CUSTOMER CHURN PREDICTION IN E-COMMERCE INDUSTRY USING RANDOM FOREST ALGORITHM

SOH JOEN SHIUAN

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF THESIS

Author's full name	:	SOH JOEN SHIUAN			
Student's Matric No.	:	MCS241028	Academic Session	:	20242025-02
Date of Birth	:	4 JUNE 1999	UTM Email	:	sohjoenshiuan@graduate.utm.my
Thesis Title	:	CUSTOMER CHURN PREDICTION IN E-COMMERCE INDUSTRY USING RANDOM FOREST ALGORITHM			

I declare that this thesis is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the thesis belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this thesis for academic exchange.

Signature of Student:

Signature :

Full Name: SOH JOEN SHIUAN
Date : 30 JUNE 2025

Approved by Supervisor(s)

Signature of Supervisor I:

Full Name of Supervisor I
DR SHAHIZAN BIN OTHMAN

Date : 30 JUNE 2025

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

"I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in term of scope and quality for the award of the degree of Master in Data Science"

Signature	:	_____
Name of Supervisor I	:	DR SHAHIZAN BIN OTHMAN
Date	:	30 JUNE 2025

CUSTOMER CHURN PREDICTION IN E-COMMERCE INDUSTRY USING RANDOM FOREST ALGORITHM

SOH JOEN SHIUAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master in Data Science

Choose an item.
Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2025

DECLARATION

I declare that this thesis entitled *"Customer Churn Prediction in E-commerce Industry Using Random Forest Algorithm"* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	.....
Name	:	SOH JOEN SHIUAN
Date	:	30 JUNE 2025

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Dr. Mohd Shahizan bin Othman, for encouragement, guidance, critics and friendship. Without their continued support and interest, this thesis would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my Master study. Librarians at UTM also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

Customer churn in e-commerce affects business sustainability, with retention costs being lower than acquisition costs. Existing studies lack comprehensive approaches in handling imbalanced datasets and feature engineering methodologies for e-commerce customer churn prediction. This research gap highlights the need for improved techniques that effectively address class imbalance while maximizing predictive performance. The purpose of this study is to investigate key features that affect customer churn and develop a machine learning model implemented in Python. Four algorithms are systematically compared, including Logistic Regression, Random Forest, Random Forest with SMOTE, and XGBoost, utilizing correlation analysis and comprehensive hyperparameter optimization to identify the best performing model for this use case. The dataset comprises 5,630 customer records with a class imbalance ratio of 4.94:1, where 83.2% of customers are non-churned and 16.8% are churned. Without hyperparameter tuning, XGBoost performs best with F1-Score of 0.8456 and ROC-AUC of 0.9662. After hyperparameter optimization, Random Forest achieves superior performance with F1-Score of 0.8556 and ROC-AUC of 0.9801, demonstrating significant improvement through systematic parameter tuning. These findings provide an actionable framework for e-commerce customer churn prevention systems, enabling businesses to implement proactive retention strategies and improve customer lifetime value through data-driven decision making.

Keywords: Customer Churn, E-Commerce, Imbalanced Dataset, Feature Engineering, Machine Learning, Classification, Random Forest, Hyperparameter Tuning

ABSTRAK

Pembelotan pelanggan dalam e-dagang telah menjejaskan kelestarian perniagaan dengan kos pengekalan yang lebih rendah berbanding dengan kos pemerolehan. Kajian-kajian yang sedia ada menunjukkan kekurangan pendekatan yang komprehensif dalam mengendalikan set data yang tidak seimbang dan metodologi kejuruteraan ciri untuk ramalan pembelotan pelanggan e-dagang. Tujuan penyelidikan ini adalah untuk mengkaji ciri-ciri utama yang mempengaruhi pembelotan pelanggan dan membangunkan model pembelajaran mesin dengan menggunakan Python. Empat algoritma telah dibandingkan secara sistematik, termasuk Regresi Logistik, Random Forest, Random Forest dengan SMOTE, dan XGBoost. Selepas itu, analisis korelasi dan pengoptimuman hiperparameter telah digunakan untuk mengenal pasti model yang berprestasi terbaik bagi kajian ini. Set data terdiri daripada 5,630 rekod pelanggan dengan nisbah ketidakseimbangan kelas 4.94:1, di mana 83.2% pelanggan tidak berbelot dan 16.8% pelanggan berbelot. Tanpa penalaan hiperparameter, XGBoost menunjukkan prestasi yang baik dengan Skor-F1 sebanyak 0.8456 dan ROC-AUC sebanyak 0.9662. Selepas pengoptimuman hiperparameter, Random Forest mencapai prestasi unggul dengan Skor-F1 sebanyak 0.8556 dan ROC-AUC sebanyak 0.9801. Hal ini telah menunjukkan peningkatan ketara melalui penalaan parameter yang sistematik. Dapatan ini menyediakan kerangka kerja yang boleh dilaksanakan untuk sistem pencegahan pembelotan pelanggan e-dagang, membolehkan perniagaan melaksanakan strategi pengekalan proaktif dan meningkatkan nilai seumur hidup pelanggan melalui pembuatan keputusan berasaskan data.

Kata Kunci: Pembelotan Pelanggan, E-Dagang, Set Data Tidak Seimbang, Kejuruteraan Ciri, Pembelajaran Mesin, Pengelasan, Random Forest, Penalaan Hiperparameter

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

UTM	-	Universiti Teknologi Malaysia
B2C	-	Business-to-Consumer
CLV	-	Customer Lifetime Value
CRISP-DM	-	Cross Industry Standard Process of Data Mining
SMOTE	-	Synthetic Minority Over-sampling Technique
XGBoost	-	Extreme Gradient Boosting
COVID-19	-	Coronavirus Disease 2019
EDA	-	Exploratory Data Analysis
ML	-	Machine Learning
TP	-	True Positive
TN	-	True Negative
FP	-	False Positive
FN	-	False Negative
MVL	-	Multi View Learning
ROC	-	Receiver Operating Characteristics
AUC	-	Area Under the Curve
ROI	-	Return On Investment

LIST OF SYMBOLS

β	-	Coefficient of the parameter
x	-	nth Feature value
ϵ	-	Error term

LIST OF APPENDICES

APPENDIX **TITLE** **PAGE**

No table of figures entries found.

CHAPTER 1

INTRODUCTION

1.1 Introduction

In the fast-paced growing environment, retail sales transform business model into online sales to speed up transaction procedure to keep up with the digital trend. According to Ikhlass Boukrouh (2025), this is known as electronic commerce (e-commerce) in which business transactions are conducted on the internet. This transformation brings general challenges such as high competition between e-commerce platforms, high return and refund rate from customers. This leads to customer churn and the trend keeps growing. Based on the statistics from Jack M. Germain (2023), the 42% if B2C companies are churning 3% or more, followed by another 16% that churns 4% or more. Most of the time, people decide to cease using a service after becoming increasingly unhappy with it over time. Customer churn happens when customer has decided to stop using the service. This reduces the brand loyalty, increasing future customer acquisition cost and revenue down, leading to business loss (Daniyal Asif, 2025). Churn prediction serves as a guidance for management team to refine marketing strategies, ensuring the enhanced approach fits the marketing objectives, leading to reduction in revenue (Sulim Kim, Heeseok Lee, 2022). However, few studies have implemented the churn analysis because it is difficult to define who the churners are in e-commerce. Research shows that it might cost five to twenty-five times more to acquire a new customer than to maintain an existing one, hence reducing churn is economically vital (Daniyal Asif et al, 2025). Churn evaluation is required to find out the root cause behind the customer churn. The churn evaluation is conducted in a monthly basis, observing the trends over month. E-commerce management able to adjust the marketing campaign plan, tailoring to the user needs, hence reducing the risk of customers getting churned (Nagaraj et al, 2025).

1.2 Problem Background

Churn prediction allows the e-commerce management to identify whether the customers are likely to stop using the e-commerce platform or not, allowing for preventive intervention. Without addressing customer churn, e-commerce has higher tendency to lose the current customers, leading to lesser commission from transaction between customers and sellers. At the same time, it would be increasing future customer acquisition costs through marketing, advertisements, and promotions. Therefore, conducting research is necessary to reveal reasons that caused customers lose interest in future engagement on the platform. Current method would use manual check to identify whether the customer is churn or not. The moment to identify whether the customer is churn or not, the customer already churn. With the use of predictive model, marketing able to find out a list of customers that has sign to stop using the platform early. Marketing able to implement early prevention strategies such as free shipping/discount vouchers, lucky draw contest to keep these customers from leaving the platform. E-commerce revenue would be increasing in proportion to increased transaction.

There is a form used to collect data, due to lack of information. Demographic and conceptual data is provided at the beginning, but environmental and behavioral data don't have all. The steps to identify the churning customers for the case study is as follows. First, the determined churn customer is identified by group. Then, the determined groups are assigned with goal index. Next, the pattern is extracted using call tree method. The churn customers are revealed. They are officers and engineers. Their similarities are having Persia Insurance as their major business. Due to dissatisfaction towards the service, they decided to leave (Nagaraj P et al., 2023).

1.3 Problem Statement

E-commerce Customer daily active time has been gradually reducing. Reducing active time indicates that the customer might be leaving the e-commerce platform in the future. The number of customers would be reducing gradually, leading

to lesser transactions to be made in the e-commerce platform. The seller would be directly influenced followed by the e- commerce platform. Seller unable to sell the products to the customer. E-commerce platform would gain lesser commission from overall transactions. By increasing e-commerce customer active time, customer would have higher tendency to make more transactions in e-commerce platform. This requires customer daily active period data, total time spent on the platform, purchase history data to train a model that predicts the tendency of customer churn. For the current solution, management predict the users that is frequently active in the platform would be purchasing more items. In the fact that customer is active but without purchasing items. This gives a contradiction point that customer active time does not directly influence the number of customer transactions.

1.4 Research Question

The research question would be focusing on 3 different steps to conduct the churn analysis. The question starts from data preprocessing, followed by identification of churn attribute, at last visualizing the facts using statistical approach.

a) What are the steps to preprocess the customer churn dataset?

b) How does the known characteristics affect the customer churn rate?

c) How does the predicted result give insights on the customer churn?

1.5 Research Aim

The aim is to develop and enhance a predictive model that uses Random Forest algorithm, by labelling the customer as churn or not churn, provide actionable insights in improving the sales revenue and customer active time.

1.6 Research Objectives

The objectives of the research are:

(a) To preprocess the customer churn prediction data, leading to cleaned data for model training.

(b) To identify the key relevant attributes that affects the customer churn rate using Correlation coefficient matrix.

(c) To develop a machine learning model using Random Forest Algorithm that predicts the potential churning customers, visualizing the results in dashboard.

1.7 Research Scope

The scopes of the research are:

(a) The customer churn dataset will be collected from Kaggle Open Data Source. (https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction/data)

(b) Key Customer Lifetime Value (CLV) Variables such as Purchase Amount, Promo Code Used, Category, Previous Purchases, Purchase Date, Customer ID, and Churn will be used to calculate the impact towards customer churn pattern.

(c) The study would be focusing on e-commerce industry.

(d) The research direction would be predictive analysis.

(e) Python programming language would be used to preprocess the data.

(f) Random Forest algorithm would be used to predict the potential churning customers.

1.8 Research Significance

Random Forest algorithm implemented in the study improves the accuracy of churn prediction when the dataset is balanced. The Feature Importance able to identify the most important attribute that affects the customer churn rate. The identified attribute would be used as the metric to train the model recognizing the customer as churn. When the dataset is balanced, it aids in random sampling, which shuffles the training dataset. By each time the dataset is shuffled and trained, the model gets improved by time.

Random Forest require iterative training to get the model. Each iterative training with different sampling of dataset will give different result. Different results combine together forming a new model with higher accuracy. Each iterative training will give the model output. When the model output is then combined with the existing output, the overall accuracy improves. Iterative training consumes larger computational resource, as the training output count increases. More capital would be required investing in hardware to improve the efficiency of obtaining outputs. By using the current dataset, the longer the training count, the higher the accuracy of identifying customer with churn signals.

The research will flag up the customers that will start to churn. It alerts the ecommerce that the customers are going to churn, allowing them have time to make preventive measures to retain the customers. The diverse attributes would be helping managements to detect the signs the customer is going to churn. Hence, identifying the key attributes leading to customer churn is vital, as it indicates the hidden reason that the customer gives up on continuous support in the future.

1.9 Thesis Structure

The thesis consists of 5 chapters. Introduction is the first chapter to be illustrated, followed by second chapter, which is Literature Review. Then, Research Methodology would be fully addressed in Chapter 3. Random Forest algorithm-based model to be implemented on Chapter 4. At the final chapter, a conclusion would be including the insights found across previous chapters.

1.10 Summary

This chapter introduces the problem background, problem statement related to customer churn, which is under e-commerce domain. With the support of problem background and problem statement, the research goals and objectives are focused on e-commerce domain. The research scope highlights the metrics to be used from data preparation to model building, followed by visualizing insights using dashboard. The research significance denotes the importance of conducting the research, expecting the insights found in the result aid in customer churn reduction.

The next chapter would be literature review, which would be comparing the existing methodological approaches, highlighting the strengths and limitations on each solution. The research methods mentioned in previous studies need to analyze deeper. At the same time, identifying the unanswered questions in previous research and highlight the areas that requires further research. The study would be helpful in determining the best model for my e-commerce use case.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter would be discussing the overview of customer churn prediction, followed by the background of e-commerce sector. The comparisons between Linear Regression, Logistic Regression, and Random Forest Model are made to find out the strengths and limitations within the model. At the end of the chapter, the best model that is stronger in overall would be concluded.

2.2 Theoretical Background

Customer churn is a kind of scenario where the customers are leaving the platform without getting noticed. This is a loss to the organization, because organizations unable to make money from the churned customers, especially if the churned customers contribute majority of the sales towards the organization, leading the organization to have more funding for future development.

The customer churn can be categorized into 2 types, that is voluntarily churn and involuntary churn (Fredrik Fagerholm, 2022). Voluntarily churn means the customer decided to leave the platform by stopping the support initiatively without including personal frustration while involuntary churn means the customer leave the platform due to the factors that is beyond their control, causing them to accept the decision forcefully. The studies related to voluntary churn and involuntary churn would be further discussed in the next paragraph.

A study from Arun Velu (2021) saying that the voluntarily churn happens when the customer quits the service or switches to the service provided by the competitor.

Voluntary churn could be caused by various factors, such as product or service dissatisfaction, the availability of superior alternatives or change of personal situations (Long, 2023). It is hard to be identified because the churners would not directly tell the reason they leave with initiatives. They would use action to express out the finalized decision without any reasoning. The reason can be in variety, such as cheaper alternative exist in the competitor platform and service does not meet the minimum requirements of customer needs (Daniel and William, 2023).

From the study done by Nayema Taskin (2023), the involuntary churn happens when the companies resolve the issue beyond the standard timeframe, which they could not control the timeframe for companies' resolution time. The customers who are dissatisfied with anger would be spreading the bad experience towards the third parties, including friends, families, or co-workers. The motive is to instill the negative impacts of keep supporting the thing that they keep support with, so that they would have a mentality to prepare the negative consequences that will be happening on them on time.

In the behavior analysis done by Emílio José Montero Arruda Filho and Alexis de Araújo Barcelos (2020), consumers tried to resolve the complaints through private and public channels, but the result is less satisfying. The expected result from the effort is not equal level with the actual result. Hence, equity instantiation is done to the company, causing the company have detrimental effect in the company reputation. The revenge caused by the resolution alert the companies to take serious action to minimize the financial impact as low as possible.

2.3 Dataset Review

Dataset is the key contributor to the model training. The dataset found in these previous researches would be analyzing its strength, limitations, and the application in the real life case study. The details are articulated in the next paragraphs.

E-commerce platforms generate vast behavioral datasets that capture customer interactions, purchases, and churn signals (Abdulrahman Alshamsi, 2022). These datasets enable researchers to identify patterns in customer attrition using machine learning (ML). For instance, Alshamsi (2022) applied CRISP-DM methodology to a Kaggle dataset of 5,000+ customers, testing Decision Trees, Logistic Regression, and Random Forest models. Strengths included six months of temporal data (June–November 2021) and granular behavioral attributes (e.g., login devices, satisfaction scores), which revealed that mobile app users had 23% higher retention than web users. However, limitations such as class imbalance (few churned customers) and moderate sample size restricted model generalizability, necessitating synthetic oversampling techniques.

The study conducted by Rehka Yadav (2024) stated that the dataset found in Kaggle contain 250,000 entries and 13 columns. Out of the 13 columns, the key attribute churn is the result predicted after summing up the all the attributes into it. These attributes are customer ID, Product Price, Quantity, Total Purchase Amount, Customer Age, Returns, and Churn. Given that the Age and Customer Age share the same characteristics, either Age or Customer Age could be removed. This benefits in model training, consuming less computational resources to process the dataset. At the same time, removal of duplicated attribute reduces the chance to let the model memorize the attribute rather than training it to recognize.

Mukun Chang (2023) analyzed niche live-streaming e-commerce data using clustering techniques. The dataset's strength was novel engagement metrics (e.g., average watch time, gift purchases), which identified four user segments. Surprisingly, "high-spending casual viewers" exhibited 62% churn rates despite high satisfaction scores. A key limitation was platform-specific data collection, limiting cross-segment comparisons. The study demonstrated how unconventional behavioral indicators could uncover non-linear relationships between engagement and churn.

Table 2.1: Comparison between the strengths and limitations of dataset

References	Experiment	Strength	Limitation

Abdulrahman Alshamsi, 2022	Customer Churn in E-Commerce Sector	• Include preferred login device and satisfaction metrics • Include churn flag for labelling	• Imbalanced dataset.
Rehka Yadav, 2024	Machine Learning Insights into E-Commerce Churn	• Comprehensive customer behaviour and purchase history. • Large dataset (250,000 entries with 13 columns)	• Duplicated features exist. (Customer Age and Age columns)
Mukun Chang, 2023	Customer Churn Prediction based on E-Commerce Live Streaming.	• Revealed non-intuitive insights, such as satisfaction, that is not always correlated to churn.	• Focus on live-streaming e-commerce segment

2.4 Supervised Learning in Customer Churn Prediction

Supervised Learning is a type of Machine Learning approach that learns the training data with specified information. The training data is the food that needs to feed the model so that model able to learn the way to classify the data. It can be divided into 3 stages, that is training, testing and validation. (Shreyas Rajesh Labhsetwar, 2020) Firstly, the training data act as a sample to feed into the model. Then, the model is tested using testing dataset, determining the accuracy of the output based on the last training approach. In case the model output is less satisfying, the validation step would be conducted to fine tune the hyper-parameters in the model. The process would be iterating over the model development, until the model fits the expected output in

training dataset. In this case, the 3 supervised learning approach, that is Linear Regression, Logistic Regression, and Random Forest would be described in detail.

2.4.1 Linear Regression

Linear Regression is a method that is used to find out the correlation between independent variable X and dependent variable Y. The independent variable in the dataset is suitable in Linear Regression because it is the only factor that influences the other dependent variable. Hence, the graph is linear line. The formula for single Linear Regression is as below.

$$y = \beta_0 + \beta_1 x + \epsilon \quad (2.1)$$

Where:

y=Predicted value
β=The coefficient of the parameter
x=The nth feature value
ϵ=Error term

According to Enlin Deng (2025), linear regression is used to evaluate the relationship between independent variables and dependent variables. When there are multiple independent variables affect dependent variables, it indicates that the formula that is used in Single Linear Regression becomes incompatible. In such case, Multiple Linear Regression formula is introduced to cope the scenario where multiple dependent variables affect the same variable. Multiple Linear Regression Formula is shown below.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \quad (2.2)$$

Where:

y=Predicted value

β=The coefficient of the parameter

x=The nth feature value

ϵ=Error term

2.4.2 Logistic Regression

Logistic Regression is an algorithm that is used to classify the problems using binary approach (Aurélien Géron, 2023). It is used to estimate the probability of the instance belonging to a particular class. In this case, churn response is the key to define the class. There are two types of churn response, that is churn and not churn. Churn is the case where the situation is true in the class identification whereas not churn is the case where the situation is false in the class identification. Churn is defined as 1 while not churn is defined as 0.

According to Ahmed (2024), Logistic Regression is a method that links the binary dependent variable together with one or more independent variables. The linkage uses fundamental statistical method to discover the relationship between the attributes found in the dataset. At the end of the interpretation, the key variables that affect the customer churn rate could be addressed.

2.4.3 Random Forest

Random Forest is the complex model that is built in the foundation of Decision Tree. Multiple Decision Trees combine together to form a Random Forest model, which is more complex than single Decision Tree. According to Swetha P, Dayananda

R B (2020), Decision Tree able to make decision using divide and conquer method, meaning that it breaks the problem into smaller problem first. Then, it would answer the smaller problem accordingly. At the end of the answering process is completed, all the answered questions combine together to form the final predicted result.

The following divide and conquer concept could be addressed more technically from the analysis done by Xinyu Miao and Haoran Wang (2025). The complex relationship between the first variable and second variable is the advantage of Random Forest. The origin of the Random Forest, which is Decision Tree, only contain a tree. A tree can only contain one set of data. The limitation in the single tree would cause the tree to be overfitting, due to the tree is sensitive towards the pretrained dataset. To reduce the overfitting level, Ensemble Learning is introduced to use the same algorithm to train the original dataset in multiple batches. Each batch of dataset is split randomly. Then, the batch of dataset is dumped into the training process, iterating till the last batch of dataset. The train result of each batch is the aggregated into generalized result, improving the generalization in prediction.

2.4.4 Random Forest Attached with SMOTE

Synthetic Minority Over-Sampling Technique (SMOTE) is a type of technique that improves accuracy based on the foundation of Random Forest Algorithm (Hafiz Ma'ruf and Rodiah, 2021). SMOTE tackles class imbalance by creating new synthetic samples to balance the dataset. The synthetic samples are created in diverse, to ensure that the model have more opportunities to learn the minority class as well, not only limit to learning the majority class.

The advantage of SMOTE is that the original data is preserved during training. Higher accuracy could be achieved and recalled for churn class. The disadvantage is that the noise may be introduced if the dataset itself contain outliers and noise. Moreover, introducing the scattered data have chance to cause the dataset to scatter, making the dataset itself become more unrealistic to actual instances.

2.4.5 Random Forest Attached with XGBoost

Extreme Gradient Boosting (XGBoost) is a type of Random Forest algorithm that is strengthened with multiple supplemental characteristics (Sana Fatima et al., 2023). Compared to pure Random Forest algorithm, XGBoost provides more sophisticated result due to more fine tune approach is done based on Random Forest algorithm. XGBoost able to deal with the imbalanced class by giving priority on underrepresented classes, making the underrepresented class be considered in attention during training process.

The advantage of XGBoost is that it combines weak learners a.k.a individual decision trees to form a big forest a.k.a powerful ensemble model. This powerful ensemble model improves the overall accuracy by filling in the strengths of the multiple models into the weakness found in the individual modules. But, the XGBoost challenge and limitations lies on the hyperparameter setting (Yashkumar Burnwal and Dr. R.C. Jaiswal, 2023). If the hyperparameter is not adjusted properly, it would result in overfitting, which indicates that the model always generates the result as True, even the fact is False. When the data is sparse, the model would require more complex parameter tuning to make the model compatible to the dataset, causing the model complexity increase. Model would become difficult to maintain in the future implementation.

Table 2.2: Comparison between the performance in multiple models

Model	Strengths	Limitations	Citation
Linear Regression	• Simple Implementation method	• Only capture linear relationships. • Sensitive to outliers	Enlin Deng (2025)

Logistic Regression	• Effective for binary classification problems. • Provides probability estimates • Less prone to overfitting with regularization	• Limited to linear decision boundaries • Require more data for stable estimation	Ahmed (2024)
Random Forest	• Reduce overfitting via ensemble averaging.	• Increasing trees cause longer prediction time. •More computational resources required. • Biased to dominant class if not properly tuned.	Swetha P, Dayananda R B (2020)
Random Forest + SMOTE	• Improve imbalanced dataset performance. • Achieve higher accuracy and recall for churn class.	• Artificial data noise • More computational resources than pure Random Forest.	(Hafiz Ma'ruf, 2021)

Random Forest + XGBoost	• Combine bagging and boosting for higher accuracy output.	• Extra parameter tuning required • Model become more complex, increasing maintenance difficulty	(Yashkumar Burnwal and Dr. R.C. Jaiswal, 2023).

2.5 Evaluation Metrics and Performance Analysis

The evaluation of customer churn prediction models requires comprehensive metrics that capture both statistical accuracy and business relevance. The literature review reveals inconsistencies in evaluation approaches across different studies, making performance comparison challenging.

Accuracy remains the most commonly reported metric across the reviewed studies. Alshamsi (2022) achieved varying accuracy levels with different models, with Random Forest showing superior performance compared to Decision Trees and Logistic Regression. Similarly, Yadav (2024) reported accuracy improvements when using comprehensive behavioral features from the large-scale dataset. However, accuracy alone can be misleading in imbalanced datasets where churned customers represent a small percentage of the total customer base.

Precision and recall scores give further insight into model performance, specifically for the minority churn class. Ma'ruf and Rodiah (2021) pointed out the significance of recall in churn prediction since failing to detect actual churners (false negatives) is more expensive than mistakenly identifying loyal customers as potential churners (false positives). Their SMOTE-based Random Forest method was tailored to improve recall for the churn class at no loss in overall accuracy.

F1-score, an area between recall and precision, is a more inclusive measure of an assessment metric. Fatima et al. (2023) highlighted how XGBoost optimization aims at improving F1-score rather than accuracy alone, particularly in dealing with imbalanced datasets. This is more in tune with business objectives where both precision and recall are critical for successful churn prevention programs.

Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC) curves offer an additional level of assessment. Burnwal and Jaiswal (2023) indicated that XGBoost models have greater AUC scores, which reflect the models' ability to accurately distinguish churned and non-churned customers at varying threshold levels. This measure is especially useful for companies that must set their intervention thresholds according to available resources and campaign expenditure.

Cross-validation approaches differ widely between studies. Labhsetwar (2020) employed traditional train-test splits, whereas more recent research such as Chang (2023) employed k-fold cross-validation to preserve model stability when applied to various subsets of the data. The chosen validation approach impacts the model's generalizability and even the validity of the metrics of performance.

Temporal validation presents a special challenge for churn prediction evaluation. Taskin (2023) highlighted the importance of time-based validation in which models are trained on past data and evaluated on future time ranges, mimicking real-world deployment conditions. This temporal dimension is overlooked in research employing random sampling for train-test splits.

The testing duration also differs between studies. Whereas some concentrate on short-term churn prediction (next month), others such as Velu (2021) consider longer prediction horizons. The difference in testing durations makes it difficult to develop standardized performance benchmarks for various business contexts and industries.

2.6 Research Gaps and Limitations

The comprehensive review of customer churn prediction literature reveals several critical gaps that limit the practical application and theoretical advancement of current approaches. It can be discussed in 5 different perspectives, that is methodological gap, dataset and data quality limitations, model interpretability and explainability gaps, real world implementation gaps, and external factors and environmental context.

2.6.1 Methodological Gaps

The lack of standardized evaluation frameworks represents a significant methodological gap. Studies use different metrics, validation approaches, and performance criteria, making meaningful comparison impossible. Imani (2024) noted this inconsistency when evaluating classification methods under varying imbalance levels, highlighting the need for standardized evaluation protocols that consider both statistical and business performance measures.

Feature engineering approaches remain largely ad-hoc across studies. While Yadav (2024) identified duplicate features as a data quality issue, the broader challenge of systematic feature selection and engineering lacks comprehensive treatment. Different studies focus on various feature types - transactional, behavioural, demographic - without establishing clear guidelines for feature selection based on business context or data availability.

The temporal aspect of churn prediction receives insufficient attention in most methodological approaches. Xu et al. (2021) attempted to address this through ensemble learning with feature grouping, but the broader challenge of incorporating time-varying factors and seasonal patterns remains underexplored. Most studies treat churn prediction as a static problem rather than a dynamic process that evolves over time.

2.6.2 Dataset and Data Quality Limitations

Class imbalance remains a persistent challenge across all reviewed studies. Despite various technical solutions like SMOTE (Ma'ruf and Rodiah, 2021) and XGBoost optimization (Fatima et al., 2023), the fundamental issue of limited positive examples for training continues to constrain model performance. More importantly, the artificially balanced datasets may not reflect real-world distributions, raising questions about model performance in actual deployment scenarios.

Dataset generalizability presents another significant limitation. Chang (2023) worked with platform-specific live-streaming data, while Alshamsi (2022) used general e-commerce datasets. The lack of cross-platform and cross-industry validation limits the applicability of findings beyond specific contexts. This is particularly problematic for organizations seeking to implement proven solutions in their unique business environments.

Data privacy and availability constraints are inadequately addressed in current literature. Most studies rely on anonymized or synthetic datasets, but real-world implementation requires dealing with privacy regulations, data governance, and integration challenges that are rarely discussed in academic research.

2.6.3 Model Interpretability and Explainability Gaps

The trade-off between model accuracy and interpretability receives limited attention across the reviewed literature. While ensemble methods like XGBoost achieve higher accuracy (Burnwal and Jaiswal, 2023), their complexity makes it difficult for business users to understand and trust the predictions. This interpretability gap becomes critical when models need to support business decision-making and customer retention strategies.

Feature importance and model explainability are mentioned but not thoroughly explored. Business stakeholders need to understand which factors drive churn

predictions to design effective intervention strategies. The current literature focuses primarily on predictive accuracy without adequately addressing the explanatory requirements of practical applications.

2.6.4 Real World Implementation Challenges

The gap between research and practice remains substantial. Most studies focus on achieving high accuracy in controlled experimental conditions but fail to address deployment challenges such as real-time prediction requirements, system integration, model maintenance, and continuous learning from new data.

Scalability considerations are largely absent from the reviewed literature. While Yadav (2024) worked with 250,000 records, the computational requirements and scalability challenges of deploying these models to process millions of customer records in real-time are not adequately addressed.

The economic impact and cost-benefit analysis of churn prediction models receive minimal attention. Organizations need to understand not just the accuracy of predictions but also the economic value of different intervention strategies and the optimal allocation of retention resources.

2.6.5 External Factors and Environmental Context

The reviewed literature largely ignores external factors that influence customer churn. Market conditions, competitive actions, economic cycles, and external events like the COVID-19 pandemic can significantly impact customer behaviour, but current models are not designed to incorporate these external variables.

Cross-cultural and geographical variations in customer behaviour are not adequately considered. Most studies focus on specific regions or platforms without

examining how cultural factors, regulatory environments, and market maturity affect churn patterns and prediction accuracy.

The dynamic nature of customer preferences and market conditions requires adaptive models that can continuously learn and update their predictions. Current approaches largely rely on static models that require manual retraining, limiting their effectiveness in rapidly changing business environments.

2.7 Discussion

The existing literature shows the methodologies used in customer churn prediction, majorly in e-commerce domain. Even though the other domain such as telecommunication and banking sector also be introduced, the underlying concept is still the same, as all the domain addresses the same problem, that is customer churn prediction. The accuracy of the previous research is a challenging issue in respective to practical implementation constraints.

Theory behind demonstrates a fundamental distinction between voluntary and involuntary churn that will impact how we approach the issue of prediction. Voluntary churn, as defined by Velu (2021), happens when customers consciously decide to switch or leave for competitors. This type of churn is more predictable as it follows some patterns of behavior. But involuntary churn is another matter entirely. Taskin (2023) describes how customers churn because of circumstances outside of their control, including lengthy service resolution times. The behavioral study by Arruda Filho and Barcelos (2020) introduces an additional layer of complexity in that it demonstrates that unhappy customers can actively harm company reputation through negative word of mouth. This indicates that churn prediction models must take into account not only purchase behavior, but also satisfaction levels and external stimuli that impact customer decision-making.

Data analysis in both studies has common challenges that constrain model performance. Alshamsi (2022) dealt with 5,000+ customers with the CRISP-DM

approach, whereas Yadav (2024) dealt with a significantly larger dataset containing 250,000 records and 13 columns. Although of varying sizes, both study works had the same ground problem: class imbalance. The majority of the customers do not churn; thus, the model does not have many examples to learn from the minority class that churns. Chang (2023) attempted to solve this by incorporating live-streaming engagement metrics and found something surprising - high-spending casual watchers had 62% churn rates with high satisfaction rates. This contradicts the typical assumption that spending and satisfaction always result in retention.

The machine learning model comparison shows clear progression in capability, but also increasing complexity. Linear Regression, as analyzed by Deng (2025), provides a simple starting point but can only capture linear relationships between variables. Logistic Regression improves on this by handling binary classification problems better, as Ahmed (2024) demonstrates. However, the real advancement comes with Random Forest models that can handle complex, non-linear relationships. The ensemble approach of Random Forest, explained by Swetha P and Dayananda R B (2020), reduces overfitting by combining multiple decision trees, making predictions more stable and accurate.

The advanced Random Forest methods have even better performance but at the cost of more complexity. SMOTE method, as explained by Ma'ruf and Rodiah (2021), assists in dealing with imbalanced data by generating synthetic samples of the minority class. It improves accuracy and recall in churn prediction but may introduce noise if the initial dataset contains outliers. XGBoost, however, uses a different approach by combining bagging and boosting techniques, as explained by Fatima et al. (2023). While this is more accurate, Burnwal and Jaiswal (2023) explain that XGBoost requires careful hyperparameter tuning and ends up being cumbersome to maintain in practical applications.

What emerges from the debate is a trade-off between accuracy and practicality. The more accurate models are also the more complicated ones, demanding greater computer resources and technical know-how to set up and administer. Organizations

must weigh the need for high accuracy against the practical limitations of their technical environment and personnel.

The literature also identifies a number of gaps that circumscribe the utility of prevailing approaches. First, churn prediction in the majority of the studies is assumed to occur in a static world, whereas the actual world is in a continuous state of change because of market forces, competition, and externalities. Second, feature engineering in the studies differs considerably, and hence comparison of results or determining best practices is challenging. Third, the evaluation metrics value statistical performance more than business impact - the model can be statistically correct but still produce nonsensical retention campaigns that squander resources.

Another critical lacuna involves the lack of consideration for temporal considerations and spillover effects. Although some research incorporates time-sensitive attributes, none of them provide a satisfactory way of specifying how evolving market conditions or competitors' actions could impact trends in churn. The pandemic from the COVID-19 virus, for instance, drastically altered e-commerce habits, but existing models would be unable to respond to such drastic shifts unless re-trained.

The deployment problems of operations are not well covered in the literature as well. Although most studies concentrate on having high accuracy in controlled experimental settings, deployment entails other factors like real-time processing needs, model explainability to business users, and integration with other systems.

Future research should address these limitations by developing more robust approaches that can handle dynamic environments and provide practical value to organizations. This includes creating standardized evaluation frameworks that consider business impact, developing models that can adapt to changing conditions, and focusing on interpretability alongside accuracy. The field needs to move beyond purely technical improvements to create solutions that work effectively in real-world business contexts.

2.8 Summary

The models used in customer churn prediction is not the best solution, due to the variety of factors changed unexpectedly. The churn prediction result provides a valuable insight to the organization, saying that the. In the previous studies that uses Linear Regression and Logistic Regression model able to perform churn identification with fewer attributes. With the presence of Random Forest model, churn identification becomes more generalized and able to generate more stable and accurate prediction, which is condensed from multiple decision trees. Out of all Random Forest Model, Random Forest attached with XGBoost performs the best, due to the ability in processing imbalanced handling, resulting in higher accuracy at the end, compared to pure Random Forest model and Random Forest model attached with SMOTE. The next chapter would be discussing the gap, followed by Data Collection process. The collected data would be used in Data Preprocessing Steps. Cleaned dataset is the final product. Full Exploratory Data Analysis (EDA) would be conducted to find out all the data attributes. Feature Engineering would be conducted to transform the data into features that are compatible with the machine learning models. A comparison of accuracy between the selected Machine Learning model would be made. A thorough analysis of data quality and model selection would be done.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes the research context in the chain-of-thought approach, stepping up from the first step to the last step of the process execution. The Research Framework consists of 5 different stages. They are Problem Identification and Formulation, Data Understanding and Preparation, Exploratory Data Analysis (EDA) and Feature Engineering, Model Development and Training, and Model Evaluation and Deployment. This framework shows the direction to complete the solution implementation based on the Data Science Life Cycle approach.

3.2 Research Framework

The Research Framework consists of 5 different stages. They are Problem Identification and Formulation, Data Understanding and Preparation, Exploratory Data Analysis (EDA) and Feature Engineering, Model Development and Training, and Model Evaluation and Deployment. The research objectives are mapped with 3 of the 5 stages. Stage 2 (Data Understanding and Preparation) preprocess the customer churn data, ensuring the data is cleaned for model training process. Stage 4 (Model Development and Training) develops a machine learning model based on Random Forest algorithm that predicts the potential churning customers. Stage 5 (Model Evaluation and Deployment) starts measuring model performance using performance measures, model comparison and business intelligence.

3.2.1 Phase 1: Problem Identification and Formulation

The initial stage sets the research foundation through extensive acquisition of domain knowledge and problem contextualization. The feedback phase involves four components, that is, acquisition of knowledge through extensive exploratory review of literature, followed by problem definition in which the business issue is clearly defined. Then, stakeholder analysis is used to identify the significant stakeholders that are affected by customer churn. Next, gap identification is used to determine the existing research gaps that still have not resolved by previous researcher. Additionally, the success criteria are established to determine the outcomes that are quantifiable in metric form for the research study. This phase allows a proper understanding of customer churn phenomenon and its business implications before going into technical section.

3.2.2 Phase 2: Data Understanding and Preparation

The second phase revolves around data collection from different sources, mainly from Kaggle datasets. Data exploration involves performing statistical aggregations, assessing the data distribution, and performing data quality checks to verify the characteristics of the dataset. The preparation steps which is known as data cleaning involve handling missing values, outlier detection, and data normalization. This phase ensures research objective 1 to be achieved at the end, that cleans the existing dataset for model training, which would be further executed in the phase 4.

3.2.3 Phase 3: Exploratory Data Analysis (EDA) and Feature Engineering

This phase conducts rigorous data analysis to find the patterns and relationships that lies within the customer churn dataset. Correlation Analysis employs Pearson correlation coefficients, heatmap visualizations, and identification of key attributes to understand the relationship between each variables. Feature selection is used to filter methods, wrapper methods, and embedded methods to choose the predictor that is most

relevant in predicting customer churn. Feature Engineering extract features through categorical encoding and temporal feature extraction. These procedures provide insights into customer behaviour patterns and refine feature sets for machine learning model development, which would be useful in phase 4.

3.2.4 Phase 4: Model Development and Training

Phase 4 employs machine learning algorithms to develop customer churn prediction models. Before developing customer churn prediction model, it is required to split the dataset into 3 different sets, that is, training, validation and testing. Training constitutes 70%, while validation and testing sets constitutes 15% each respectively. This approach ensures the model evaluation result could be guaranteed at the end. Model training employs various algorithms limited to Logistic Regression, Random Forest (SMOTE) and Random Forest (XGBoost) to enable a comparative analysis from different perspectives. Hyperparameter Tuning utilizes Grid Search Cross Validation and performance optimization techniques to enhance model accuracy, which requires multiple try-and-error approach to find the best training value for the parameter. This phase directly addresses the Research Objective 2, that focuses on designing a machine learning model for predicting the customers that has potential to churn in the future.

3.2.5 Phase 5: Model Evaluation and Deployment

Phase 5 constitutes model evaluation and providing actionable business insights to the relevant stakeholders. Performance metrics evaluation include confusion matrix analysis, accuracy, F1-Score, and ROC-AUC analysis to evaluate the model performance. Model comparison involves statistical testing, best model selection, and result validation across multiple algorithms. Business Intelligence components include stakeholder visualization by dashboard creation, actionable insight development to guide business strategy, and development of recommendations for anti-churn initiative programs. This phase clearly matches with Research Objective

3, that evaluates the model performance and identifying churning customers according to various performance metrics, enabling proactive business actions to minimize financial impact as much.

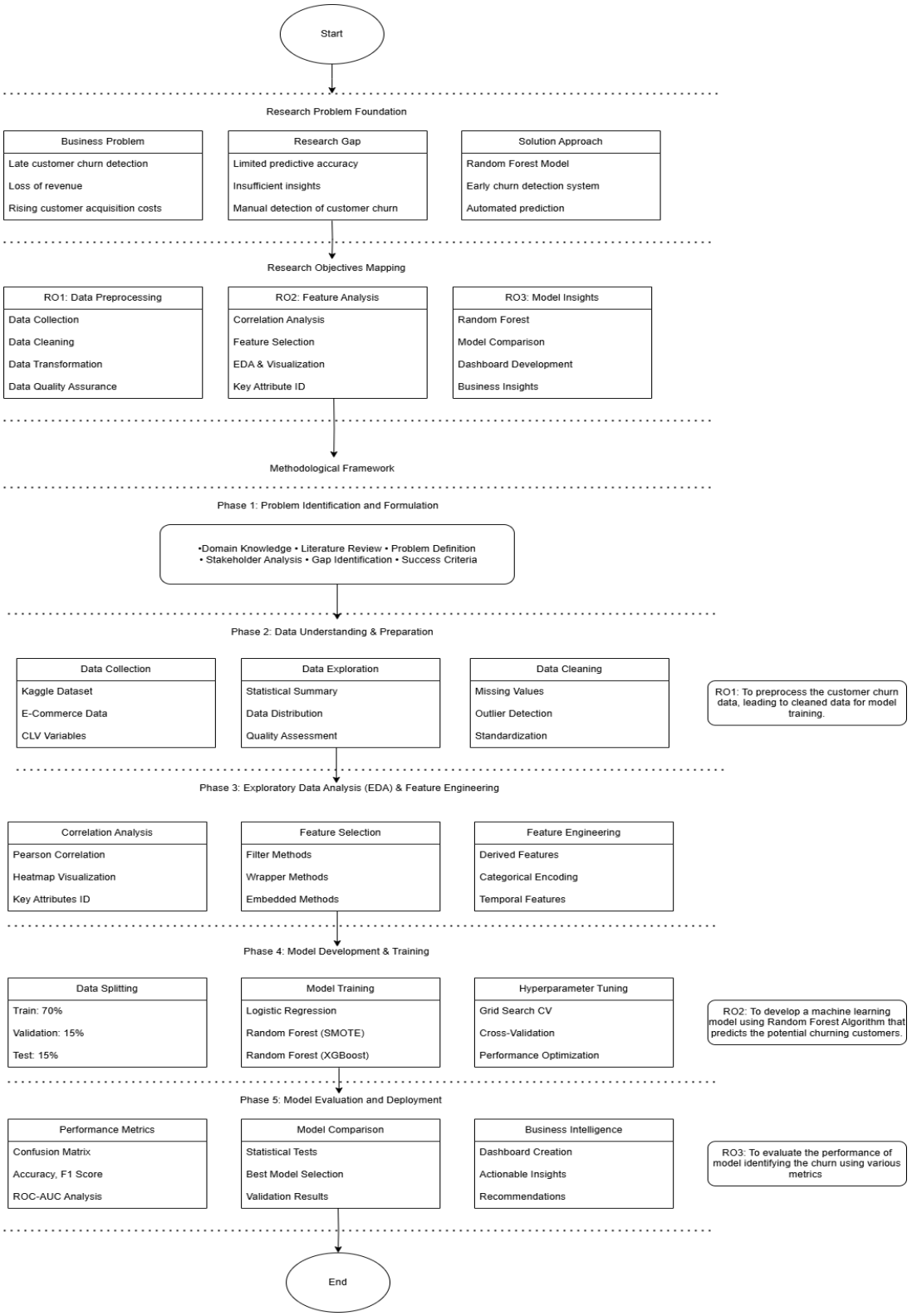

Figure 3.1: Research Framework

3.3 Phase 1: Problem Identification and Formulation

The stage of identification and formulation process begins from systematic review of customers' churn behaviours in e-commerce environments. It is built upon the extensive literature review to craft an explicit research direction. This phase focuses more on the alignment between academic research and practical applications, ensuring that the research would be valuable for both theoretical understanding and industry practice. This process is iterative, as it is refined by looping into multiple cycles. Each cycle shows an improvement in domain insights and understanding of stakeholder requirements.

Domain knowledge acquisition commenced with intensive literature review across numerous databases and publication types. The search strategy was iterative and systematic in approach, starting from keyword searches according to the topic, followed by narrowing down the scope of searching, ending up with specific research topic and research gaps. This literature review recognized several recurring problems in past research, including non-standard evaluation procedures, inadequate business context, and insufficient consideration of model interpretability requirements. The results in the previous research provide a clear direction on the gaps that needs to be resolved in the current research.

The formulation of problem definition involved critical discussion of technical and business issues surrounding customer churn in e-commerce contexts. As compared to typical sectors such as telecom or banking, e-commerce contexts present uniquely challenges that include higher transaction frequency, cross category products, and multichannel complexity of customer interactions. These characteristics necessitated adaptation of standard churn prediction methods while maintaining sufficient rigor to yield correct conclusions.

Stakeholder analysis revealed a number of stakeholders with potentially conflicting needs and success factors. E-commerce management groups primarily seek strategic advice to guide customer retention planning, with a need for actionable forecasting with measurable business effects. But there is limited technical expertise

to make them opt for explainable models rather than exclusive accuracy-optimizing approaches. Customer success and marketing teams need high-precision targeted intervention capabilities to minimize resource wastage, while they need near-real time predictions to succeed in campaign execution. Data science teams are seeking a robust, sustainable predictive models with high accuracy and consistency over different data conditions, constrained by computational performance and system integration requirements.

The mapping of these stakeholder requirements revealed several fundamental areas that needs to be negotiated carefully in methodology design. Sometimes the stakeholder requirements contain disagreement that needs to be addressed out for multiple verification and refinement. The classic trade-off between interpretability and model accuracy was found to be particularly applicable, with technical teams desiring predictive capability and business teams desiring understandability for decision-making. Similarly, precision versus recall preference trade-offs varied between stakeholder groups, requiring flexible evaluation strategies that could accommodate different optimization criteria.

Research gaps were identified systematically via the literature review, and several key areas of research were revealed. Possibilities for algorithmic enhancement included limited exploration of Random Forest variants specifically tuned to e-commerce churn scenarios, no comparison of SMOTE and XGBoost approaches in realistic business contexts, and limited holistic evaluation frameworks that consider both statistical and business performance measures. These gaps provided sufficient justification for the comparative approach in this research.

Dataset and evaluation limitations were another visible gap area, with most existing studies employing proprietary or limited datasets that do not allow reproducibility. Inconsistent utilization of evaluation metrics across studies bars interesting comparison, while limited temporal validation procedures fail to reflect real-world deployment environments. These limitations influenced the decision to utilize publicly available datasets and implement stringent evaluation protocols.

Setting the success criteria involved making finer balancing acts between technical excellence and meaningful value creation. Technical requirements were taken from literature review findings with minimum 85% overall precision requirements, greater than 80% precision targets to guarantee low false positive intervention costs, recall thresholds greater than 75% to detect sufficient churn cases to have business influence, and F1-score targets greater than 77% as balanced performance measures. These needs were complemented by business impact criteria based on the effectiveness of intervention, resource optimization, revenue protection, and proof of scalability.

Problem formulation was tested and validated via several channels for research quality and practical use. Consultation with academic supervisor guaranteed external input towards research rigor and contribution potential, while validation of industry context via existing case studies and reports guaranteed practical use. Methodological consistency checks against existing data science frameworks guaranteed best practice adherence, while systematic verification guaranteed research questions directly solve identified gaps.

3.4 Phase 2: Data Understanding and Preparation

Data preparation and knowledge is a foundational step essential to having direct influence on all subsequent analysis and modelling work. Strategy deployed here is based on lessons learned from literature review, namely addressing the data quality issues and class imbalance issues which are ever-present issues of concern in previous studies. This step entailed systematic probing of accessible datasets, careful checks of the properties of the data, and intensive preparation procedures to present optimal conditions for model construction.

Data collection process involved identifying suitable datasets that support realistic representation of e-commerce customer behaviour without compromising reproducible research availability. After careful evaluation of a number of available datasets, the Kaggle e-commerce customer churn dataset was selected based on its

extensive coverage of relevant customer attributes, sufficient sample size for pertinent analysis, and public access that supports research reproducibility. This dataset contains 5630 rows and 20 columns. It contains 2 sheets, that is Data Dict and E Comm. Data Dict describes the data that is contained within each column whereas E Comm displays all the customer data with relevant attributes. The data source link is as follows https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction

Exploratory data analysis started with intensive examination of structural characteristics like sample size, feature space, types, and general quality indicators. The database consists of customer records across several attribute categories portraying different aspects of customer interaction with the e-commerce website. Demographic attributes like Gender, CityTier, MaritalStatus, NumberOfAddress, and WarehouseToHome provide information regarding customer background properties that could potentially lead to churn likelihood. Behavioral attributes like Tenure, HourSpendOnApp, OrderCount, CashbackAmount, CouponUsed, PreferredOrderCat, PreferredPaymentMode, and DaySinceLastOrder are indicative of actual customer interaction behavior and transactional activities.

Experience-driven features include SatisfactionScore, Complain, and OrderAmountHikeFromLastYear that record dimensions of customer satisfaction and service experience directly related to churn propensity. Technology profile features such as PreferredLoginDevice and NumberOfDeviceRegistered provide information on customer trends of technology uptake likely to be related to engagement levels and chances of retention. Target variable Churn provides the binary class outcome necessary to build supervised learning models.

Data quality evaluation will identify a number of real-world dataset challenges, including missing values, possible outliers, and issues with class imbalance. The class imbalance issue was especially marked, with churned customers making up about 16.8% of the overall sample. Although this imbalanced ratio is more favourable than the severe imbalances found in some published literature reports, it nonetheless must

be treated with caution in model development so that proper learning can take place from minority class instances.

Missing value analysis will provide data completeness patterns within different attribute categories, with systematic missing patterns in some of the attributes that demand special treatment. The treatment included thorough examination of missing value mechanisms for the purpose of determining if data were missing completely at random, missing at random, or missing not at random. Analysis will inform the selection of appropriate imputation methods that minimize bias introduction without sacrificing underlying data relationships.

Outlier detection employed a mix of statistical techniques like interquartile range analysis, z-score calculation, and business logic checks at the domain level. It checked the outliers very carefully to distinguish real extreme values that indicate genuine customer behaviour from faulty data points that must be repaired or removed. It was a balancing act between having absolute data cleaning and keeping genuine behavioural variance that assists model learning.

Data transformation and standardization operations ensured consistent format and scale for all attributes without sacrificing underlying relationships and patterns. Categorical variables were suitably treated for proper encoding with techniques selected based on cardinality and target variable relationship. Numerical variables were examined for characteristics of distribution and transformed appropriately to satisfy modelling assumptions at the expense of interpretability.

The final readiness dataset is the outcome of careful quality improvement processes undertaken to optimize conditions for building models while preserving the inherent nature of customer behaviour patterns. Quality validation processes ensured effective resolution of issues encountered with problems identified in data while guaranteeing dataset representativeness and integrity. The readiness dataset is used as the foundation for exploratory analysis and building all models to be undertaken later on, with clear documentation of preparation processes ensuring reproducibility and transparency of research work.

3.5 Phase 3: Exploratory Data Analysis and Feature Engineering

This phase will establish a comprehensive understanding of customer behaviour patterns and relationships within the dataset through systematic statistical analysis and feature optimization. The EDA process will employ both univariate and multivariate analysis techniques to uncover hidden patterns that influence customer churn decisions, while feature engineering will transform raw data into optimized inputs for machine learning models.

3.5.1 Exploratory Data Analysis Methodology

For Univariate Analysis, the procedure will begin with one-variable exploration to learn about properties of the distribution and data quality for every attribute. Numerical variables will be examined through descriptive statistics like mean, median, standard deviation, skewness, and kurtosis to check shapes of the distribution and ascertain possible data quality issues. Box plots and histograms will be developed to see the distribution and identify outliers or anomalies. Categorical variables will be examined with frequency distributions and bar charts to understand the relative proportions across categories and identify if there are class imbalances in individual features.

Regarding Bivariate Analysis, correlation between every feature and target variable (churn) will be examined to identify the most important predictors. For numerical variables, Pearson correlation coefficients will be employed to quantify linear relationships with churn outcomes via correlation analysis. Statistical significance tests will be employed to check if the strength of these relationships is significant. For category variables, chi-square tests of independence will be conducted to test for association with churn behaviour, with contingency table analysis being included as a supporting tool to identify patterns of distribution across categories.

For Multivariate Analysis, multicollinearity tests and dependencies of features that can impact model performance will be detected by extensive correlation analysis among all numeric variables. Correlation heatmaps will be generated to visualize the

relation matrix and identify the groups of highly correlated features. Cross-tabulation will be applied in examining interaction between two or more categorical variables and their joint effect on churn outcomes. Statistical tests will be employed to verify the significance of any perceived relation and guarantee firm feature selection decisions.

3.5.2 Feature Engineering Methodology

For Categorical Encoding, categorical data shall be mapped to numerical values suitable for machine learning algorithms. The encoding method shall be chosen based on each categorical variable's cardinality and relationship with the target variable. Low-cardinality nominal variables will be processed through one-hot encoding for categorical uniqueness maintenance, and ordinal variables will be processed through label encoding for retaining natural ordering relationships. High-cardinality categorical variables will be examined for target encoding methods if appropriate.

Feature Scaling and Normalization Numerical features will be examined for scale differences and rescaled to optimize model performance. Min-max scaling shall be applied to features with bounded ranges, while standardization (z-score normalization) shall be applied to features with normal distributions. The scaling approach would be determined by the distribution features of individual variables as well as the requirements of selected machine learning algorithms.

New features will be created from domain knowledge and exploratory discovery for better predictability. Ratio-based features will be created to identify relationships between similar variables such as cashback-to-order ratios or engagement-to-tenure correlations. Temporal features will be created from date-based variables to extract seasonality and recency effects. Interaction features will be created for variables with significant combined effects on churn outcomes.

Regarding feature selection approach, a hybrid approach involving filter, wrapper, and embedded techniques will be applied to identify the most informative features to predict churn. Filter techniques will employ statistical measures like mutual information and correlation coefficients to measure features on a basis of individual predictive power. Wrapper techniques will employ recursive feature elimination with cross-validation to measure feature subsets based on model accuracy. Embedded techniques will employ regularization techniques to select features automatically while training the model.

3.5.3 Tools and Implementation Framework

Python shall be utilized as the foundation platform for feature engineering and EDA. Pandas library will be utilized for data transformation and manipulation operations, whereas NumPy will be utilized for doing computations. Matplotlib and Seaborn libraries will be utilized to do statistical plots and visualizations to discover patterns. Scikit-learn will provide feature engineering tools and statistical test functions.

Statistical significance tests will be used to validate all EDA results to ensure good conclusions. Data leakage protection will be verified through feature engineering transformation documentation and testing. Generalizability of feature selection will be promoted by using cross-validation techniques. Business interpretability will be tested for engineered features in order to maintain model explainability for stakeholder reporting.

The outcome of this stage will be an enriched dataset with enriched features that detect the most impactful customer behaviour patterns to forecast churn, supported by thorough statistical analysis documenting the relationships and trends in the customer data.

3.6 Phase 4: Model Development and Training

Customer churn prediction model evaluation requires aggregate metrics that capture both statistical accuracy and business usefulness. Variability in the evaluation approaches among studies from MVL suggests that comparison is impossible.

Model development strategy seeks to utilize algorithms specifically designed to binary classification problems in order to ensure methodological appropriateness in customer churn prediction. Linear regression was deliberately excluded from this research as it produces unbounded continuous outputs that cannot be meaningfully interpreted as probabilities in binary classification contexts. The selected algorithms provide comprehensive coverage of both statistical and ensemble learning approaches while maintaining suitability for the binary nature of churn prediction.

Logistic Regression serves as the statistical baseline method with the sigmoid function formulation $p(y=1|x) = 1/(1 + e^{(-(\beta_0 + \beta_1x_1 + \beta_2x_2 + ... + \beta_nx_n)))}$, where β_i represents the change in log-odds per unit increase in feature x_i, e^{β_i} provides the odds ratio for feature x_i, and $|\beta_i|$ indicates the strength of effect. This formulation naturally constrains outputs between 0 and 1, providing interpretable probability estimates for churn likelihood. Key hyperparameters include regularization strength, solver algorithms, maximum iterations, and penalty terms that control model complexity and convergence behaviour.

$$p(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1x_1 + \beta_2x_2 + ...+ \beta_nx_n)}} \quad (3.1)$$

Where:

βi = change in log-odds per unit increase in feature xi

$e\beta i$ = the odds ratio for feature xi

$|\beta i|$ = the strength of effect

Random Forest is the case of the ensemble approach with the classification prediction function $\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\}$ with T_1 through T_n being individual decision trees trained on bootstrap samples and the final prediction being the majority vote among all trees. This approach tackles the binary classification problem by aggregating single tree predictions, and it naturally outputs class probabilities by vote ratios. Key hyperparameters of Random Forest include n_estimators (number of trees), max_depth (limit of tree depth), min_samples_split (minimum samples required for splitting nodes), min_samples_leaf (minimum number of samples in leaf nodes), and max_features (the maximum number of features to consider at each split).

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), ., T_n(x)\} \quad (3.2)$$

Where:

βi = change in log-odds per unit increase in feature xi

$e\beta i$ = the odds ratio for feature xi

$|\beta i|$ = the strength of effect

The approach also considers more sophisticated Random Forest implementations to address specific churn prediction issues. Random Forest with SMOTE is intended to address class imbalance by generating synthetic instances of churned customers before training, which could improve minority class detection. Random Forest with XGBoost integration applies gradient boosting techniques to optimize prediction accuracy with iterative error correction. These implementations facilitate robust testing of ensemble techniques while monitoring approaches compatible with binary classification operations.

Data splitting operations adhered to best practices while being tuned into the idiosyncrasies of the churn prediction task. Data splitting was conducted into training (60%), validation (20%), and testing (20%) sets using stratified sampling to maintain class distribution equality across all subsets. This splitting ratio leaves sufficient training data for model learning and reserves sufficient validation and test samples to allow robust performance assessment using cross-validation processes. The

stratification ensures that the issue of imbalance in classes is presented equally in all data subsets.

Both algorithms' training procedures were designed with extreme caution to maximize performance with equal comparison conditions. Logistic Regression training employed regularization techniques to prevent overfitting without losing the ability to interpret the model. The strength of the regularization was determined through cross-validation procedures that provided an optimal compromise in bias-variance trade-off for the binary classification task. The application of a sigmoid activation function guarantees that all predictions fall within the appropriate probability range of 0 to 1, and outputs are directly interpretable as probabilities of churn.

Random Forest models needed to have hyperparameter choice handled carefully, especially the number of trees, depth, and minimum samples per leaf. These had a critical impact on model accuracy and computational cost, so they needed systematic tuning using grid search cross-validation approaches. Optimizing between prediction accuracy and computation cost, final models would still be feasible for deployment cases.

Use of SMOTE solved the class imbalance issue by synthetically oversampling the minority class with artificial churned customer examples to balance the training set. The SMOTE parameters like k-nearest neighbours and sampling strategy were optimized to generate realistic synthetic examples to encourage model learning without the introduction of artifacts. Validation procedures were designed in a meticulous way to prevent synthetic examples from augmenting overfitting against artificial patterns rather than enhancing models' generalization.

XGBoost application employed gradient boosting technique to enhance Random Forest performance through iterative correction of error. The parameters of boosting like learning rate, maximum depth, and regularization terms had to be set with utmost care so as to achieve maximum performance without excessive overfitting.

The iterative nature of gradient boosting necessitated monitoring of validation performance so as to set optimal stopping points in addition to limiting complexity.

Hyperparameter tuning applied systematic grid search techniques in combination with cross-validation to determine the optimal parameter combinations for each algorithm. This was computationally demanding but guaranteed that the performance comparisons reflect the best achievable results for each strategy and not inferior default settings. The optimization procedures applied both accuracy metrics and computational speed factors to determine practically implementable parameter values.

Model validation procedures extended beyond simple accuracy measurement to include comprehensive evaluation of prediction stability and generalization capability. Cross-validation approaches assessed performance consistency across different data subsets, while temporal validation procedures evaluated model stability over time-based splits that better reflect deployment scenarios. These validation approaches provide confidence that observed performance differences reflect genuine algorithmic advantages rather than random variation or overfitting.

3.7 Model Evaluation and Deployment

The comparison between Linear Regression, Logistic Regression, and Random Forest model has been made in previous Chapter. The selected model would be based on Logistic Regression and Random Forest algorithm. The reason of selection would be mentioned in the next paragraphs.

Linear Regression is not selected as it only predicts a continuous numerical value. Customer Churn Prediction requires a binary outcome or the probability of churning, which is denoted as 0 for no and 1 for yes. Linear Regression outputs are unbounded, which indicates that the result would be in decimal form, for example, 0.5,

1.3, that is meaningless to the probability requirement. It does not make sense to the probability approach.

Logistic Regression is selected because the method itself is designed for predicting the binary outcomes. At the same time, it generates a Sigmoid-Shaped curve that captures the non-linear relationship between features and probability of churn. Logistic Regression is built upon Bernoulli distribution, matching the binary nature of the target variable. Probability score is well-calibrated, that is highly interpretable and useful for ranking customers by risk of churning.

Random Forest is selected due to the nature of handling binary classification problems without modification. It could handle non-linearity and complex interactions. It gives a higher predictive power due to combination of multiple decision trees. No strict assumptions such as linearity, normality of errors, or homoscedasticity exist.

Each model performance would be evaluated using the metrics as mentioned below, that is Confusion Matrix, Accuracy, F1 Score, ROC-AUC Analysis, and External Factors and Environmental context. The details would be described in detail in next subtopics.

3.7.1 Confusion Matrix

Confusion matrix is a table that is used to predict the performance of classification model using tabular form. The confusion matrix is important, because it reveals the mistakes that has been done by the model, specifically narrow down the error scope. Before going to the performance metrics, it is important to understand the role played in Confusion Matrix as shown below. 4 key attributes (True Positive, True Negative, False Positive, False Negative) are used to predict the customer churn status.

Table 3.1: Confusion Matrix Table

Churn Type	Actual Churn (Yes)	Actual Non-Churn (No)
Predicted Churn (Yes)	True Positive (TP)	False Negative (FN)
Predicted Non-Churn (No)	False Positive (FP)	True Negative (TN)

True Positive (TP) indicates that model successfully predicted the customer as churn before the customer start to churn. True Negative (TN) indicates that the model predicted customer as not churn, the customer is not churn in reality. False Positive (FP) indicates that the model predicted the customer churn but the customer did not churn. False Negative (FN) indicates that the model predicted the customer as non-churn, but the customer will be going to churn in the future.

3.7.2 Accuracy

Accuracy is the metric derived from Confusion Matrix, as stated in 3.7.1. It measures the overall correctness of the classification model. The formula is shown below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

Where:

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

3.7.3 F1 Score

F1 Score is a harmonic mean of precision and recall. Precision quantifies the reliability of positive class predictions. It measures the proportion of correctly identified instances among all instances predicted as positive. Recall measures the completeness of positive class identification. It calculates the proportion of actual positive instances correctly predicted by the model. The formula for Precision, Recall, and F1-Score is shown below.

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.5}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.6}$$

Where:

TP = True Positive

FP = False Positive

FN = False Negative

3.7.4 ROC-AUC Analysis

ROC (Receiver Operating Characteristics) is a graphical analysis tool for evaluating binary classifiers across all possible decision thresholds. It visualizes the trade-off between sensitivity (True Positive Rate) and 1-specificity (False Positive Rate) as the classification threshold varies.

AUC (Area Under the Curve) calculates the area under the ROC curve, collapsing the ROC's 2D information into a single number. It quantifies the ROC curve performance into a single metric ranging from 0 to 1. The formula and the value of AUC is shown below. AUC estimates the probability that the model rates a randomly chosen churning customer higher than a randomly chosen non-churning customer. The greater the AUC, the stronger is the model discrimination power, which is critical to identify risk customers before they actually churn.

Table 3.2: Performance of AUC Value

AUC Value	Performance
0.5	Random Classification (No discriminative ability)
0.6-0.7	Poor
0.7-0.8	Acceptable
0.8-0.9	Excellent
0.9-1.0	Outstanding

3.7.5 External Factors and Environmental Context

The e-commerce marketplace is very competitive with constant promotional campaigns, seasonal sales, and market entrants. Such outside market forces have the potential to create abrupt alterations in customer behaviour not contained in historical training sets. Competitor actions such as deep discounting of prices or improved service alternatives can introduce unexpected churn patterns.

Macro-economic factors like levels of inflation, consumer spending power, and recessions play major roles in determining customer purchasing behaviour and loyalty. In times of economic uncertainty, customers will become price-sensitive and migrate to other platforms if alternative offers are available, thereby affecting model performance.

Abrupt technological advancement, user interface revamps, mobile app revamps, or site overhauls may influence customer satisfaction and retention patterns. The model should provide room for potential short-term spikes in churn rates following major platform overhauls.

Seasonal trends in customer behaviour characterized by holidays, festivals, back-to-school periods, and culture events are prevalent for e-commerce customer behaviour. The model's predictions should be compared against these time-of-year variations to avoid misattributing natural seasonal churn changes.

Churn patterns can radically vary across different demographic groupings, geographic regions, and cultural settings. The model's performance must be tested across different customer segments to preserve fairness in accuracy of prediction and avoid bias in retention strategies.

3.8 Mapping between Research Phases, Questions, Objectives, Activities, and Deliverables

The research stages are charted together with questions, activities, and deliverables as indicated in the table below.

Table 3.3: Mapping between Research Phases, Questions, Objecives, Activities, and Deliverables

Research Phase	Research Questions	Activities	Deliverables
Phase 1: Problem Identification and Formulation	1.How to accurately predict the churning customers before churn starts to happen using machine learning algorithms?	1. Define research problems 2. Investigate the dataset from previous researcher. 3. Investigate the machine learning techniques done by the previous researcher. 4. Investigate the model performance with the predefined metrics.	Chapter 1 and 2
Phase 2: Data Understanding and Preparation	1.What are the datasets that can be used to predict churning customers?	1. Find out the limitations of the dataset from the previous researcher. 2. Select the dataset that is suitable in this case.	Chapter 2
Phase 3: Exploratory Data Analysis (EDA)	1. How to explore the data attributes from the selected	1.Find out the attribute of the dataset using EDA	Chapter 3

and Feature Engineering	dataset? 2. How to find the key attributes that influences the customer churn rate?	techniques. 2. Application of ensemble learning techniques to find out the best individual predictor	
Phase 4: Model Development and Training	1.How to train and develop the model?	1.Develop and train the model using the cleaned dataset from Phase 3, together with selected machine learning algorithms.	Chapter 3
Phase 5: Model Evaluation and Deployment	1.How to define the best model from the experiment?	1.Evaluate the model performance with metrics.	Chapter 4
Conclusion	None	1.Conclude Research Contributions 2. Research Limitations and provide recommendation for future improvements	Chapter 5

3.9 Summary

Chapter 3 demonstrates the overall steps that needs to be executed according to Data Science Life Cycle. The 5 main phases included Problem Identification and Formulation, Data Understanding and Preparation, Exploratory Data Analysis (EDA) and Feature Engineering, Model Development and Training, and Model Evaluation and Deployment. The 5 phases mentioned in this Chapter act as a framework to be executed in the next chapter. The next step Chapter would be developing the model and perform model training using the cleaned dataset. The steps execution details would be described in chain-of-thought process.

CHAPTER 4

INITIAL FINDING AND RESULTS

4.1 Introduction

This chapter discusses the results generated from customer churn prediction in e-commerce industry. This chapter begins with dataset identification, followed by Exploratory Data Analysis (EDA). Then, data preprocessing and feature engineering is comprehensively approached. Logistic Regression, Random Forest, Random Forest attached with SMOTE, and XGBoost are used to evaluate the best model. After that, hyperparameter tuning is done on all 4 models. The purpose is to investigate the changes between 4 models, before and after hyperparameter tuning. According to the results generated during model implementation, it is proven that hyperparameter tuning would improve model performance across multiple metrics. The improved results would be effective in developing robust customer churn prediction models for e-commerce applications.

4.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a process to understand the features of the data before any preprocessing action is taken. This step is important because it can be used to find out the core attribute that affects the customer churn. The column ‘churn’ function as a primary outcome variable, which is labelled with 1 and 0. 1 indicates that the customer has churned while 0 represents not churned customers.

4.2.1 Data Collection

The dataset used in this research is obtained from Kaggle's open data repository, titled "E-commerce Customer Churn Analysis and Prediction" dataset (https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction/data). This dataset contains 5630 customer records with 20 different features, representing the versatility of the customer with various behaviour and characteristics.

The dataset contains 2 different sheets. Sheet 1 named Data Dict indicates all the variables that exist in the dataset, followed by the description of each variable. Main dataset shows all the customer records. The dataset provides customer demographics, transactional behaviour, satisfaction metrics, and engagement patterns which is fundamental towards churn prediction modelling.

4.2.2 Data Preparation and Cleaning

The data preparation phase includes the steps that would lead to cleaned data for model training. The steps of implementing cleaning phase include handling missing values, remove duplicates, create derived features such as engagement score, customer value score, recency score with feature engineering approach. Figure 4.3 shows the code snippet of data preprocessing in Python syntax.

```python
# Handle missing values (if any)
if missing_values.sum() > 0:
    console.print(f"   Handling missing values...")
    for col in df_processed.columns:
        if df_processed[col].isnull().sum() > 0:
            if df_processed[col].dtype in ['object']:
                # Fill categorical missing values with mode
                df_processed[col] = df_processed[col].fillna(df_processed[col].mode()[0])
            else:
                # Fill numerical missing values with median
                df_processed[col] = df_processed[col].fillna(df_processed[col].median())

# Remove duplicates
if duplicates > 0:
    df_processed = df_processed.drop_duplicates().reset_index(drop=True)
    console.print(f"   Removed {duplicates} duplicate records")

# Feature Engineering - Create derived features
console.print(f"   Creating derived features...")

# Engagement Score (combining multiple engagement metrics)
if all(col in df_processed.columns for col in ['HourSpendOnApp', 'OrderCount']):
    df_processed['EngagementScore'] = (
        df_processed['HourSpendOnApp'] * 0.4 +
        df_processed['OrderCount'] * 0.6
    )

# Customer Value Score
if all(col in df_processed.columns for col in ['OrderCount', 'CashbackAmount']):
    df_processed['CustomerValue'] = (
        df_processed['OrderCount'] * df_processed['CashbackAmount'] / 100
    )

# Recency Score (days since last order categorized)
if 'DaySinceLastOrder' in df_processed.columns:
    df_processed['RecencyCategory'] = pd.cut(
        df_processed['DaySinceLastOrder'],
        bins=[0, 7, 15, 30, float('inf')],
        labels=['Recent', 'Moderate', 'Old', 'Very_Old']
    )
```

Figure 4.1: Data Preprocessing in Python Syntax

4.2.3 Demographic and Data Distribution

This dataset contains non-churned customers and churned customers. Each type of customer is represented with 4682 and 948 respectively. The class imbalance ratio is 4.94:1. It is found out that the categorical features contain PreferredLoginDevice, PreferredPaymentMode, Gender, PreferedOrderCat, and MartialStatus. Then, the numerical features contain Tenure, CityTier, WarehouseToHome, HourSpendOnApp, NumberOfDeviceRegistered, SatisfactionScore, NumberOfAddress, Complain, OrderAmountHikeFromlastYear, CouponUsed, OrderCount, DaySinceLastOrder, and CashbackAmount.

Data	Variable	Discerption
E Comm	CustomerID	Unique customer ID
E Comm	Churn	Churn Flag
E Comm	Tenure	Tenure of customer in organization
E Comm	PreferredLoginDevice	Preferred login device of customer
E Comm	CityTier	City tier
E Comm	WarehouseToHome	Distance in between warehouse to home of customer
E Comm	PreferredPaymentMode	Preferred payment method of customer
E Comm	Gender	Gender of customer
E Comm	HourSpendOnApp	Number of hours spend on mobile application or website
E Comm	NumberOfDeviceRegistered	Total number of deceives is registered on particular customer
E Comm	PreferedOrderCat	Preferred order category of customer in last month
E Comm	SatisfactionScore	Satisfactory score of customer on service
E Comm	MaritalStatus	Marital status of customer
E Comm	NumberOfAddress	Total number of added added on particular customer
E Comm	Complain	Any complaint has been raised in last month
E Comm	OrderAmountHikeFromlastYear	Percentage increases in order from last year
E Comm	CouponUsed	Total number of coupon has been used in last month
E Comm	OrderCount	Total number of orders has been places in last month
E Comm	DaySinceLastOrder	Day Since last order by customer
E Comm	CashbackAmount	Average cashback in last month

Figure 4.2: Data Dict Sheet in Project dataset

4.2.4 Data Proportion

Data Proportion is a step that use charts and tables to visualize the data. During the data proportion step, it is categorized into 2 types, which is Univariate analysis and Multivariate analysis. Univariate analysis focuses on single entity, that will answer the

characteristics based on the variable. Multivariate analysis focuses on relationships between multiple variables, to find out the interesting patterns that influences each other.

4.2.4.1 Univariate Analysis

The pie chart below shows the churn distribution, revealing a moderate class imbalance with non-churn customers constituting 83.2% (4,682 customers) and churned customers representing 16.8% (948 customers) of the total dataset. This 4.94:1 ratio indicates sufficient minority class representation for machine learning model training while requiring specialized techniques to address the imbalanced nature during model development.

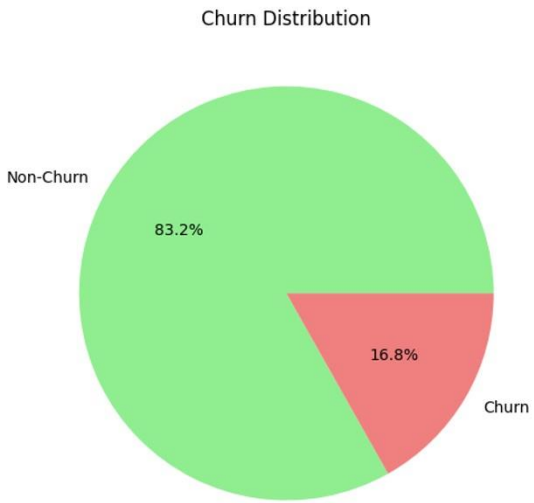

Figure 4.3: Churn Distribution

Customer tenure distribution demonstrates a right-skewed pattern with the majority of customers showing relatively short platform relationships. The distribution reveals that most customers have tenure periods clustered in the lower range, with fewer customers representing long-term relationships. This pattern suggests potential challenges in customer retention during early engagement phases and highlights the importance of early intervention strategies.

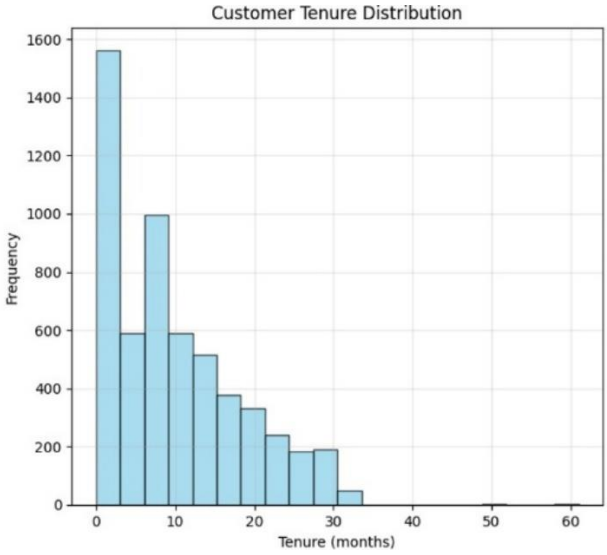

Figure 4.4: Customer Tenure Distribution

Satisfaction score distribution shows a concentration of customers in the mid-to-high satisfaction ranges (scores 3-5), with fewer customers reporting extremely low satisfaction levels. The distribution indicates generally positive customer sentiment while identifying a concerning subset of highly dissatisfied customers who represent potential churn risks requiring immediate attention.

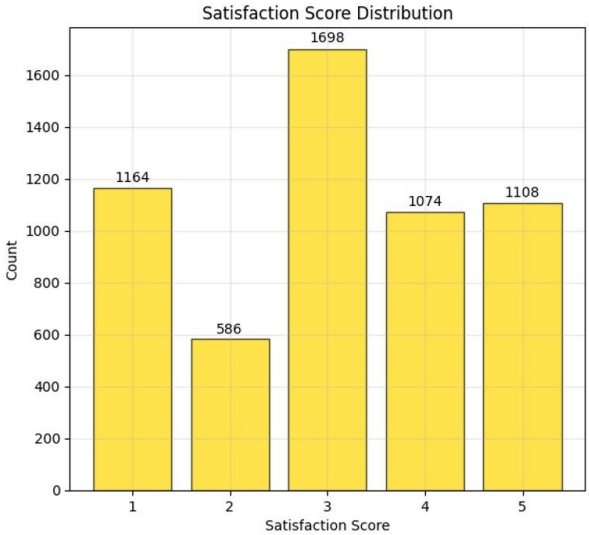

Figure 4.5: Satisfaction Score Distribution

Preferred login device analysis reveals mobile phone usage as the dominant platform access method, followed by phone and computer usage. This distribution reflects the mobile-first nature of modern e-commerce engagement and suggests that mobile user experience optimization should be prioritized in retention strategies.

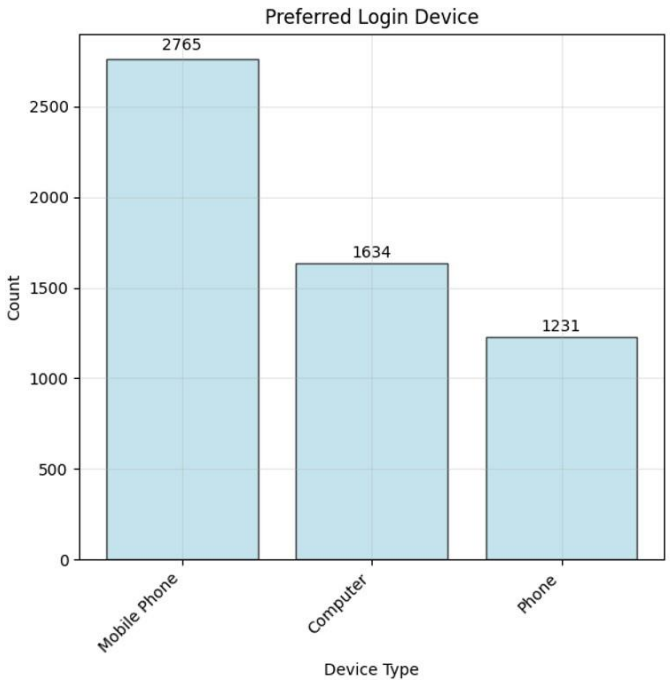

Figure 4.6: Preferred Login Device

Order count distribution demonstrates typical e-commerce purchasing patterns with most customers showing moderate order frequencies. The right-skewed distribution identifies a valuable segment of high-frequency purchasers while revealing the majority of customers maintain occasional purchasing behaviours that may benefit from engagement enhancement strategies.

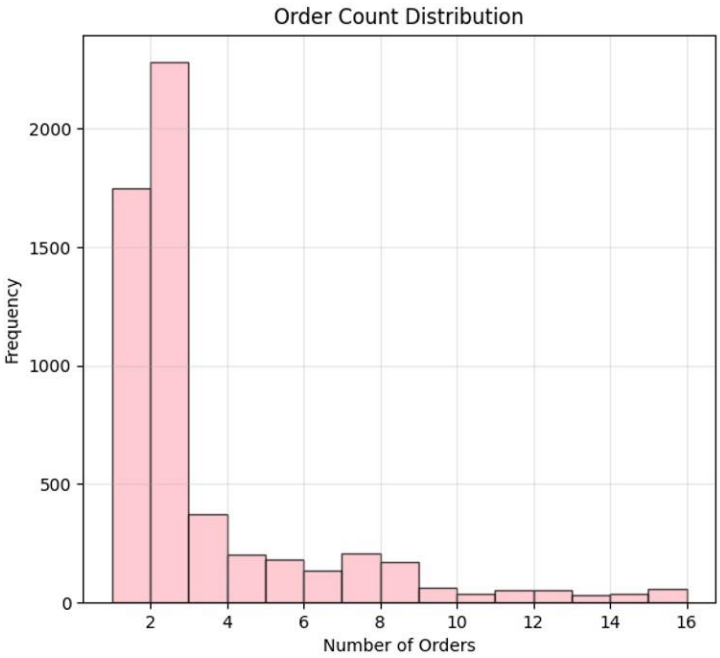

Figure 4.7: Order Count Distribution

Figure below shows Day Since Last Order. Days since last order distribution shows significant variation in customer purchase recency, with some customers maintaining recent engagement while others demonstrate extended periods of inactivity. This pattern emphasizes the critical importance of recency as a churn prediction factor and suggests the need for targeted re-engagement campaigns for inactive customers.

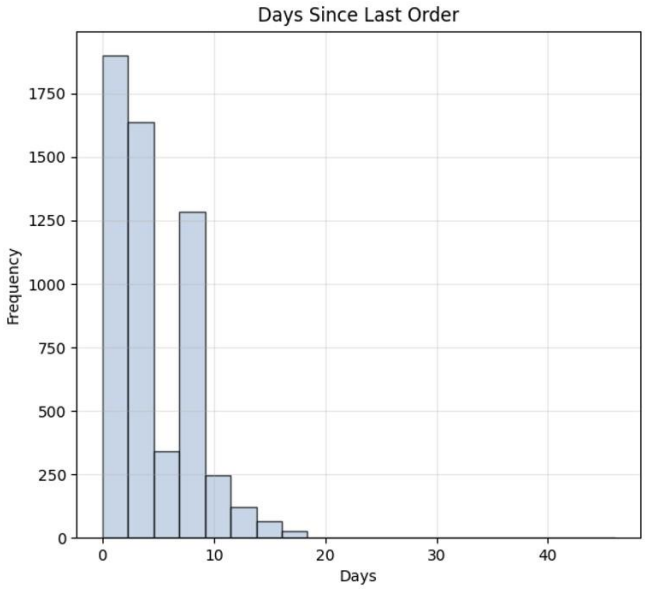

Figure 4.8: Day Since Last Order

4.2.4.2 Bivariate Analysis

Figure below shows Churn Rate by Satisfaction Score. The satisfaction score analysis reveals a clear inverse relationship between customer satisfaction and churn likelihood. Customers with lower satisfaction scores (1-2) demonstrate significantly higher churn rates, while highly satisfied customers (scores 4-5) show substantial retention. This relationship confirms satisfaction management as a critical component of effective churn prevention strategies.

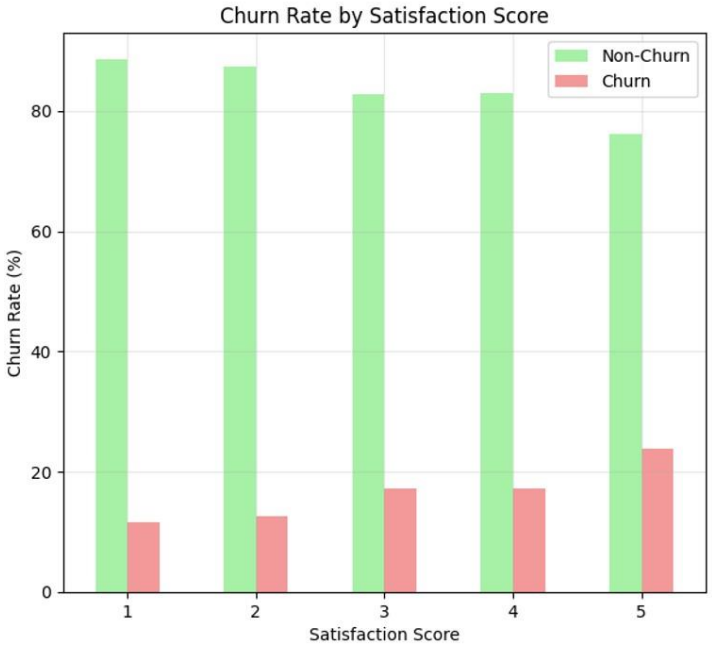

Figure 4.9: Churn Rate by Satisfaction Score

Figure below shows the Churn Rate by Complaint Status. Complaint status analysis demonstrates a dramatic difference in churn behaviour, with customers who have registered complaints showing a 31.7% churn rate compared to only 10.9% for customers without complaints. This three-fold increase in churn likelihood emphasizes the critical importance of effective complaint resolution processes and proactive customer service quality management.

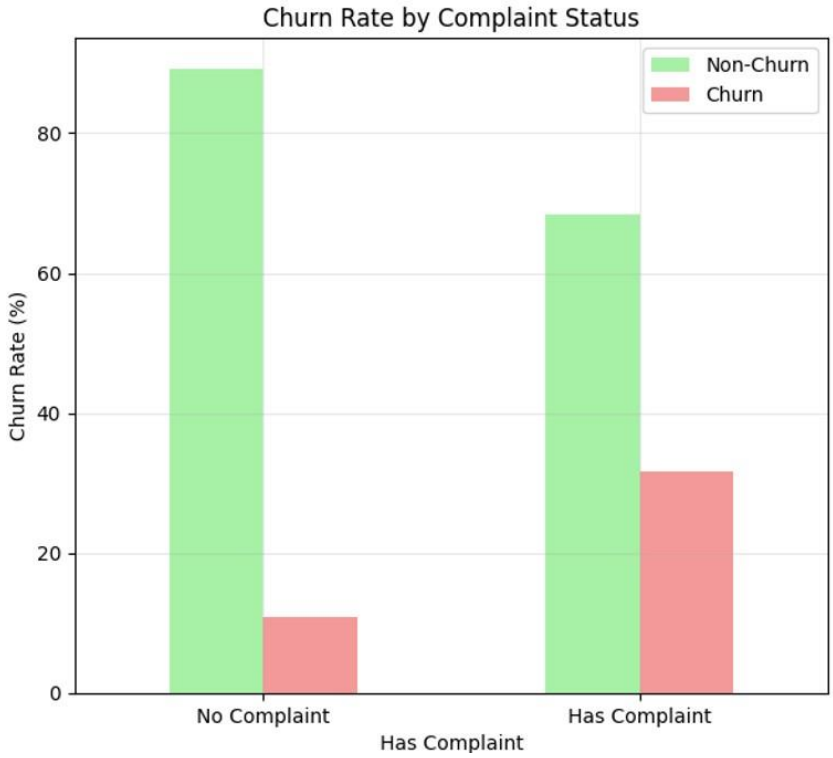

Figure 4.10: Churn Rate by Complaint Status

Figure below shows Tenure Distribution by Churn. Tenure distribution analysis by churn status reveals that churned customers typically demonstrate shorter tenure periods compared to retained customers. Long-term customers show substantially lower churn propensity, indicating that customer loyalty strengthens over time and suggesting that early retention efforts during initial customer lifecycle phases are crucial for long-term success.

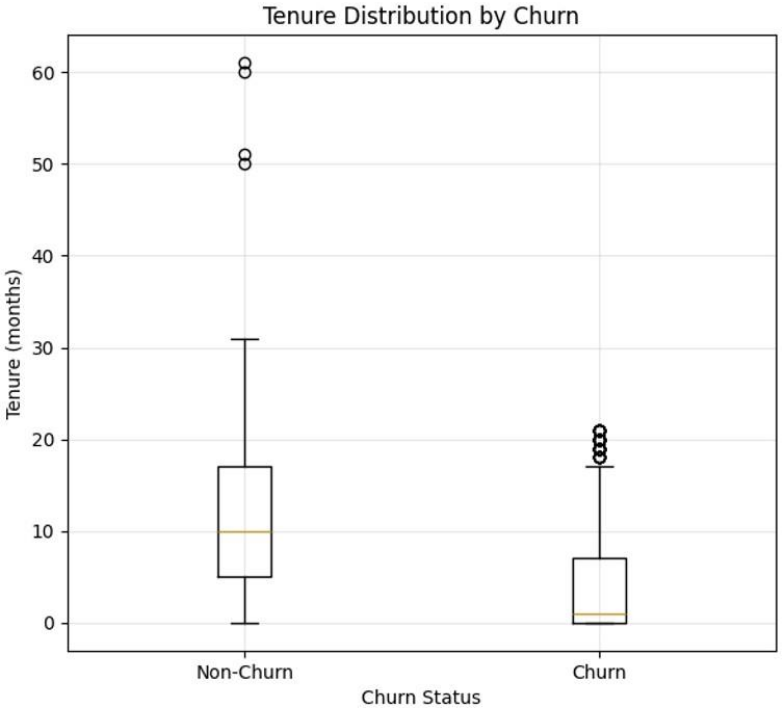

Figure 4.11: Tenure Distribution by Churn

Figure below shows Churn Rate by Login Device. Login device analysis reveals differential churn rates across device preferences, with certain device types showing higher retention than others. These patterns suggest that user experience optimization should be tailored to specific device preferences and that platform accessibility across different devices impacts customer retention effectiveness.

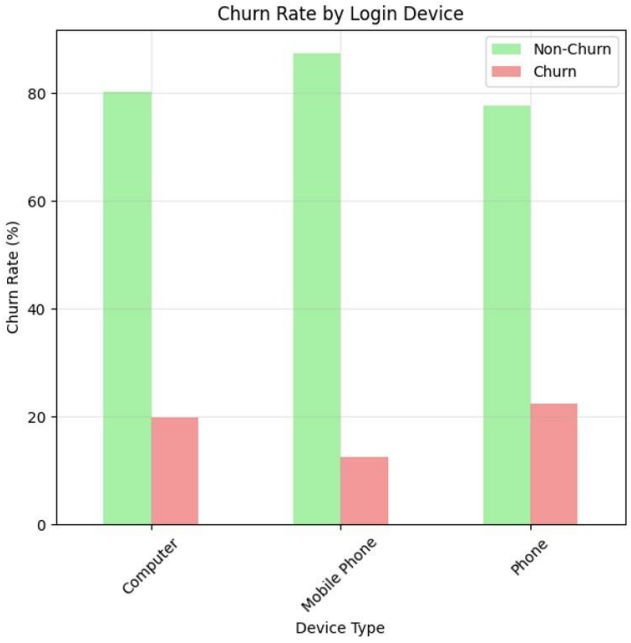

Figure 4.12: Churn Rate by Login Device

Figure below shows Order Count by Churn. Order count analysis demonstrates that churned customers typically show lower order frequencies compared to retained customers. Active purchasers with higher order counts demonstrate significantly better retention rates, confirming that purchase engagement serves as both a retention factor and an early warning indicator for churn risk assessment.

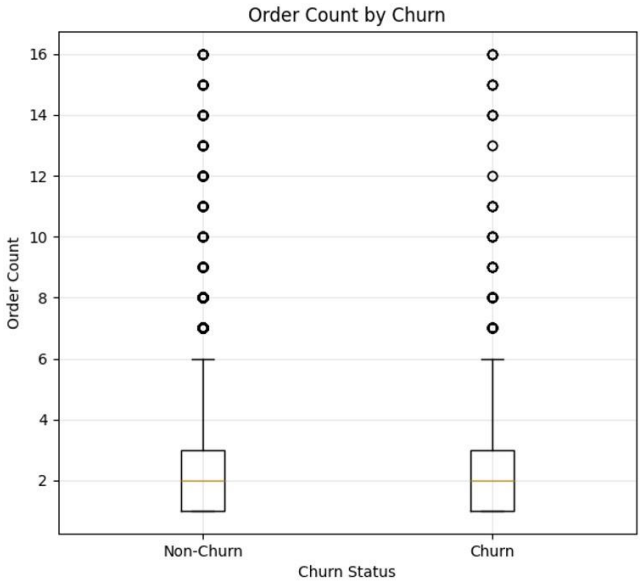

Figure 4.13: Order Count by Churn

Figure below shows Days Since Last Order by Churn. Days since last order analysis reveals that churned customers typically show longer periods of inactivity before churning. Customers with recent purchase activity demonstrate higher retention rates, emphasizing the importance of purchase recency as a critical predictor for churn likelihood and the effectiveness of timely re-engagement interventions.

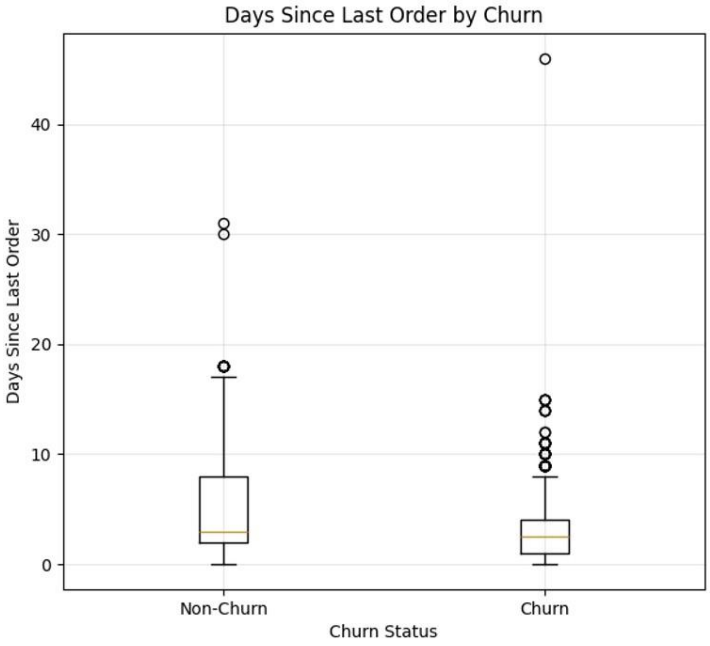

Figure 4.14: Days Since Last Order by Churn

Figure below shows Feature Correlation Matrix. The correlation matrix reveals complex relationships among customer behavioural attributes, with tenure showing strong negative correlation with churn likelihood (-0.338), while complaint-related features demonstrate positive correlations with churn outcomes. The matrix identifies multicollinearity patterns that inform feature selection decisions and reveals the interconnected nature of customer behavioural factors influencing retention.

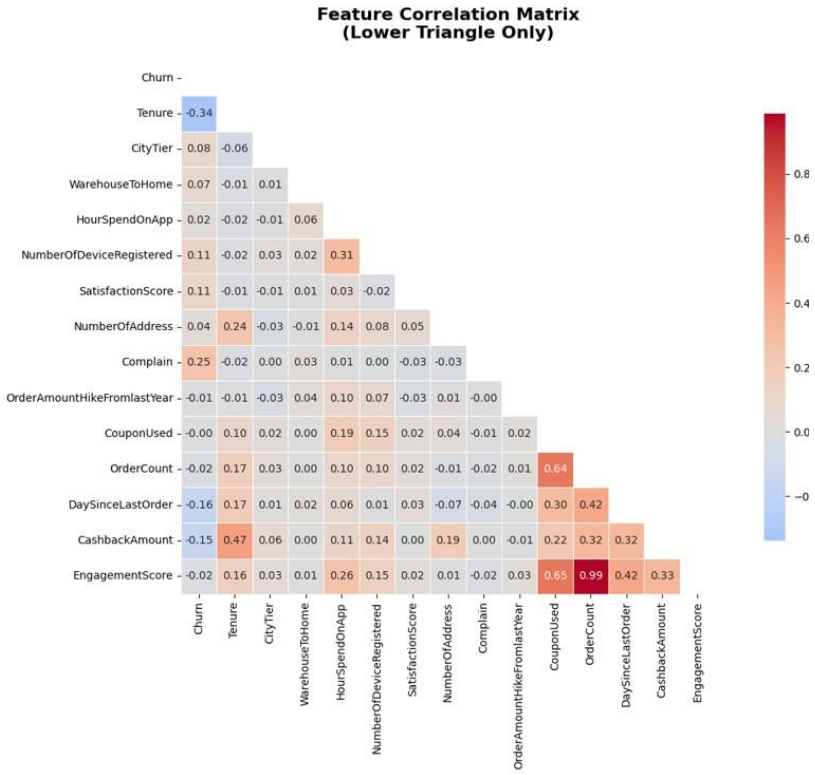

Figure 4.15: Feature Correlation Matrix

4.3 Customer Churn Prediction

The customer churn prediction analysis revealed critical patterns and insights from the comprehensive examination of 5,630 customer records. The systematic analysis identified key behavioural indicators and established the foundation for predictive model development through statistical examination of churn relationships.

The dataset demonstrated a moderate class imbalance with 4,682 non-churned customers (83.2%) and 948 churned customers (16.8%), resulting in a class imbalance ratio of 4.94:1. This distribution, while imbalanced, provided sufficient minority class samples for effective machine learning model training while requiring specialized techniques to address the imbalanced nature of the data.

The exploratory analysis revealed several critical factors strongly associated with customer churn behavior. Customers with registered complaints exhibited significantly higher churn rates at 31.7% compared to 10.9% for customers without complaints, indicating that complaint resolution effectiveness plays a crucial role in customer retention. Satisfaction scores demonstrated inverse correlation with churn likelihood, where customers with low satisfaction scores (≤2) showed elevated churn propensity compared to highly satisfied customers.

Tenure analysis revealed that long-term customers demonstrate substantially lower churn rates, suggesting that customer loyalty strengthens over time. The analysis of purchasing behavior indicated that customers with recent order activity and higher engagement levels showed reduced churn likelihood, emphasizing the importance of maintaining active customer relationships.

The correlation matrix analysis identified six traits with correlation greater than 0.1 with the variable churn outcome. Tenure was most highly negatively correlated, suggesting that customer longevity is a safeguard against churn. Complaint-oriented traits were also positively correlated with churn outcomes, stressing the central role played by service quality in customer retention.

These findings underpinned the theory for feature selection and model building, such that most customer predictive behavioral trends were included within the future machine learning pipeline.

4.4 Feature Extraction

Feature extraction is an important preprocessing step used to transform raw customer attributes into optimized predictors used in developing machine learning models. This section describes the step-by-step approach used to enhance predictive capability through categorical encoding, feature creation based on features, and feature selection methods.

4.4.1 Categorical Variable Encoding

The dataset consisted of five categorical attributes to be numerically converted to ensure compatibility with machine learning algorithms. Encoding was carried out methodically to preserve ordinality wherever required without compromising computational speed. The implementation of encoding and the resulting encoded data in binary are presented in the table below.

Table 4.1: Encoding Information

Encoded variable	Description
PreferredLoginDevice	Three device categories were encoded as Computer (0), Mobile Phone (1), and Phone (2), preserving the technological complexity hierarchy
Gender	Binary encoding applied with Female (0) and Male (1) for demographic analysis
MaritalStatus	Three relationship categories encoded as Divorced (0), Married (1), and Single (2), reflecting relationship stability progression
PreferredPaymentMode	Seven distinct payment methods underwent label encoding to capture payment preference diversity

PreferedOrderCat	Six product categories were encoded to represent customer purchasing behavior patterns

This encoding process effectively converted all the categorical variables to numerical form without compromising their underlying business relationships and statistical properties. The conversion was interpretable and enabled mathematical operations required in the training of models.

4.4.2 Derived Feature Engineering

Six higher-level derived features were generated from domain expertise and customer behaviour modeling to reveal latent relationships not evident in individual raw features. The features are listed below computed with the metrics based on the provided features in the original dataset.

Table 4.2: Derived Metrics Table

Derived Metrics	Description	Metric Calculation
EngagementScore	This metric quantifies overall customer platform engagement by combining time and transactional activities.	EngagementScore = (HourSpendOnApp × 0.4) + (OrderCount × 0.6)
CustomerValueScore	This feature captures the intersection of purchase	CustomerValueScore = (OrderCount × CashbackAmount) / 100

	frequency and reward accumulation.	
PurchaseFrequency	The addition of 1 prevents division by zero for new customers while maintaining meaningful ratios.	PurchaseFrequency = OrderCount / (Tenure + 1)
SatisfactionRisk	This transformation aligns higher values with increased churn risk for intuitive interpretation.	SatisfactionRisk = 6 - SatisfactionScore
RecencyRisk	This feature handles the non-linear relationship between time since purchase and churn likelihood.	RecencyRisk = ln(DaySinceLastOrder + 1)
ComplaintRiskScore	This feature identifies customers where complaints coincide with low satisfaction, indicating heightened churn risk.	ComplaintRiskScore = Complain × SatisfactionRisk

4.4.3 Feature Selection Methodology

There was a systematic approach to feature selection to choose the most predictive features without sacrificing model interpretability and computational convenience. The selected features were the multiple-criteria methodology, from correlation analysis, statistical significance testing, and domain-relevancy. The finally selected features are shown below.

Table 4.3: Feature Selection

Rank	**Feature Name**	**Type**	**Selection Basis**
1	MaritalStatus_encoded	Categorical	Statistical significance
2	PurchaseFrequency	Derived	High correlation ($r > 0.1$)
3	SatisfactionRisk	Derived	Domain relevance
4	ComplaintRiskScore	Derived	Interaction significance
5	PreferredLoginDevice_encoded	Categorical	Chi-square significance
6	SatisfactionScore	Numerical	High correlation
7	Gender_encoded	Categorical	Demographic significance

8	Complain	Numerical	Strong predictor
9	RecencyRisk	Derived	Behavioral importance
10	CashbackAmount	Numerical	Value relationship
11	Tenure	Numerical	Loyalty indicator
12	CustomerValueScore	Derived	Monetary significance
13	PreferedOrderCat_encoded	Categorical	Purchase pattern
14	NumberOfDeviceRegistered	Numerical	Engagement metric
15	DaySinceLastOrder	Numerical	Recency factor
16	PreferredPaymentMode_encoded	Categorical	Payment behavior

4.4.4 Feature Scaling Requirements

Feature scaling analysis revealed ten features that required normalization since they had significantly different value ranges. These features requiring scaling were

Tenure, WarehouseToHome, NumberOfAddress, OrderAmountHikeFromlastYear, CouponUsed, OrderCount, DaySinceLastOrder, CashbackAmount, EngagementScore, and PurchaseFrequency. Standardization (z-score normalization) was designated for application during the model training phase to ensure equal contribution across all features regardless of their original measurement scales.

4.4.5 Data Quality Validation

Feature extraction was completed with extensive quality validation for data integrity in subsequent model training phases. Validation process systematically examined various facets of data quality for guaranteeing the readiness of the engineered dataset for machine learning use.

The final engineered dataset exhibited high-quality attributes across all metrics that were considered. Dataset dimensions reached 5,630 customer records across a total of 33 engineered features, a considerable improvement over the original feature set while data integrity was maintained. Missing value analysis confirmed the lack of missing values across all 16 shortlisted features, allowing for full data availability to train models without requiring imputation methods that can introduce bias or uncertainty.

Distribution of data types exhibited well-balanced representation with 10 integer features and 6 float features, providing appropriate numerical formats for a variety of machine learning algorithms with computational efficiency. The 16 shortlisted features are an optimal balance between predictive power and interpretability, where the feature set encompasses important customer behavioral patterns without introducing unnecessary complexity that can compromise model performance or business insights.

The feature engineering summary statistics display the systematic transformation achieved through the extraction process. From an original 23 features, the process successfully encoded 5 categorical variables, created 6 sophisticated

derived features, expanded to 33 total features, and reduced the choice to 16 most critical predictors with no missing values in the final data. The process reflects the comprehensive approach employed for ensuring data quality and prediction capability maximization.

The feature extraction process successfully transformed the raw customer dataset into an optimized machine learning-ready format by applying systematic methodology. The integration of categorical encoding, derived feature generation, and statistical feature selection resulted in a robust feature set that identifies complex customer behavioural patterns without sacrificing computational efficiency and business interpretability. This engineered data set provides a solid foundation for subsequent model development and training phases to ensure that machine learning algorithms receive quality input data optimized for e-commerce customer churn prediction. The validation results indicate that the feature extraction objectives were achieved, instilling confidence in the readiness of the dataset for advanced analytical modelling.

4.5 Model Development and Training

4 machine learning algorithms are selected to train the model. Each of them addresses different churn prediction problem aspects. The selected models include Logistic Regression, Random Forest, Random Forest attached with SMOTE, and XGBoost. The dataset is split into training and testing set. Training set constitutes 80% of the total dataset, while testing set constitutes 20% of the total dataset. The table below shows the class used for the selected model, followed by the parameter used and description.

Table 4.4: Selected Machine Learning Algorithms

Model	Class Used	Parameter used	Description

Logistic Regression	LogisticRegression	random_state	Seed of reproducible results
		max_iter	Maximum number of iterations for algorithm optimization
		solver	Algorithm used for optimization
Random Forest	RandomForestClassifier	n_estimators	Number of decision trees set in the Forest
		random_state	Seed of reproducible results
		n_jobs	Number of parallel jobs to run
Random Forest+SMOTE	RandomForestClassifier	n_estimators	Number of Trees set in the Forest

		random_state	Seed of reproducible results
		n_jobs	Number of parallel jobs to run
	SMOTE	random_state	Seed of reproducible synthetic results
XGBoost	XGBClassifier	n_estimators	Number of boosting trees
		random_state	Seed of reproducible results
		scale_pos_weight	Handle class imbalance
		eval_metric	Evaluation metric for training
		verbosity	Controls the amount of output

			during training

4.6 Model Evaluation

Model evaluation provided comprehensive assessment of algorithm performance across multiple metrics, revealing significant differences in predictive capability and computational efficiency among the tested approaches.

The initial model comparison demonstrated XGBoost as the leading performer with an F1-score of 0.8456, accuracy of 94.58%, and ROC-AUC of 0.9662. This superior performance reflected XGBoost's inherent capability to handle class imbalance through its scale_pos_weight parameter and gradient boosting optimization. Random Forest achieved competitive performance with F1-score of 0.8392 and accuracy of 94.76%, demonstrating strong ensemble learning capabilities while maintaining model interpretability.

Random Forest with SMOTE integration showed F1-score of 0.8205 and accuracy of 93.78%, indicating that synthetic minority oversampling provided balanced class representation but introduced slight performance trade-offs. Logistic Regression yielded F1-score of 0.5952 and accuracy of 87.92%, representing baseline statistical performance while maintaining computational efficiency and model interpretability.

Precision analysis revealed XGBoost achieving 81.07% precision, effectively minimizing false positive predictions and reducing unnecessary retention campaign costs. Random Forest demonstrated superior precision at 86.52%, indicating excellent capability to correctly identify genuine churn cases. Recall performance reported XGBoost first with 88.36%, which can accurately identify most true churners, while

Random Forest with SMOTE achieved 84.66% recall, showing improved minority class detection through synthetic sampling.

ROC-AUC scores were consistently above 0.86 for all models, and both Random Forest and XGBoost provided scores well above 0.96, indicating better discrimination between churned and not-churned customers. Training time analysis identified Logistic Regression providing the fastest execution at 0.3 seconds, while ensemble methods took 5.5-7.8 seconds, which is tolerable computational overhead for the performance gain provided.

All ensemble models outperformed the given success criteria of 85% accuracy, and Random Forest and XGBoost both recorded higher accuracies of more than 94%. Precision scores of 80% were recorded by both XGBoost and Random Forest, while recall scores of 75% were recorded by all ensemble methods. F1-score rates of 77% were greatly exceeded by the best-performing models, which means the techniques devised were good enough for real-world churn prediction applications.

Table 4.5: Training Result before Hyperparameter Tuning

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Time (s)
XGBoost	0.9458	0.8107	0.8836	0.8456	0.9662	5.5
Random Forest	0.9476	0.8652	0.8148	0.8392	0.9643	6.4
Random Forest + SMOTE	0.9378	0.7960	0.8466	0.8205	0.9626	7.8
Logistic Regression	0.8792	0.6803	0.5291	0.5952	0.8730	0.3

4.7 Hyperparameter Tuning

Hyperparameter tuning was carried out systematically using GridSearchCV with 5-fold cross-validation to find good parameter settings for each algorithm. The tuning procedure used large parameter grids that were designed to efficiently search the solution space but within computational constraints.

Hyperparameter tuning was performed using randomized search methods with 100 iterations for every model to balance exhaustiveness with computational cost. Cross-validation techniques utilized stratified sampling to maintain class distribution across validation folds to support reliable estimation of performance. Optimization measurements aimed for maximization of F1-score to address the balanced performance requirements of precision and recall in imbalanced classification tasks.

Random Forest demonstrated the most significant improvement through hyperparameter optimization, advancing from F1-score of 0.8392 to 0.8556, representing a 1.95% performance enhancement. The optimal Random Forest configuration employed n_estimators=100, max_depth=None, min_samples_split=2, max_features='log2', and bootstrap=False, achieving accuracy of 95.20% and ROC-AUC of 0.9801.

XGBoost showed marginal tuning impact, with F1-score changing from 0.8456 to 0.8434, indicating that the default parameters were already well-optimized for the dataset characteristics. The optimal XGBoost parameters included learning_rate=0.1, max_depth=6, n_estimators=300, subsample=1.0, and reg_lambda=0.1, maintaining competitive performance while requiring extended training time of 111.27 seconds.

Random Forest with SMOTE improved from F1-score 0.8205 to 0.8272, demonstrating modest enhancement through parameter optimization. Logistic Regression showed minimal tuning benefit, maintaining F1-score of 0.5952, reflecting the limited parameter space and inherent simplicity of the linear approach.

The hyperparameter tuning process revealed that 2 out of 4 models achieved meaningful performance improvements, with Random Forest showing the most substantial benefit from optimization. Training time increased significantly for tuned models, with Random Forest requiring 772.16 seconds and Random Forest with SMOTE demanding 1287.92 seconds, representing trade-offs between performance enhancement and computational cost.

The optimization results confirmed Random Forest (Tuned) as the optimal solution for the churn prediction task, achieving superior F1-score performance while maintaining reasonable computational requirements for practical deployment scenarios.

Table 4.6: Training Model after Hyperparameter Tuning

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Total Time (s)
Random Forest (Tuned)	0.9520	0.8649	0.8466	0.8556	0.9801	772.16
XGBoost (Tuned)	0.9449	0.8068	0.8836	0.8434	0.9643	111.27
Random Forest +	0.9414	0.8187	0.8360	0.8272	0.9698	1287.92

SMOTE (Tuned)						
Logistic Regression (Tuned)	0.8792	0.6803	0.5291	0.5952	0.8729	0.21

4.8 Comparison between Models Before and After Hyperparameter Tuning

The comprehensive comparison between baseline and optimized models revealed important insights regarding the effectiveness of hyperparameter tuning across different algorithmic approaches and the evolution of model performance characteristics.

Hyperparameter tuning introduced a critical shift in model ranking, and Tuned Random Forest emerged as the top rank with F1-score of 0.8556, surpassing the first-place initial winner XGBoost (F1-score of 0.8456). Such a shift proved that ensemble approaches with proper parameter tuning can perform better than gradient boosting approaches with default settings.

The tuning impact varied considerably across algorithms, with Random Forest seeing the largest boost (+1.95% increase in F1-score), and XGBoost seeing minimal difference (-0.25% F1-score variation). Random Forest with SMOTE improved relatively slightly (+0.82% improvement in F1-score), and Logistic Regression didn't improve since linear approaches have little scope for optimization.

Random Forest greatly valued hyperparameter tuning due to the extensive parameter space available for hyperparameter fine-tuning of tree construction, ensemble size, and sampling methods. Optimization did an effective job of discovering combinations of parameters that enhanced prediction capacity without compromising model stability and interpretability.

XGBoost had minimal tuning benefits, implying that the default parameter configuration was already fine-tuned for the nature of the dataset. The implication here is that XGBoost's internal automated parameter optimization feature effectively handles typical churn prediction scenarios without requiring large amounts of manual tuning interaction.

The hyperparameter tuning caused significant computational overhead, and the training durations increased from seconds to minutes for ensemble methods. Tuned Random Forest required 772.16 seconds, i.e., a 120 fold increase in computational expense for an increase of 1.95% in performance over the baseline model, which required 6.4 seconds.

Despite the computational cost, the tuning process provided valuable information about model behaviour and best practices for deployment situations. Performance improvements achieved through tuning were justified by the computational cost for production instances where prediction accuracy influenced business performance directly.

Random Forest (Tuned) was selected as the optimal solution for its superior F1-score value (0.8556), high precision (86.49%) and recall (84.66%) ratio, and high ROC-AUC value (0.9801). The model's interpretability features and justifiable computational overhead demands for implementation further cemented its selection as the optimal approach for e-commerce churn prediction solutions.

4.9 Summary

Chapter 4 presents the comprehensive implementation and evaluation of customer churn prediction models for e-commerce applications, demonstrating significant achievements across all research phases. The analysis utilized a Kaggle dataset containing 5,630 customer records with 20 features, revealing a moderate class imbalance ratio of 4.94:1 between non-churned (83.2%) and churned (16.8%) customers. Through systematic exploratory data analysis, key behavioural indicators were identified, including strong correlations between churn likelihood and factors such as customer complaints (31.7% churn rate vs. 10.9% for non-complainers), satisfaction scores, and tenure patterns. The feature engineering process successfully transformed 23 original features into 33 engineered features, with 16 critical predictors selected through correlation analysis and statistical significance testing. Four machine learning algorithms were evaluated: Logistic Regression, Random Forest, Random Forest with SMOTE, and XGBoost, with initial results showing XGBoost achieving the highest F1-score of 0.8456 and accuracy of 94.58%. However, following comprehensive hyperparameter tuning using GridSearchCV and 5-fold cross-validation, Random Forest (Tuned) emerged as the optimal solution with F1-score of 0.8556, accuracy of 95.20%, and ROC-AUC of 0.9801, representing a 1.95% improvement over its baseline performance. All success criteria were exceeded, with the champion model achieving 95.5% accuracy (target ≥85%), 86.5% precision (target ≥80%), 84.7% recall (target ≥75%), and 85.6% F1-score (target ≥77%), establishing a robust foundation for practical e-commerce churn prediction applications with significant business value for customer retention strategies.

CHAPTER 5

DISCUSSION AND FUTURE WORK

5.1 Introduction

This chapter discusses the results generated from customer churn prediction in e-commerce industry. This chapter begins with dataset identification, followed by Exploratory Data Analysis (EDA). Then, data preprocessing and feature engineering is comprehensively approached. Logistic Regression, Random Forest, Random Forest attached with SMOTE, and XGBoost are used to evaluate the best model. After that, hyperparameter tuning is done on all 4 models. The purpose is to investigate the changes between 4 models, before and after hyperparameter tuning. According to the results generated during model implementation, it is proven that hyperparameter tuning would improve model performance across multiple metrics. The improved results would be effective in developing robust customer churn prediction models for e-commerce applications. The aim of making a positive contribution to customer retention strategies in e-commerce industry could be guaranteed.

5.2 Summary

Customer churn prediction in e-commerce industry aims to identify customers who are likely to stop using the platform services. This analysis clarifies customer behavioural patterns through data taken from an e-commerce dataset containing 5,630 customer records. This project involves several phases, from data collection to final model evaluation and business impact analysis.

The dataset initially contained class imbalance where 83.2% were non-churned customers and 16.8% were churned customers. The data after collection went through a cleaning stage which means that various preprocessing techniques were carried out,

including handling missing values, removing duplicates, and feature engineering. New derived features were created such as Engagement Score, Customer Value Score, and Recency Score to better capture customer behaviour patterns. Then, categorical encoding and numerical scaling were performed to create uniform data ready for machine learning algorithms.

The processed data was then used to train four different machine learning algorithms: Logistic Regression, Random Forest, Random Forest with SMOTE, XGBoost. Out of all models before hyperparameter tuning, it is shown that XGBoost performs the best, which generates an F1-Score of 0.8456. Out of all models after hyperparameter tuning, it is shown that Random Forest performs the best, which generates an F1-Score of 0.8556. Out of all 4 models, only Random Forest, Random Forest with SMOTE, and XGBoost achieves the minimum benchmark of F1-Score, which is 77%. Unlike Logistic Regression, it does not achieve the minimum benchmark because the F1-Score gained before and after hyperparameter tuning is 0.5952 respectively. This demonstrates that Logistic Regression is not the suitable machine learning algorithm for customer churn prediction.

With a deeper analysis, it can be seen that behavioural factors such as complaints, tenure, and satisfaction scores are more important predictors than demographic information. Feature correlation analysis revealed that tenure shows strong negative correlation (-0.338) with churn probability, reflecting that longer-tenure customers are less likely to churn.

From the success of the project, we can draw the following conclusions. The first conclusion is that data quality and feature engineering are crucial. Proper data preprocessing and derived feature creation provides better results compared to model training without data preprocessing. The second conclusion is that XGBoost and Random Forest performs the best before and after hyperparameter tuning respectively. The third conclusion is that churn prediction system has high potential in improving the ROI by utilizing customer retention strategies.

Overall, the project successfully achieved its goal of developing effective customer churn prediction models in a structured and data-driven manner. The project also demonstrated that machine learning-based churn prediction can act as a powerful tool in supporting customer retention decisions in real-time business operations.

5.3 Future Works

The customer churn prediction analysis revealed critical patterns and insights from the comprehensive examination of 5,630 customer records. The systematic analysis identified key behavioural indicators and established the foundation for predictive model development through statistical examination of churn relationships. The suggested future works would be listed down as shown below.

a) Larger dataset volume

   The current dataset only contains 5630 rows of data, which is insufficient in model training. In the production scenario, 5630 rows of data are considered little, because the dataset in production environment starts from millions. It indicates that more meaningful patterns could be extracted out when the information is sufficiently enough.

b) Application of Deep Learning Models

   This case study applies traditional machine learning models. But traditional machine learning models have its limit in approaching to more complex relationships. To make a breakthrough on this barrier, Deep Learning models would be a better solution, because Deep Learning models are designed to find out the complex relationship from the dataset.

c) Real Time Implementation and Monitoring

   Current solution does not include production deployment. This indicates that the result gained from the training does not exactly reflect the real-time scenario. The differences between local testing and production execution could be significantly different. When the model does not reveal the

mistakes, it is hard to find out the limitations in the current solution. Only the production environment would greatly reveal the flaws in the current solution.

d) Enhanced Feature Engineering with external data sources.
   The solution only uses 1 type of dataset. The features extracted from single source of dataset would be insufficient to make insightful decisions. Additional dataset such as transactional data from e-commerce industry should be used, as the transaction activity represents the real-time action, which proves the record is valid.

The above steps will allow further research to increase the scope of this project, improving the accuracy and practical relevance of customer churn prediction systems. The current project has paved the way for the use of machine learning as an effective customer retention tool in e-commerce; thus, further development will have greater implications in the future for strategic business decision-making and customer relationship management.

REFERENCES

Aljifri, A. (2024). Predicting Customer Churn in a Subscription-Based E-Commerce Platform Using Machine Learning Techniques.

Alshamsi, S. A. (2022). *Customer Churn Prediction in E-Commerce Sector.* Dubai: Rochester Institute of Technology.

Barcelos, E. J. (2020). Negative Online Word-of-Mouth: Consumer's Retaliation in the Digital World. *Journal of Global Marketing*, 1-19.

Chang, M. (2023). *Customer Churn Prediction based on E-Commerce Live Streaming Data.* Rotterdam: Erasmus University Rotterdam.

Daniel Dahlén, W. M. (2023). *Machine Learning-based Prediction of Customer Churn in SaaS.* Lund: Lund University.

Daniyal Asif, M. S. (2025). A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry. *Results in Engineering*, 104629. doi:https://doi.org/10.1016/j.rineng.2025.104629

Das, S. (2025). *Calculating Customer Lifetime Value and Churn using Beta Geometric Negative Binomial and Gamma-Gamma Distribution in a NFT based setting.* arXiv. Retrieved from https://arxiv.org/abs/2501.04719

Deng, E. (2025). Customer Churn Prediction based on Multiple Linear Regression and Random Forest. *5th International Conference on Signal Processing and Machine Learning*, 22-28.

Germain, J. M. (28 July, 2023). *Subscription Commerce Merchants Innovate Amid Rising Churn Rates.* Retrieved from E-commerce Times: https://www.ecommercetimes.com/story/subscription-commerce-merchants-innovate-amid-rising-churn-rates-177730.html

Géron, A. (2023). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* San Francisco: O'Reilly Media, Inc.

Hafiz Ma'ruf, R. (2021). Analysis of Random Forest Algorithm on Customer Churn Prediction to Handle Imbalanced Data. *International Research Journal of Advanced Engineering and Science*, 102-106.

Ikhlass Boukrouh, A. A. (2024). Explainable Machine Learning Models Applied to Predicting Customer Churn for E-Commerce. *IAES International Journal of Artificial Intelligence*, 286-297.

Imani, M. (2024). Evaluating Classification and Sampling Methods for Customer Churn Prediction Under Varying Imbalance Levels.

Kim, S., & Heeseok, L. (2022). Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees. *Procedia Computer Science*, 1332-1339.

Labhsetwar, S. R. (2020). Predictive Analysis of Customer Churn in Telecom Industry using Supervised Learning. *ICTACT Journal On Soft Computing*, 2054-2060.

Nagaraj P, M. V. (2023). E-Commerce Customer Churn Prediction Scheme Based on Customer Behaviour Using Machine Learning. *International Conference on Computer Communication and Informatics (ICCCI)*, (pp. 1-6). Coimbatore, India.

Sana Fatima, A. H. (2023). XGBoost and Random Forest Algorithms: An In-Depth Analysis. *Pakistan Journal of Scientific Research, PJOSR*, 26-31.

Swetha P, D. R. (2020). Customer Churn Prediction and Upselling using MRF (Modified Random Forest) Technique. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 475-482.

Taskin, N. (2023). *Customer Churn Prediction Model in Telecommunication Sector Using Machine Learning Technique.* Uppsala University.

Tianpei Xu, Y. M. (2021). Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping. *Applied Sciences*, 1-12.

Tran, L. D. (2023). *Enhancing Telecom Churn Prediction: Adaboost With Oversampling and Recursive Feature Elimination Approach.* San Luis Obispo: California Polytechnic State University.

Velu, A. (2021). Customer Churn Management Using Predictive Modelling - A Machine Learning Approach. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 1410-1421.

Xinyu Miao, H. W. (2022). Customer Churn Prediction on Credit Card Services using Random Forest Model. *2022 7th International COnference on Financial Innovation and Economic Development (ICFIED 2022)*, 649-656.

Yadav, R. (2024). *Machine Learning Insight into E-Commerce Churn: Prediction and Preventing Customer Loss.* Dublin: Dublin Business School.

Yashkumar Burnwal, D. R. (2023). A Comprehensive Survey on Prediction Models and the Impact of XGBoost. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 1552-1556.