ensemble Classification. *Principles of Data Mining*, 209–220. https://doi.org/10.1007/978-1-4471-4884-5_14

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Brownlee, J. (2016, August 10). *5 Step Life-Cycle for Neural Network Models in Keras*. Machine Learning Mastery. https://machinelearningmastery.com/5-step-life-cycle-neural-network-models-keras/

Capraro, A. J., Broniarczyk, S., & Srivastava, R. K. (2003). Factors Influencing the Likelihood of Customer Defection: The Role of Consumer Knowledge. *Journal of the Academy of Marketing Science*, *31*(2), 164–175. https://doi.org/10.1177/0092070302250900

Chai, W., Ehrens, T., & Kiwak, K. (2020, September). *What is CRM (customer relationship management)?* SearchCustomerExperience. https://www.techtarget.com/searchcustomerexperience/definition/CRM-customer-relationship-management

Cooper, D. R., & Schindler, P. S. (2014). *Business Research Methods*. http://www.mim.ac.mw/books/Donald%20R%20Cooper's%20Business%20Research%20Methods,%2012th%20Edition.pdf

Erickson, B. J., Korfiatis, P., Akkus, Z., Kline, T., & Philbrick, K. (2017). Toolkits and Libraries for Deep Learning. *Journal of Digital Imaging*, *30*(4), 400–405. https://doi.org/10.1007/s10278-017-9965-6

Gachenge, B. (2020). *The Effects of Online Customer Experience On E-Commerce Adoption In Nairobi: A Case of Jumia E-Commerce Platform*. https://erepo.usiu.ac.ke/bitstream/handle/11732/5977/BEATRICE%20GACHENGE%20%20MBA%202020.pdf?sequence=1&isAllowed=y

Glaser, B. G., & Strauss, A. L. (2017). The Discovery of Grounded Theory. *The Discovery of Grounded Theory*, 1–18. https://doi.org/10.4324/9780203793206-1

Guest, G., Namey, E., & Chen, M. (2020). A simple method to assess and report thematic saturation in qualitative research. *PLOS ONE*, *15*(5), e0232076. https://doi.org/10.1371/journal.pone.0232076

Haucap, J. (2015). Editorial: Consumer behavior and telecommunications policy. *Telecommunications Policy*, *39*(8), 625–626. https://doi.org/10.1016/j.telpol.2015.07.013

Heaslip, E. (2022, April 20). *B2B vs B2C: What's the Difference?* Https://Www.uschamber.com/Co. https://www.uschamber.com/co/start/strategy/b2b-vs-b2c#:~:text=B2C%20stands%20for%20business%2Dto

International Trade Administration. (2021, September 13). *Kenya - Information, Communications and Technology (ICT)*. Www.trade.gov. https://www.trade.gov/country-commercial-guides/kenya-information-communications-and-technology-ict

Jain, H., Yadav, G., & Manoov, R. (2020). Churn Prediction and Retention in Banking, Telecom and IT Sectors Using Machine Learning Techniques. *Algorithms for Intelligent Systems*, 137–156. https://doi.org/10.1007/978-981-15-5243-4_12

Jones, T. O., & Sasser, W. E. (2014, August). *Why Satisfied Customers Defect*. Harvard Business Review. https://hbr.org/1995/11/why-satisfied-customers-defect

Jwalapuram, N. (2021, April 12). *PyTorch Library | What is PyTorch Library for Deep Learning* |. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/04/a-gentle-introduction-to-pytorch-library/

Kenton, W. (2021, October 29). *Customer: Definition and How to Study Their Behavior for Marketing*. Investopedia. https://www.investopedia.com/terms/c/customer.asp

Kriti. (2019). *Customer Churn: A Study of Factors Affecting Customer Churn using Machine Learning*. https://dr.lib.iastate.edu/server/api/core/bitstreams/963c8e0d-4209-4137-9d05-ac20968963f9/content

Krzysztof Grabczewski. (2016). *Meta-Learning In Decision Tree Induction.* Springer International Pu.

Kulkarni, A., Patil, A., Patil, M., & Bhoite, S. (2019). Customer Churn Analysis and Prediction. *International Journal of Computer Applications Technology and Research*, *8*(9), 363–366. https://doi.org/10.7753/ijcatr0809.1005

Matuszelański, K., & Kopczewska, K. (2022, January 15). *Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach*. https://www.mdpi.com/0718-1876/17/1/9

Nichols, J. A., Herbert Chan, H. W., & Baker, M. A. B. (2018). Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews*, *11*(1), 111–118. https://doi.org/10.1007/s12551-018-0449-9

Park, Y. S., Konge, L., & Artino, A. R. (2020). The Positivism Paradigm of Research. *Academic Medicine*, *95*(5), 690–694. Researchgate. https://doi.org/http://dx.doi.org/10.1097/ACM.0000000000003093

Pondel, M., Wuczyński, M., Gryncewicz, W., Łysik, Ł., Hernes, M., Rot, A., & Kozina, A. (2021). Deep Learning for Customer Churn Prediction in E-Commerce Decision Support. *Business Information Systems*, 3–12. https://doi.org/10.52825/bis.v1i.42

Pustokhina, I. V., Pustokhin, D. A., Nguyen3, P. T., Elhoseny, M., & Shankar, K. (2021, March 21). *Multi-objective rain optimization algorithm with WELM model for customer churn prediction in telecommunication sector*. Link Springler. https://link.springer.com/content/pdf/10.1007/s40747-021-00353-6.pdf

Rampasek, L., & Goldenberg, A. (2016). TensorFlow: Biology's Gateway to Deep Learning? *Cell Systems*, *2*(1), 12–14. https://doi.org/10.1016/j.cels.2016.01.009

Ramadhanti, D., Mohamad, E., Muid, A., & Larasati, A. (2023). *Building customer churn prediction models in Indonesian telecommunication company using decision tree algorithm* . Aip.org. https://pubs.aip.org/aip/acp/article-abstract/2654/1/040001/2869364/Building-customer-churn-prediction-models-in?redirectedFrom=fulltext

Reichheld, F., & Sasser, W. Earl. (2014, August). *Zero Defections: Quality Comes to Services*. Harvard Business Review. https://hbr.org/1990/09/zero-defections-quality-comes-to-services

Reinartz, W. J., & Kumar, V. (2003). The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. *Journal of Marketing*, *67*(1), 77–99. https://doi.org/10.1509/jmkg.67.1.77.18589

Saheed, Y. K., & Hambali, M. A. (2021, October 1). *Customer Churn Prediction in Telecom Sector with Machine Learning and Information Gain Filter Feature Selection Algorithms*. IEEE Xplore. https://doi.org/10.1109/ICDABI53623.2021.9655792

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2017). Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & Quantity*, *52*(4), 1893–1907. https://doi.org/10.1007/s11135-017-0574-8

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students* (8th ed.). Pearson.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 3–29. https://doi.org/10.1177/1536867x20909688

Shabankareh, M. J., Nazarian, A., Ranjbaran, A., Seyyedamiri, N., & Shabankareh, M. A. (2021). *A Stacking-Based Data Mining Solution to Customer Churn Prediction*. WestminsterResearch. https://westminsterresearch.westminster.ac.uk/download/4e063ec5b9f2da44099b0dbcd8f9ba89862ab36e820bd727496ace0b80f4c643/967319/final%20version12.pdf

Sharma, P. (2022, March 3). *A Basic Introduction to Tensorflow in Deep Learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/03/a-basic-introduction-to-tensorflow-in-deep-learning/

Swetha Amaresan. (2018). *What Is Customer Churn? Definition]*. Hubspot.com. https://blog.hubspot.com/service/what-is-customer-churn

Tariq, M. U., Babar, M., Poulin, M., & Khattak, A. S. (2021, June). *Distributed model for customerchurn prediction using convolutional neural network*. Research Gate. https://www.researchgate.net/publication/351813838_Distributed_model_for_customer_churn_prediction_using_convolutional_neural_network

Tipton, D. (2020, November 8). *What is Agile Methodology?* Eoiin Connect. https://eoiinconnect.com/what-is-agile-methodology/

Tobi, H., & Kampen, J. K. (2017). Research design: the methodology for interdisciplinary research framework. *Quality & Quantity*, *52*(3), 1209–1225. springer. https://doi.org/10.1007/s11135-017-0513-8

Verhoef, P. C. (2003). Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development. *Journal of Marketing*, *67*(4), 30–45. https://doi.org/10.1509/jmkg.67.4.30.18685

Wai-Ho Au, Chan, K. C. C., & Xin Yao. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation*, *7*(6), 532–545. https://doi.org/10.1109/tevc.2003.819264

Wu, X., & Meng, S. (2016, June 1). *E-commerce customer churn prediction based on improved SMOTE and AdaBoost*. IEEE Xplore. https://doi.org/10.1109/ICSSSM.2016.7538581

Xiahou, X., & Harada, Y. (2022). B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(2), 458–475. https://doi.org/10.3390/jtaer17020024

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. https://doi.org/10.1016/j.eswa.2008.06.121

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

Yasar, K. (2021). *PyTorch*. SearchEnterpriseAI. https://www.techtarget.com/searchenterpriseai/definition/PyTorch

Yu, L., Li, B., & Jiao, B. (2019). Research and Implementation of CNN Based on TensorFlow. *IOP Conference Series: Materials Science and Engineering*, *490*, 042022. https://doi.org/10.1088/1757-899x/490/4/042022

Yuksel, A. (2008, January). *(PDF) Consumer Satisfaction Theories: A Critical Review*. ResearchGate. https://www.researchgate.net/publication/258224400_Consumer_Satisfaction_Theories_A_Critical_Review

Zhang, J., Chen, W., Petrovsky, N., & Walker, R. M. (2021). The Expectancy-Disconfirmation Model and Citizen Satisfaction with Public Services: A Meta-analysis and an Agenda for Best Practice. *Public Administration Review*, *82*(1). https://doi.org/10.1111/puar.13368

Zhao, M., Zeng, Q., Chang, M., Tong, Q., & Su, J. (2021). A Prediction Model of Customer Churn considering Customer Value: An Empirical Research of Telecom Industry in China. *Discrete Dynamics in Nature and Society, 2021*, 1–12. https://doi.org/10.1155/2021/7160527

Žukauskas, P., Vveinhardt, J., & Andriukaitienė, R. (2018). Philosophy and Paradigm of Scientific Research. *Management Culture and Corporate Social Responsibility*. Intechopen. https://doi.org/https://doi.org/10.5772/intechopen.70628

Appendices

Appendix A: Summary of Literature Review

Table A.1 Summary of the algorithms that have been widely used in customer churn prediction.

Algorithm	Author	Advantages	Limitations
Random Forests	(Breiman, 2001)	-It can handle big data with numerous variables. -It can automatically balance datasets when a class is more infrequent than other classes in the data.	Large number of trees can make the algorithm too slow.
K-Means	(Ahmed et al., 2020)	-Scales to large datasets. -Easy to implement	It has problem clustering data where clusters are of varying sizes and density.
Decision Trees	(Krzysztof Grabczewski, 2016)	-It requires less data preparation.	-It can easily overfit because it lacks an inherent mechanism to stop hence creating complex decision rules.

Table A.2 Summary of research work in comparison to the churn prediction

Author	Industry	Techniques	Evaluation Metrics	Limitations
(Xiahou & Harada, 2022)	E-commerce	K-Means, Support Vector Machines	Accuracy: 90% Recall: 95% Precision:85%	-This study only uses K-means algorithm to segment and divide customer type which might not provide convincing results compared to using two segmentation methods
(Matuszelaski & Kopczewska, 2022)	E-commerce	XGBoost Logistic Regression Models	Not Stated	-Only two algorithms were tested. -Study was conducted during pre-covid period hence some factors driving ecommerce during covid were ignored.
(Wu & Meng, 2016)	E-commerce	Synthetic Minority Oversampling Technique (SMOTE), Adaboost Algorithm	Not Stated	-Study was conducted during pre-covid period and the features used does not include factors after covid.

(Pondel et al ,2021)	E-commerce	Deep Learning	Accuracy:74% Recall:68% Precision:78%	-Low accuracy, recall and precision levels.
(Saheed & Hambali's ,2021)	Telecom	Support Vector Machine (SVM), the Multi-Layer Perceptron (MLP), the Random Forest (RF), and Naive Bayes (NB).	Accuracy:95%	-Prediction performed in contractual settings where churn is known.
(Shabankareh et al ,2021)	Telecom	Support Vector Machine	Not Stated	-Only one algorithm was used for prediction. - Prediction done in contractual settings.
(Jain et al, 2020)	Telecom	Logistic Regression, Random Forest, SVM and XGBoost	Not Stated	- Prediction done in contractual settings.

Appendix B: Project Gantt Chart

Link: https://sharing.clickup.com/9003112570/g/h/8ca163u-241/16b3e6d27a500d4

Snip view:

Table B.1 Project Gantt chart showing key stages and dates in 2022 - 2023

	5/1-12/22	6/7-26/22	7/8-21/22	11/1-24/22	1/2-27/23	2/2-21/23	3/3-29/23	4/1-5/23
Abstract								
Proposal								
Risk and ethics								
Literature review								
Implement model								
Test model								
Evaluate and conclude								
Mid-point review								
Submission								
Signing								
Write-up								

Appendix C: Dataset

Due to privacy reasons, the data set from any and all institutions used in this study remains confidential.

However, as the training data set was from Kaggle, one can access it via the below link:

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

In case the data is deleted, or the link is inaccessible, kindly contact:

Papetua.narina@strathmore.edu

Appendix D: Data Analysis

This study followed a mixed method approach that combined both quantitative and qualitative data analysis to comprehensively examine the requirements of an effective stock market forecasting system.

Quantitative data collected through interviews with individual investors was subjected to statistical analysis using descriptive statistics such as mean, standard deviation and frequency.

The results of our quantitative analysis shed light on effectively estimating client turnover, ecommerce businesses may maximize customer retention, customer loyalty, and ultimately their bottom line.

By applying a mixed-methods approach, this study ensured a full synthesis of quantitative and qualitative data analysis, thereby facilitating a comprehensive and differentiated understanding of the research question.

Appendix E: Consent Form

Dear Participant,

We are conducting a customer churn prediction study to understand the factors that lead to customer attrition in Jumia Kenya. This study is being carried out by Papetua Narina.

The purpose of this research is to analyze customer data and identify patterns that may help to predict when a customer is likely to terminate their relationship with the company. The information we gather from this study will help us to develop strategies to improve customer retention and satisfaction.

Your participation in this research involves sharing your personal data with us. The data we will collect includes your name, contact details, transaction history, and other relevant information that we may obtain from Jumia Kenya. All of this data will be kept confidential and will only be used for the purpose of this study.

We want to assure you that your participation in this research is completely voluntary. You may choose to withdraw from the study at any time without any penalty or consequence. Your decision to participate or not will not affect your relationship with Jumia Kenya.

By signing below, you agree to participate in the research and give your consent for us to use your personal data for research purposes. You also understand that your data will be kept confidential and used only for this study.

Thank you for your participation in this important study.

Sincerely,

Papetua Narina

I, _____, give my consent for my personal data to be used in the [Project Name] Customer Churn Prediction Study.

Signature: _____ Date: _____

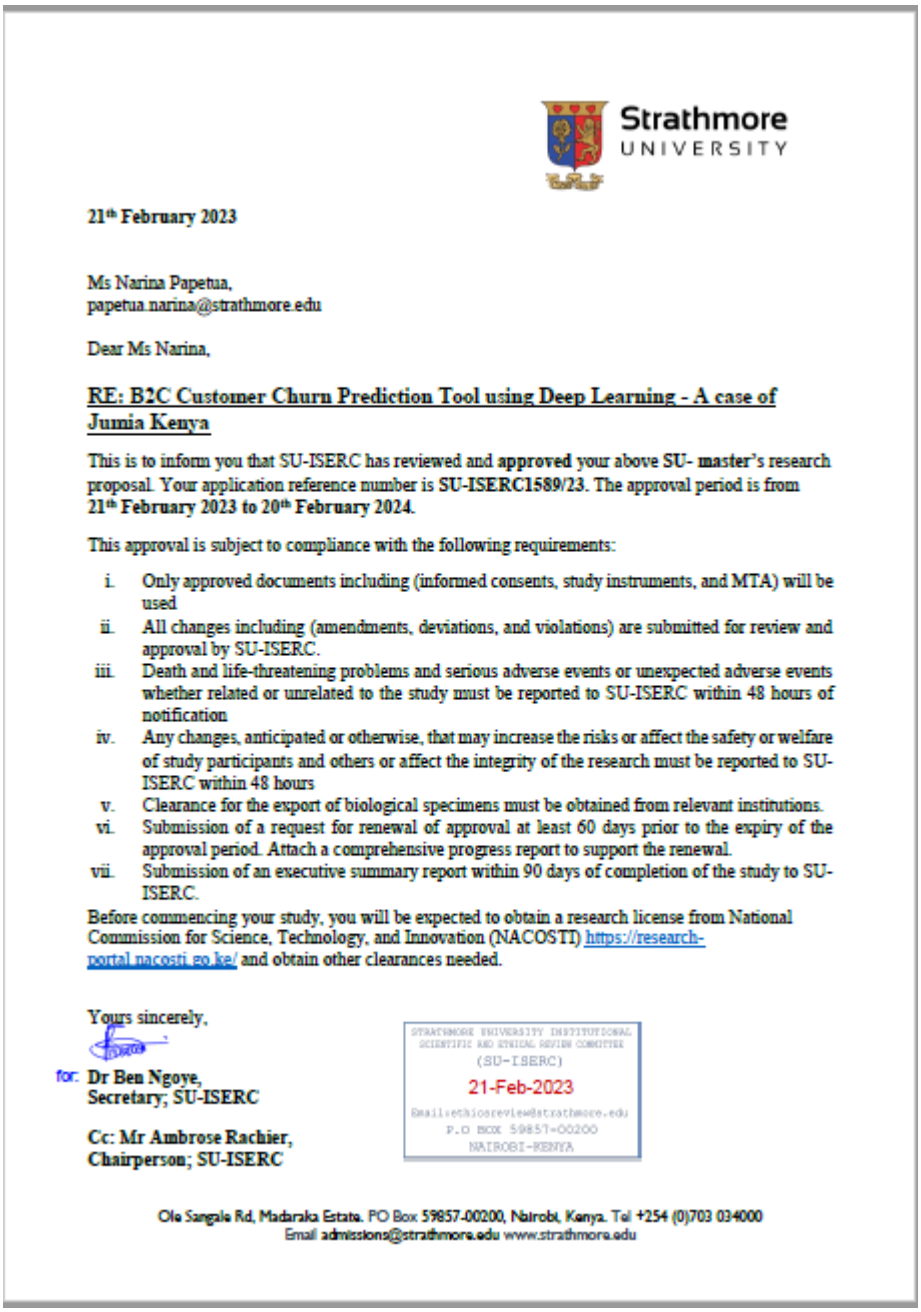
Appendix F: Budget

Item and Description	Price in Kes
A laptop with a core i7 processor, with GPU capability, 16GB of ram, and a 1terabyte SSD drive	150,000
Google colab 6 months subscription	7,800

For hosting:

Item and Description	Price in Kes
Cloud Hosting	5,000

Appendix G: Ethical Review



Strathmore UNIVERSITY

21th February 2023

Ms Narina Papetua,
papetua.narina@strathmore.edu

Dear Ms Narina,

RE: B2C Customer Churn Prediction Tool using Deep Learning - A case of Jumia Kenya

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU- master's** research proposal. Your application reference number is **SU-ISERC1589/23**. The approval period is from **21th February 2023 to 20th February 2024**.

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, and MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 48 hours of notification
iv. Any changes, anticipated or otherwise, that may increase the risks or affect the safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 48 hours
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

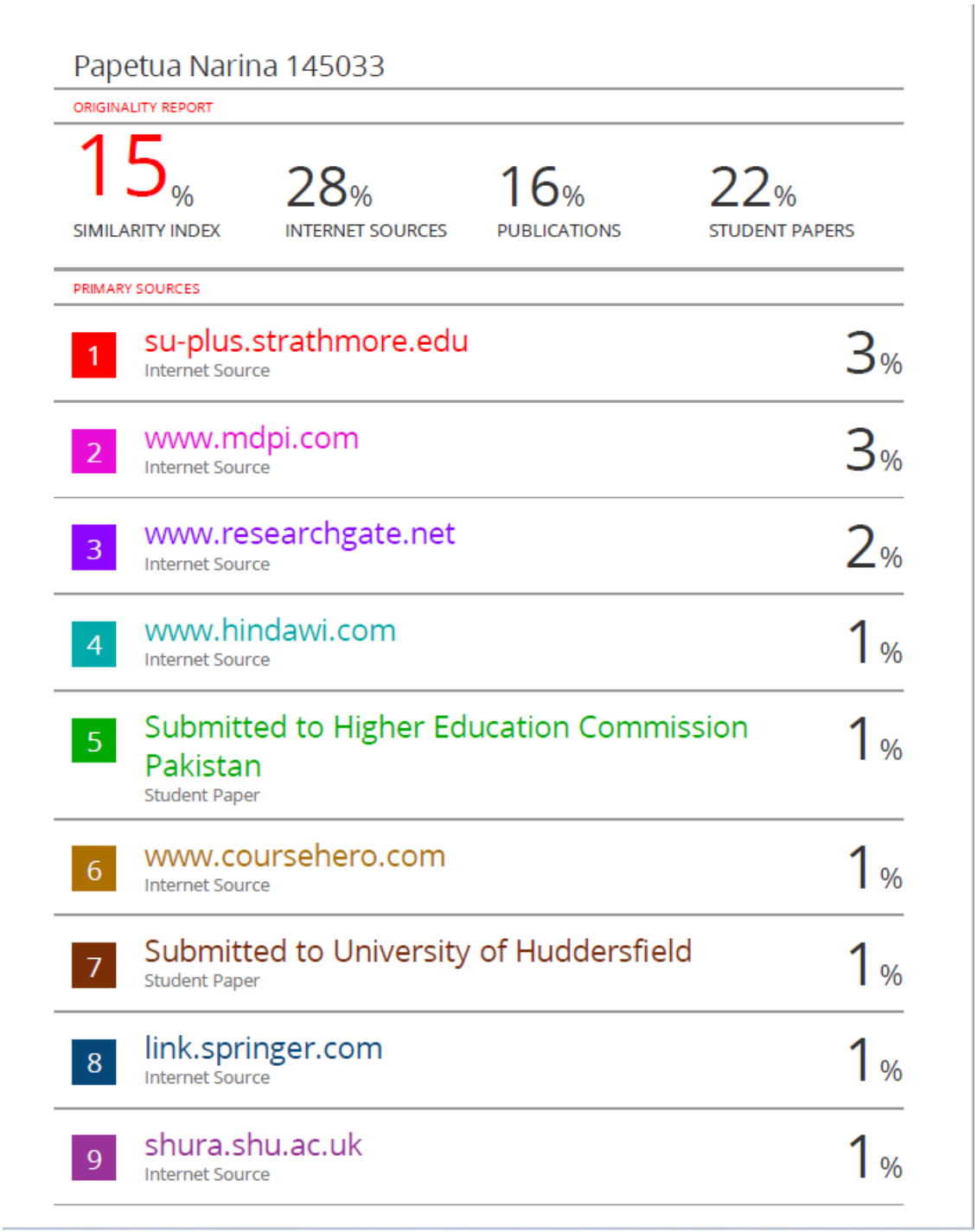
for: **Dr Ben Ngoye,**
Secretary; SU-ISERC

Cc: Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
21-Feb-2023
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix H: Similarity Report



Papetua Narina 145033

ORIGINALITY REPORT

15%	28%	16%	22%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	su-plus.strathmore.edu Internet Source	3%
2	www.mdpi.com Internet Source	3%
3	www.researchgate.net Internet Source	2%
4	www.hindawi.com Internet Source	1%
5	Submitted to Higher Education Commission Pakistan Student Paper	1%
6	www.coursehero.com Internet Source	1%
7	Submitted to University of Huddersfield Student Paper	1%
8	link.springer.com Internet Source	1%
9	shura.shu.ac.uk Internet Source	1%