E-Commerce Customer Retention Analysis

MSc Research Project
Data Analytics

Mohan Sugumaran
Student ID: x22183779

School of Computing
National College of Ireland

Supervisor: Arjun Chikkankod

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Mohan Sugumaran		
Student ID:	X22183779		
Programme:	Data Analytics	**Year:**	2024
Module:	Research Project		
Supervisor:	Arjun Chikkankod		
Submission Due Date:	12/08/2024		
Project Title:	E-commerce Customer Retention Analysis		
Word Count:	7050 **Page Count** 22		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Mohan Sugumaran

Date: 12/08/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table of Contents

E-Commerce Customer Retention Analysis

Mohan Sugumaran
Student ID: x22183779

Abstract

In the highly competitive e-commerce industry, businesses encounter substantial difficulties in maintaining their current clients due to the ease of transitioning between platforms and the multitude of options available. Consequently, corporations in this fiercely competitive market struggle to maintain customer loyalty. The customer attrition poses a significant risk to the financial stability and long-term viability of an organisation. This project aims to provide a comprehensive machine learning system that can anticipate client churn and enable proactive customer retention measures. A thorough data analysis focusses on Tenure, City_Tier, Payment methods, Gender, and Service_Score. Handling missing data, encoding categorical variables, and normalising numerical characteristics are data preparation tasks. SMOTE is used to equalise the dataset. Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Bagging, Support Vector Machine, and Naive Bayes are evaluated. GridSearchCV adjusts hyperparameters to improve model performance. Models are evaluated on accuracy, precision, recall, and F1 score, focusing on recall. Gradient Boosting, the best model with 0.95 accuracy and 0.96 recall, is used with Streamlit to provide real-time churn estimates and actionable insights.

Keywords: Customer Churn, E-commerce Retention, Data Analytics, Machine Learning Prediction, SMOTE, Gradient Boosting, Predictive Modelling

1 Introduction

In the dynamic and highly competitive realm of online e-commerce, the task of maintaining clients has grown ever more difficult. This issue is heavily influenced by the ease of platform switching for customers and the abundance of accessible alternatives. E-commerce consumer attrition or customer churn can be defined as the regular and consistent loss of consumers over a given period of time and is seen as one of the major threat factors that e-commerce business organisations face in the long run that can negatively impact their financial viability. The high rate of customer turnover leads to the negative impacts such as loss of its income, increased advertising costs, and the firm’s inability to grow. Comprehending and reducing client turnover has become a vital priority for e-commerce businesses seeking to maintain their market status and profitability. Obtaining new consumers incurs more costs compared to keeping current ones, while loyal customers create more income by making repeated purchases and promoting the brand. Churn is influenced by factors such as customer service, insufficient client involvement, and more attractive offerings from rival companies. Conventional customer retention strategies often prove inadequate since they fail to consider the unique behaviours of individual customers. Utilising data analytics and machine learning enables organisations to get valuable information about customer behaviours, detect trends of customer attrition, and forecast clients who are likely to discontinue their engagement. The objective of this project is to create an advanced churn prediction model using machine learning methods, which will allow for timely and personalized interventions.

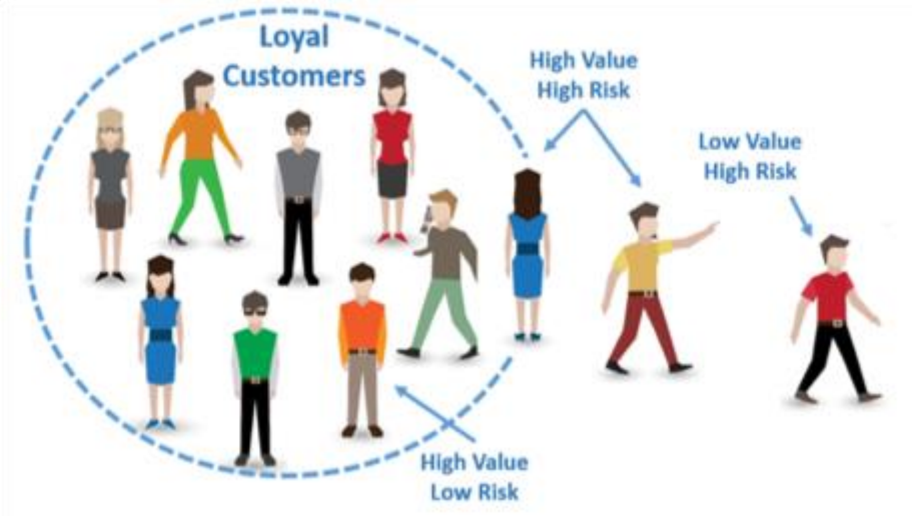


Figure 1: Customer Segmentation Model

The research includes the gathering, analysis, organisation, creation, and assessment of data. Multiple algorithms will be evaluated, and the most effective model will be implemented on a user-friendly platform such as Streamlit to provide real-time forecasts and practical insights. This technique aims to improve client retention, decrease churn rates, and facilitate the long-term expansion and viability of e-commerce enterprises.

1.1 Problem Statement

Customer satisfaction and its management is also turning out to be one of the challenging objectives in the e-commerce business environment where the competition is always high. E-commerce consumer attrition or customer churn can be defined as the regular and consistent loss of consumers over a given period and is seen as one of the major threat factors that e-commerce business organisations face in the long run that can negatively impact their financial viability. The high rate of customer turnover leads to negative impacts such as loss of its income, increased advertising costs, and the firm's inability to grow. This problem is made worse by the fact that one account can represent multiple sets of customers. The cumulative effect of account churn makes it even more important to efficiently identify and reduce churn to prevent more losses. To protect market position, decrease revenue loss, and preserve the long-term profitability of the firm, it is vital to address the issue of customer turnover.

1.2 Problem Solution

An extensive churn prediction model will be created based on this project where the focus will be on using state-of-art machine learning tools to work on the issue of customer churn. In other words, components that are most crucial in case of customer churn will have to be accessed by the model for predicting which accounts are most likely to churn. These characteristics include tenure, service interactions, payment methods, and customer demographics. The model will be trained on a balanced dataset to improve its accuracy of prediction. The given model will be fine-tuned using a balanced data set to increase its rate of prediction. This shall be done through data pre-processing approaches and by applying the Synthetic Minority Oversampling Technique (SMOTE) to cater for class imbalance issue.

Algorithms that fit within the machine learning bracket will be considered as there is more than one algorithm that could be used, and the most effective one must be determined. Some of these algorithms for example are Logistic Regression, Decision Tree, Random Forest, Gradient Boosting and so on. Out of all the created models, the best one will be used to present churn forecasts in real-time to manage churn rates and develop strategies for customer loyalty.

1.3 Research Questions

i. In the e-commerce sector, what are the most important variables that cause customers to leave?
ii. What are the best practices for optimising machine learning models to forecast e-commerce client churn?
iii. How can a churn prediction model play a role in influencing client retention and profitability for businesses?

1.4 Ethical Consideration

There are several ethical issues with the suggested research on e-commerce customer departure prediction. This project's dataset was obtained via Kaggle:

Privacy of Data: The dataset includes consumer purchase patterns and interaction history, but it does not contain any personal information. Ensuring secure storage and data confidentiality remains important throughout the analysis process.

Machine Learning Biases: Based on the training set of data, models may display biases that result in inaccurate predictions. To guarantee impartiality and reasonable churn forecasts, assessment processes and mitigating techniques are required.

Anchors and Counterfactuals: Precise churn forecasting is crucial. While false positives (erroneous churn projections) cost resources, false negatives (missed churners) result in wasted retention chances. It's crucial to weigh these dangers.

Explainability and Transparency: Particularly intricate models like Gradient Boosting, machine learning models might be ambiguous. To increase confidence and validation, it is crucial to improve explainability and transparency, with a particular emphasis on enhancing model interpretability.

1.5 Structure of the Report

The main report consists of three sections. In the current paper, you will find a brief background of the study in relation to the objectives and research questions that lead to anticipating customer churn in e-commerce organisations. This was succeeded by a review of the literature which defines the gaps in the churn prediction methods and elaborate on new complex data analysis and more efficient and precise machine learning technologies for effectual churn prediction. Moreover, the final important section, the research design and methodology section, describes the process of constructing the churn prediction model, the data exploration, data preparation, model building, and model assessment. In this section, data privacy, elimination of bias, and explicability and transparency of the machine learning models are also discussed ethically.

2 Related Work

2.1 Overview

Customer attrition, or consumers' desertion over time, is a critical issue with regard to retaining customers and sales for any organisation, especially e-businesses. Precisely forecasting customer attrition and executing efficient methods to retain customers are essential for keeping a steady client base. This literature review examines several approaches and models employed in churn prediction, emphasizing important discoveries and contributions from relevant sources.

2.2 Models for Predicting Customer Churn

Initial investigations on churn prediction predominantly employed statistical methodologies. (Neslin et al., 2006) provided evidence for the efficacy of logistic regression in forecasting customer turnover through the examination of consumer behaviour patterns and transaction histories. In a similar vein, (Kim & Yoon, 2004) emphasized the utilisation of decision trees to categorize clients according to their likelihood of churning.

Machine learning developments have led to the exploration of increasingly intricate models. Random forests and gradient boosting machines, such as XGBoost, have become popular because of their exceptional predicting accuracy. In their study, (Lemmens & Croux, 2006) conducted a comparison of several machine learning approaches and discovered that ensemble methods, such as random forests, had superior performance when compared to standard statistical models. (Chen & Guestrin, 2016) emphasized the effectiveness and capability of XGBoost in managing extensive datasets and intricate prediction problems.

2.3 Data mining techniques

Data mining plays a crucial role in the identification of churn tendencies. (Hung et al., 2006) employed clustering methodologies to categorise consumers with similar characteristics and detect individuals who are very likely to churn. In addition, (Shaaban et al., n.d.) utilised association rule mining to reveal concealed connections between customer characteristics and churn behaviour.

2.4 Feature Engineering

Efficient feature engineering is crucial for improving the performance of the model. (Buckinx & Van Den Poel, 2005) highlighted the significance of extracting significant characteristics from unprocessed data, such as recency, frequency, and monetary value (RFM) measures. (Verbeke et al., 2011a) emphasized the application of social network analysis to identify relationship characteristics that impact customer attrition.

2.5 Managing Data Imbalance

Customer churn statistics frequently exhibit an imbalance, characterized by a smaller number of churn instances in comparison to non-churn cases. This disparity has the potential to influence the predictions made by the model. In 2002, Chawla et al proposed the Synthetic Minority Over-sampling Technique or commonly known as SMOTE. Through SMOTE, synthetic samples for the minority class are created. Japkowicz and Stephen (2002) (Chawla et al., 2002) conducted a thorough examination of methods for managing unbalanced data in churn prediction.

2.6 Evaluation Criteria

Proper assessment of churn prediction models necessitates suitable metrics. The evaluation of most models is often conducted using metrics like as accuracy, precision, recall, and F1 score. Nevertheless, Neslin et al. (2006) asserted that recall has particular significance in churn prediction, since it quantifies the model's capacity to accurately identify individuals who have already churned. In their study, (Burez & Van den Poel, 2009) conducted a comparison of several assessment metrics and emphasised the need of taking the business environment into account when choosing evaluation criteria.

2.7 Recent advancements in Churn Prediction

Recent research has investigated the use of advanced deep learning methods for predicting customer attrition. (Hinton et al., 2012) showed the capability of deep neural networks to effectively capture intricate patterns in consumer data. In addition, employed CNNs and RNNs for time-series consumer data analysis and achieved outstanding prediction results.

Another highly effective approach developed for consideration is transfer learning. This has been evidenced by (Weiss et al., 2016) in their study where the author showed that improving one model using the other provided improved results based on the transfer of information when there is limited data with labelled information.

2.8 Ethical Considerations

The issue of ethical implications in churn prediction has garnered significant attention. (Roy, 2017) examined the potential dangers of algorithmic bias and underscored the need of ensuring fairness in machine learning models. In addition, (Dwork & Roth, 2014) developed the notion of differential privacy as a means of safeguarding consumer data during the process of analysis.

2.9 Case studies and Applications

Multiple case studies exemplify the pragmatic uses of churn prediction models (Verbeke et al., 2011b) conducted a case study on telecom churn prediction, illustrating the advantages of using different data sources. In research done by (Coussement & Van den Poel, 2008), the authors examined the prediction of newspaper subscription cancellations and emphasised the efficacy of logistic regression and random forests.

2.10 Hybrid Models

Aggregating various models can improve the accuracy of predictions. Hybrid models that combine statistical methodologies with machine learning techniques have demonstrated potential (Larivière & Van Den Poel, 2005) integrated logistic regression with decision trees, yielding superior outcomes compared to standalone models. (Tsai & Lu, 2009) suggested a combination of neural networks and genetic algorithms as a hybrid technique for predicting customer turnover in the banking industry.

2.11 Ensemble Learning

Ensemble learning approaches, which amalgamate predictions from many models, have exhibited enhanced accuracy. In 2001, Breiman proposed the concepts of bagging and boosting, which serve as the foundation for widely used algorithms such as random forests and gradient boosting machines. (Caruana et al., 2004) emphasised the advantages of ensemble learning in mitigating overfitting and enhancing generalization.

2.12 Model Interpretability

Although complicated models frequently have superior accuracy, the task of achieving interpretability remains a problem. In 2016 Ribeiro, Singh and Guestrin came up with an approach known as Local Interpretable Model-agnostic Explanations (LIME) for the purpose of explaining black-box models' predictions. Moreover, (Lundberg et al., 2017) introduced SHapley Additive Explanations (SHAP) for the provision of an understanding concerning the importance of some features.

2.13 Deployment and User Engagement

Streamlit integration for e-commerce applications is a freely available application framework that enables the rapid development of data applications. Integrating machine learning models into Streamlit apps for customer churn prediction can offer dynamic and user-friendly interfaces. (Aendikov & Azayeva, 2024) showcased the integration of Geographic Information System (GIS) with machine learning analytics in a Streamlit application, emphasizing the capability for instantaneous data processing and visualisation.

3 Research Methodology

3.1 Business Understanding

The initial part of business analysis is determining the primary factors that contribute to customer attrition inside the e-commerce platform. This stage guarantees that the approach effectively tackles the crucial obstacle of client retention. By precisely delineating the issue of customer attrition and comprehending the impact of churn on revenue, we integrate the project's objectives with the broader company goals, with a specific emphasis on diminishing churn rates and enhancing customer retention techniques.

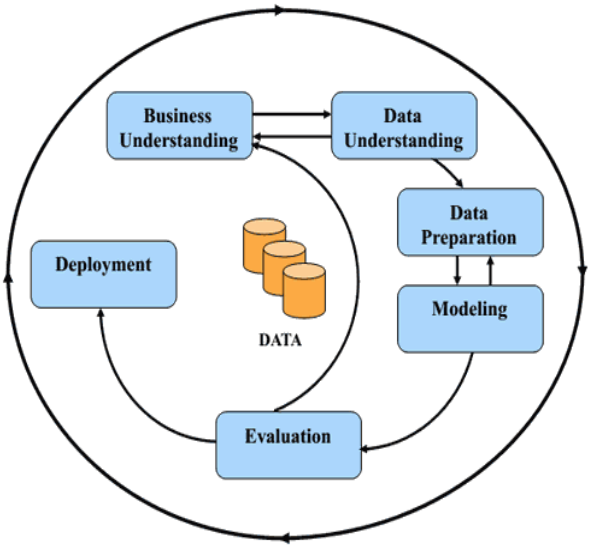


Figure 2: CRISP-DM System Flow

3.2 Data Analysis

During the data comprehension phase, the dataset is examined to reveal significant patterns pertaining to consumer behaviour and churn. This entails analysing characteristics such as the frequency of purchases, the duration of client relationships, and the history of interactions. By analysing these data points, we obtain valuable information about the aspects that strongly indicate churn. This ensures that the following phases in the process are grounded in a thorough comprehension of the data's attributes and its significance to the situation at hand.

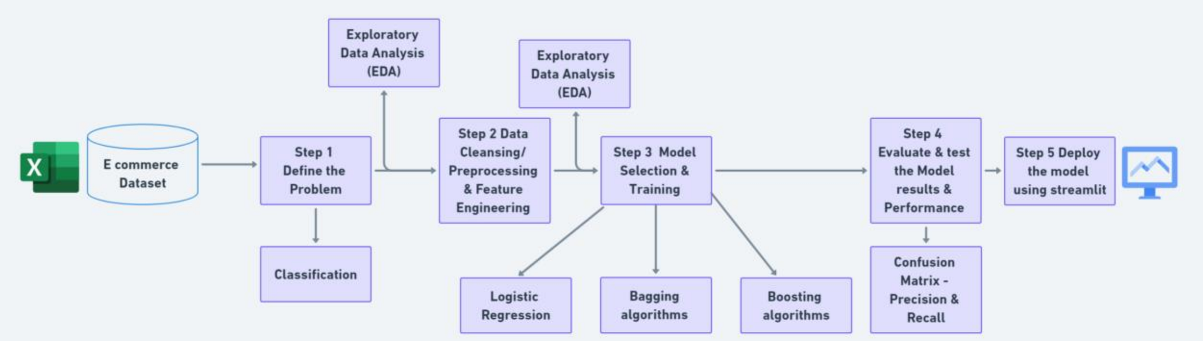

E commerce Dataset
Step 1 Define the Problem
Classification
Exploratory Data Analysis (EDA)
Step 2 Data Cleansing/ Preprocessing & Feature Engineering
Exploratory Data Analysis (EDA)
Step 3 Model Selection & Training
Logistic Regression
Bagging algorithms
Boosting algorithms
Step 4 Evaluate & test the Model results & Performance
Confusion Matrix - Precision & Recall
Step 5 Deploy the model using streamlit

Figure 3: E-commerce customer retention architectural diagram

3.3 Data Preparation

Data preparation encompasses the process of purifying and modifying the unprocessed data to guarantee its suitability for analysis. This encompasses the tasks of managing null values, converting category information into numerical representations, and generating additional characteristics that might potentially improve the accuracy of churn prediction. Thorough data preparation is crucial for enhancing the precision and dependability of the model, guaranteeing its capacity to successfully acquire knowledge from the data.

3.4 Modelling

During the modelling step, the prepared data is subjected to a variety of machine learning methods, including Random Forest, Gradient Boosting, and Logistic Regression. This stage involves choosing the best suitable model according to the company needs, training it using past data, and adjusting its parameters to enhance performance. The objective is to construct a resilient model capable of precisely detecting consumers who are prone to churn, hence facilitating proactive retention initiatives.

3.5 Evaluation

During the assessment phase, models are appraised based on important parameters like as accuracy, recall, and precision. Given the significance of recall in this specific situation, the evaluation primarily centres on the model's ability to accurately detect real churners. Through the process of comparing several models, we guarantee that the chosen model not only aligns with the business objectives but also demonstrates strong performance on unfamiliar data, hence establishing its reliability in accurately predicting customer turnover in real-life situations. Below is the confusion matrix for the best model random Forest by applying the SMOTE technique.

3.6 Deployment

During the deployment phase, the selected model is integrated into the operational processes of the e-commerce platform. This allows instantaneous or regular forecasts, allowing the company to promptly implement measures to retain consumers who are at danger of leaving. In addition, the deployment process involves establishing monitoring systems to monitor the performance of the model, assuring its ability to deliver precise forecasts even if the business and consumer behaviour undergo changes.

4 Design Specification

4.1 Architectures Used

In this step, we will be explaining about the algorithms we will be using in this research. Additionally, we will get knowledge of the interface that will be utilised for actual web deployment.

4.2 Logistic Regression

Because of its ease of use and performance in binary classification problems, Logistic Regression was utilised in this research. By fitting a logistic function to the data, this model calculates the chance of a client churning. To forecast the log chances of the target variable in this instance, churn it employs a linear combination of the input characteristics (Hosmer et al., 2013) Logistic Regression's interpretability is a major plus as it clarifies how each factor affects the churn probability. With this model as a starting point, we can better understand what causes customers to leave.

4.3 Decision Trees

The use of Decision Trees was based on its simple and clear methodology for modelling. Their functioning involves the iterative division of the data into smaller groups, using the most important characteristic at each step, resulting in a hierarchical arrangement resembling a tree (Quinlan, 1986) Every branch in the model reflects a specific decision rule, which ultimately leads to a final prediction upon reaching the leaf nodes. Decision Trees are extremely transparent, facilitating the visualisation and comprehension of the decision-making process. They successfully handle both numerical and categorical data, and their capacity to capture non-linear correlations makes them appropriate for discovering intricate patterns in consumer behaviour that lead to churn.

4.4 Random forest

The ensemble learning approach known as Random Forests was utilised to enhance both the accuracy and robustness of predictions (Breiman, 2001) This method use an ensemble of Decision Trees, constructed using random subsets of the data and characteristics, to provide a model that is more robust and can be applied to a wider range of situations. The ultimate forecast is derived by taking the average of the forecasts made by each individual tree. Random Forests alleviate the problem of overfitting, which is frequently observed in individual Decision Trees, and improve the model's performance by harnessing the collective intelligence. This model is highly efficient in managing extensive datasets with a huge number of variables, making it ideal for our churn prediction task.

4.5 Gradient Boosting

Gradient Boosting Machines (GBM) were used due of their exceptional prediction capabilities and robustness (Friedman, 2001) The GBM algorithm constructs models in a sequential manner, where each subsequent model is designed to rectify the mistakes caused by the preceding models. This iterative method enhances the overall performance of the model by minimising a designated loss function. GBM is a highly efficient method for managing complicated datasets and identifying nuanced patterns, which is why it is our favourite choice for the churn prediction challenge. Although GBM's computing requirements are higher and it necessitates meticulous hyperparameter adjustment, its exceptional accuracy and precision make it indispensable for customer churn prediction.

4.6 AdaBoost

AdaBoost, also known as Adaptive Boosting, was employed to improve the effectiveness of feeble classifiers by amalgamating them into a potent classifier (Freund & Schapire, 1997) The process involves training weak learners, usually Decision Trees, in a sequential manner and modifying their weights according on the faults of the prior models. This iterative procedure prioritises the most challenging examples to forecast, progressively enhancing the overall accuracy of the model. AdaBoost is a very successful algorithm for churn prediction due to its ability to effectively reduce both bias and volatility, making it a stable and reliable choice. The capacity to amalgamate several feeble models into a robust one facilitates the comprehension of intricate client behaviour.

4.7 Bagging

Bagging, also known as Bootstrap Aggregating, was employed to minimise variation and mitigate the risk of overfitting (Breiman, 1996) This ensemble technique generates numerous subsets of the original dataset using bootstrapping, which involves randomly sampling with replacement, then trains separate models on each of these subsets. The final forecast is obtained by averaging all the models' predictions. Bagging is useful when there is a necessity of enhancing the models for which there is a high rate of overfitting including the Decision Trees. It achieves this by enhancing their stability and accuracy. Bagging is employed in our churn prediction assignment to enhance generalisation to unfamiliar data, hence assuring more dependable predictions. The below diagram describes difference between the bagging and boosting.

4.8 Support Vector Machines

This algorithm was chosen due to their efficacy in handling high-dimensional areas and their resistance to overfitting (Cortes & Vapnik, 1995) The SVM algorithm operates by identifying the most favourable hyperplane that effectively divides the classes (churn vs. non-churn) with the widest possible margin. Kernel tactics are very valuable for datasets in which the classes cannot be separated in a linear manner. These tricks include transforming the data into higher dimensions. The SVM algorithm is optimally suitable for both linear and non-linear separable data making it capable for the prediction of customer turnover. The capacity to optimise the margin between classes contributes to reaching a high level of predicted accuracy.

4.9 Naïve Bayes

This algorithm was picked because of its straightforwardness and high computing efficiency (Murphy, 2012) This classifier utilises Bayes' theorem and assumes that the characteristics are independent of each other. Despite this robust presumption, Naive Bayes frequently exhibits impressive performance in real-world scenarios, especially when dealing with datasets that have many dimensions. The algorithm computes the likelihood of each class (churn vs. non-churn) by considering the provided characteristics and assigns the class with the highest likelihood. Naive Bayes is particularly adept at managing categorical data and serves as a valuable reference model in our churn prediction research, offering rapid insights and a benchmark for more intricate models.

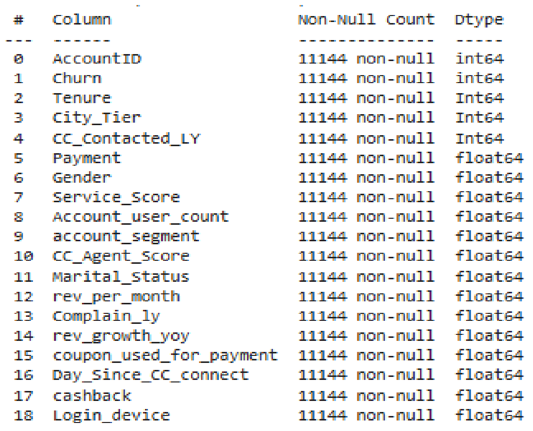
5 Implementation

5.1 Understanding Objectives

Identifying and comprehending the key goals is the first stage of every endeavor. Predicting consumer churn in the e-commerce industry and creating efficient retention measures to reduce it is the primary objective. Our specific goals are to determine what causes customers to leave, develop reliable prediction models, and come up with tailored retention methods to keep customers around and increase loyalty while decreasing attrition. With well-defined goals in place, the project stays on track to deliver measurable results that support the company's overall objectives.

5.2 Data Collection and Preprocessing

This project's data was gathered from Kaggle. To make sure the dataset is reliable and usable, preprocessing measures are conducted once data is obtained. Part of this process is cleaning the data so that it is free of duplicate or incorrect entries. The analysis considers the relevance and effect of missing data while deciding whether to impute or delete them. Furthermore, churn prediction is assisted by engineering relevant characteristics. consumer lifetime value, average transaction value, purchase frequency, and other engagement data that might shed light on consumer behaviour could be included in these features.



```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   AccountID                 11144 non-null  int64
 1   Churn                     11144 non-null  int64
 2   Tenure                    11144 non-null  Int64
 3   City_Tier                 11144 non-null  Int64
 4   CC_Contacted_LY           11144 non-null  Int64
 5   Payment                   11144 non-null  float64
 6   Gender                    11144 non-null  float64
 7   Service_Score             11144 non-null  float64
 8   Account_user_count        11144 non-null  float64
 9   account_segment           11144 non-null  float64
 10  CC_Agent_Score            11144 non-null  float64
 11  Marital_Status            11144 non-null  float64
 12  rev_per_month             11144 non-null  float64
 13  Complain_ly               11144 non-null  float64
 14  rev_growth_yoy            11144 non-null  float64
 15  coupon_used_for_payment   11144 non-null  float64
 16  Day_Since_CC_connect      11144 non-null  float64
 17  cashback                  11144 non-null  float64
 18  Login_device              11144 non-null  float64
```

Figure 4: Description of Data

5.3 Exploratory Data Analysis

Since EDA is done after data preprocessing, a basic understanding of the data is gained to get a handle on the data. This amounts to performing simple statistical operations aimed at the description of the distribution and depicting the main characteristics of the data. To find out how various attributes relate to the dependent variable, in this case customer churn, we do a correlation analysis. To get important insights from data, data visualisation techniques like histograms, box plots, and scatter plots are employed to spot trends and outliers.

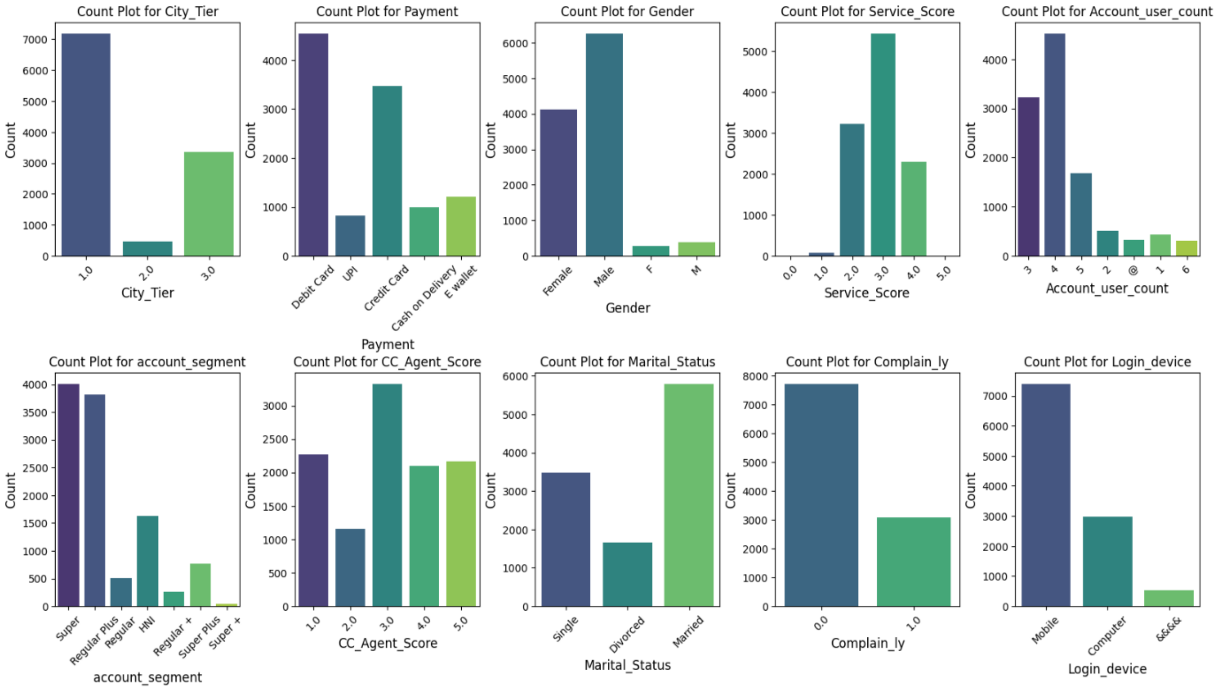


Figure 5: Univariate Analysis-Categorical Feature Distribution

The Figure 14 plots offer valuable insights into the distribution of several categorical parameters associated with customer retention. The plot depicting the 'City_Tier' reveals that a substantial number of consumers originate from Tier 1 cities, although a notable proportion also come from Tier 2 cities. The 'Payment' method plot indicates a predilection for specific payment methods, with Credit Card and Debit Card being the predominant choices. The 'Gender' plot exhibits an even distribution of clients between males and females. The plot of the 'Service_Score' indicates that most customers have a service score in the middle range, while there are fewer customers with scores at the extreme ends. The term 'Account_user_count' indicates that most accounts have only one user, while a smaller number of accounts have multiple users.

The plot of the 'account_segment' reveals a greater level of client concentration in specific segments. The metric 'CC_Agent_Score' exhibits a broad range of scores assigned by customer service agents. The 'Marital_Status' plot indicates that married consumers are the predominant group. The term 'complain_ly' indicates that a significant proportion of consumers have registered complaints in the previous year. The 'Login_device' plot displays the prevailing use of specific devices for the purpose of logging in. These data are valuable for comprehending client demographics and behaviours, which are essential for creating efficient churn prediction models.

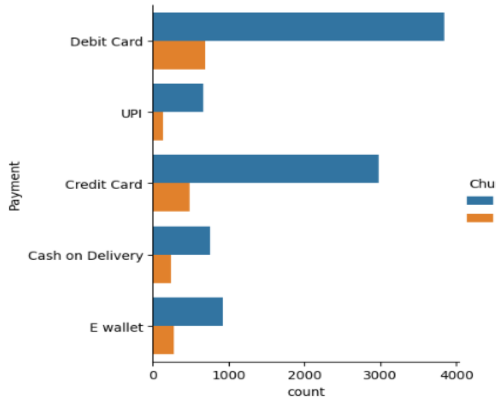


Figure 6: Payment Method

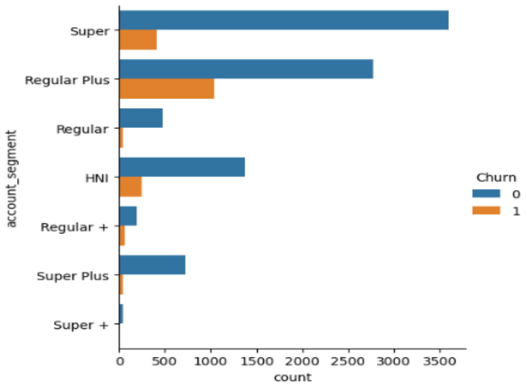


Figure 7:Account Segment

According to the Figure 15, most non-churning clients use Debit Cards and Credit Cards, with Debit Cards being the most common. E-Wallets and UPI have the fewest users, while Cash on Delivery is modest. More churn occurs with Cash on Delivery and UPI than other payment options. This visualisation reveals which payment methods are linked to client attrition, aiding retention initiatives.

The Figure 16 reveals that the 'Super' and 'Regular Plus' groups have the most consumers and some churn. The 'HNI' category has many users and a modest turnover rate. The 'Regular', 'Regular +', 'Super Plus', and 'Super +' sectors have fewer clients and low turnover.

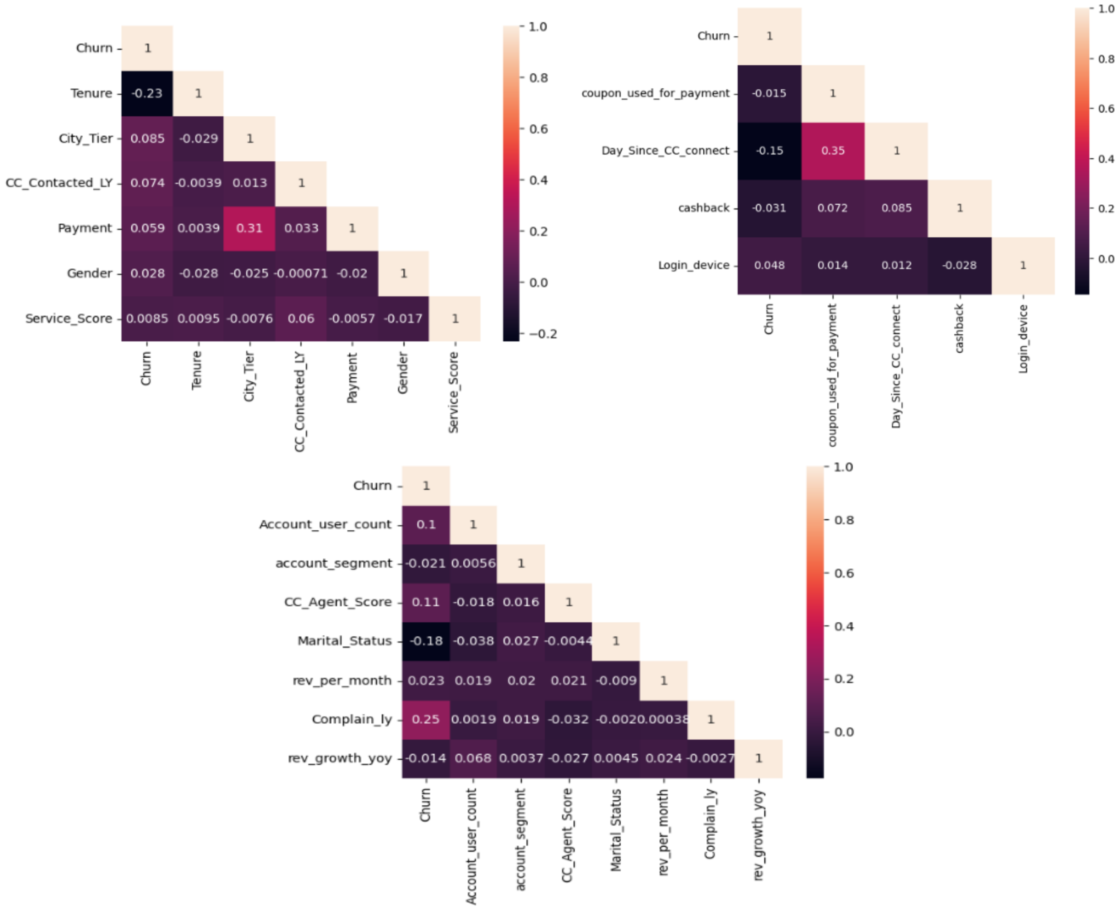


Figure 8: Correlation Heatmaps of Features

The correlation matrix provides valuable insights on customer retension. The variable 'tenure' has a robust negative connection with churn, suggesting that an extended duration of customer tenure decreases the probability of turnover. The variable 'Complain_ly' has a positive connection with churn, indicating that consumers who file complaints are more prone to churning. The variable 'Day_Since_CC_connect' has a significant negative connection with churn, suggesting that recent engagements with customer care can reduce churn rates. Furthermore, the variables 'Marital_Status' and 'Account_user_count' exhibit significant negative associations with churn, suggesting that customers who are married and have a higher number of users on their account are less prone to churn. On the other hand, there is a low association between 'Login_device' and 'coupon_used_for_payment' and churn, suggesting that these characteristics have a limited effect on customer retention. These observations emphasise the significance of customer involvement and satisfaction in decreasing retention rates.

5.4 Splitting Dataset

The data has been split into 70% training set and 30% to the testing set. The training data was utilised to cultivate the machine learning models, enabling them to acquire knowledge and discern patterns associated with client turnover.

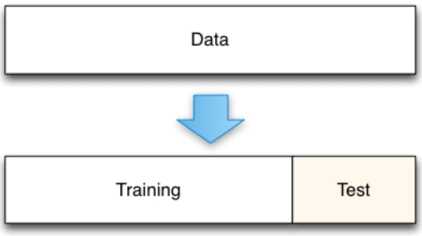
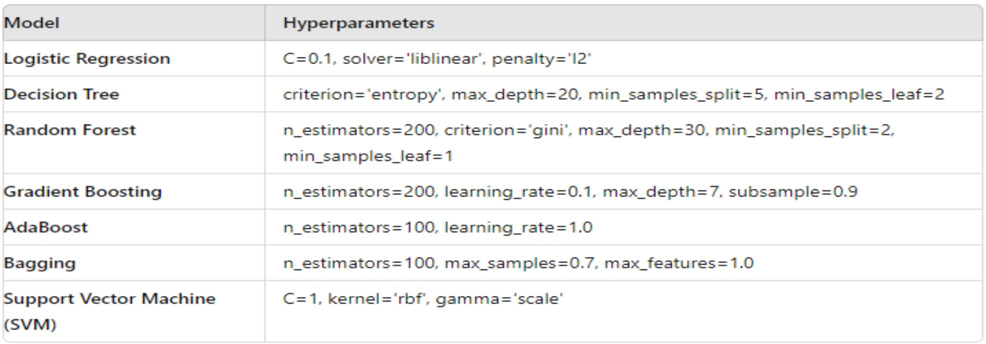


Figure 9: Splitting Data into Training and Testing Sets

5.5 Models and Hyperparameter Tuning

Machine learning models were optimised via hyperparameter tuning across various algorithms. Optimisation is necessary to create the best forecasts, especially in critical situations like customer churn prediction. Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Bagging, and SVM are studied. Grid Search tested several parameter combinations to find the best hyperparameters for each model. The models improved in accuracy and recall, which are crucial for churn prediction, after tuning. After optimisation, the Random Forest model has balanced accuracy and recall, making it a good contender for practical applications. After adjustment, the Gradient Boosting model improved recall, suggesting it may detect churners. The following table lists each model's tuned hyperparameters:



Model	Hyperparameters
Logistic Regression	C=0.1, solver='liblinear', penalty='l2'
Decision Tree	criterion='entropy', max_depth=20, min_samples_split=5, min_samples_leaf=2
Random Forest	n_estimators=200, criterion='gini', max_depth=30, min_samples_split=2, min_samples_leaf=1
Gradient Boosting	n_estimators=200, learning_rate=0.1, max_depth=7, subsample=0.9
AdaBoost	n_estimators=100, learning_rate=1.0
Bagging	n_estimators=100, max_samples=0.7, max_features=1.0
Support Vector Machine (SVM)	C=1, kernel='rbf', gamma='scale'

Figure 10: Optimised Hyperparameters for Machine Learning Models

5.6 Evaluation Process

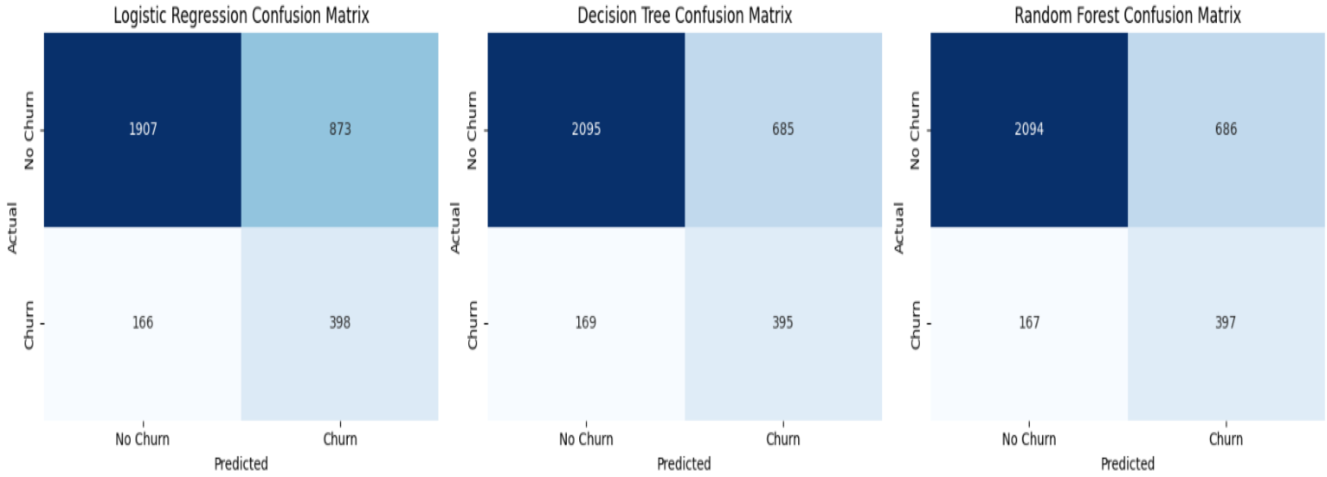
To make sure that the chosen algorithms reliably predict client retention and adapt to new data, evaluating them is an essential step. To assess the model, the predictions of a given model must be compared with the results of subsequent analysis on the test set. The performance is evaluated with several metrics: It also includes metrics like Turn Around Time, Sensitivity, Specificity, Positive Predictive Value among others; F1 Score and AUC-ROC. Regarding the second aspect, the results as true positives, true negatives, false positives, and false negatives are depicted by a confusion matrix. To further guarantee the models' resilience across various data subsets, cross-validation techniques can be employed.

5.7 Model Development

The next phase is to create churn prediction machine learning models. It was possible to split the given data set into the training and testing data set for proper evaluation of the performance of the model developed. Some of the machine learning models that are used in the process include Logistic Regression, Decision Trees, Random Forests, Gradient Boosting Machines and Neural Networks based to how effectively the task is done. For fine tuning the parameters of the model, something like Grid Search or Random Search can be used. Next, the traversal and training of the models take place through the training data set, and the obtained models' validity is through the test set. They are defined by measures that are as follows; Accuracy, Precision, Recall, F1 Score, and AUC-ROC.

5.8 Evaluation Process- Confusion Matrices

To make sure that the chosen algorithms reliably predict client retention and adapt to new data, evaluating them is an essential step. To assess the model, the predictions of a given model must be compared with the results of subsequent analysis on the test set. The performance is evaluated with several metrics: It also includes metrics like Turn Around Time, Sensitivity, Specificity, Positive Predictive Value among others; F1 Score and AUC-ROC. Regarding the second aspect, the results as true positives, true negatives, false positives, and false negatives are depicted by a confusion matrix. To further guarantee the models' resilience across various data subsets, cross-validation techniques can be employed.



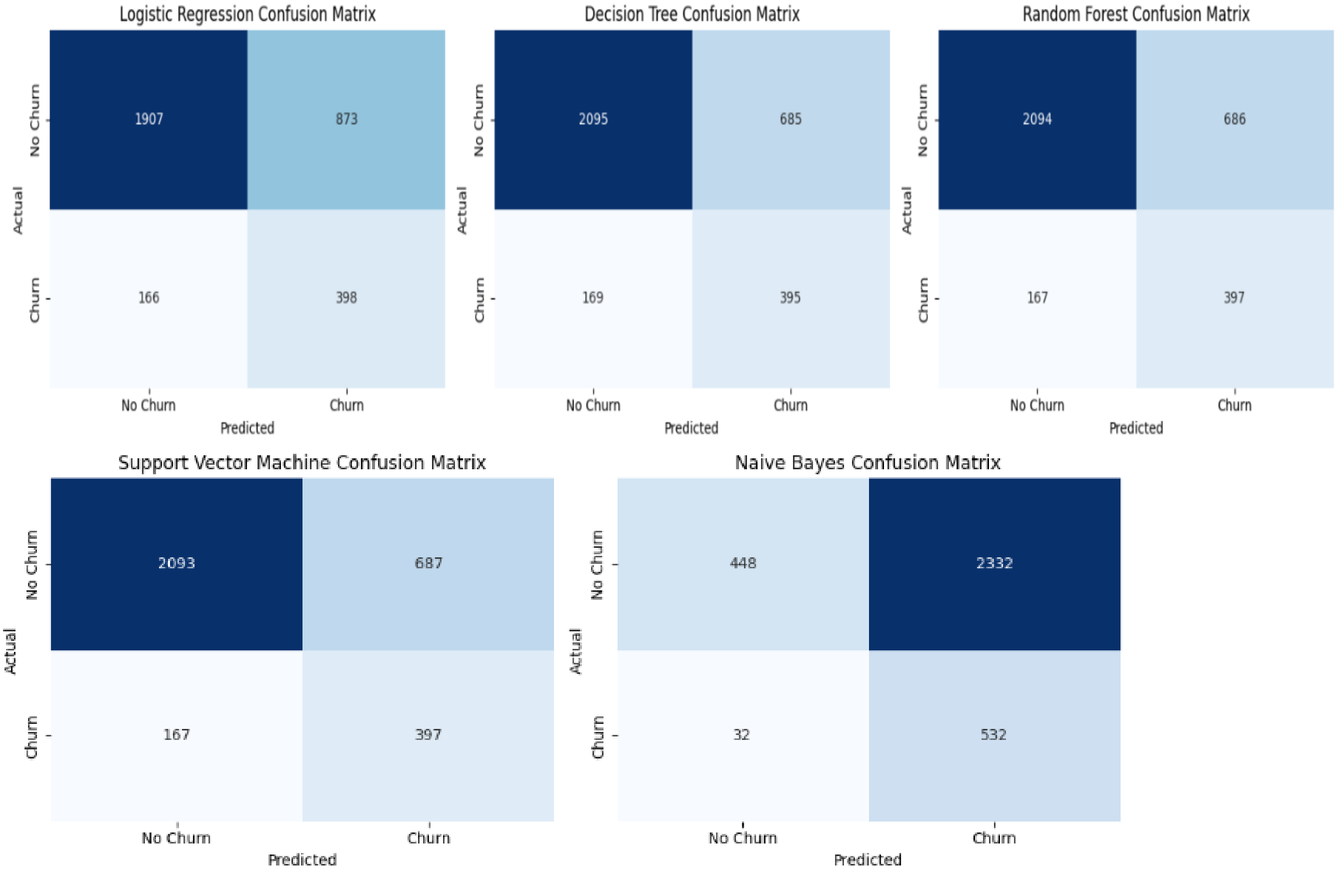
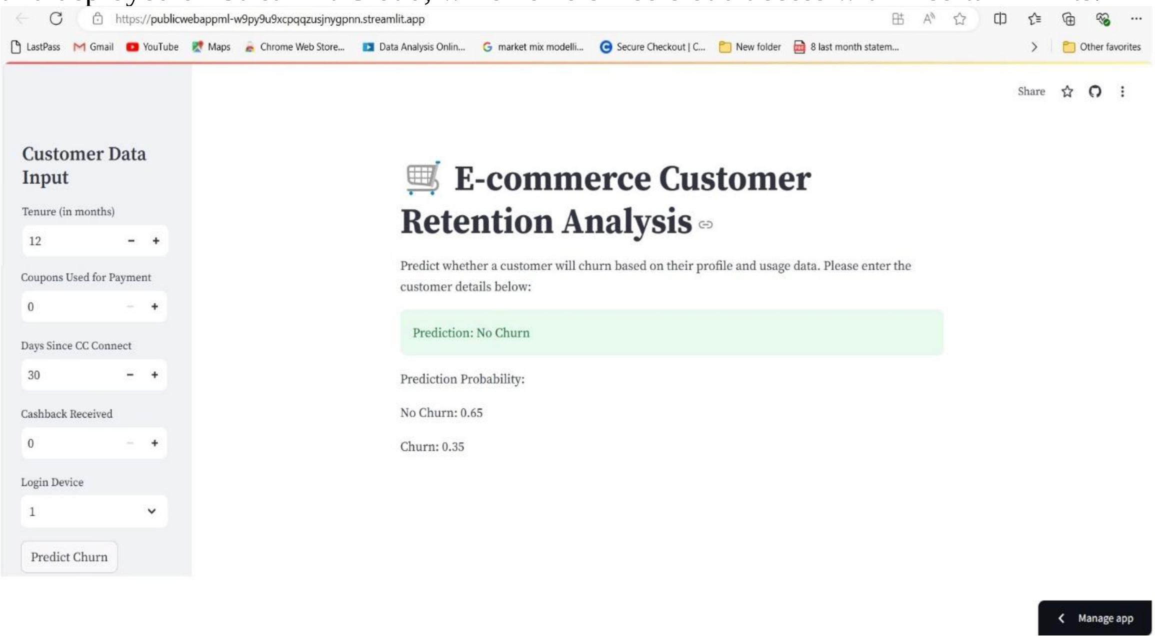


Figure 11: Confusion Matrices for All Models

5.9 Deployment Phase

The project incorporates Streamlit into the churn prediction algorithms to offer a user-friendly and interactive interface. This program makes it possible to process data in real-time and alter forecasts and strategies on the go. The UI is a great tool for businesses since users can use it to test different scenarios and retention techniques. The model has been designed and deployed on Streamlit Cloud, which offers free cloud access within certain limits.



Customer Data Input

Tenure (in months)

12

Coupons Used for Payment

0

Days Since CC Connect

30

Cashback Received

0

Login Device

1

Predict Churn

E-commerce Customer Retention Analysis

Predict whether a customer will churn based on their profile and usage data. Please enter the customer details below:

Prediction: No Churn

Prediction Probability:

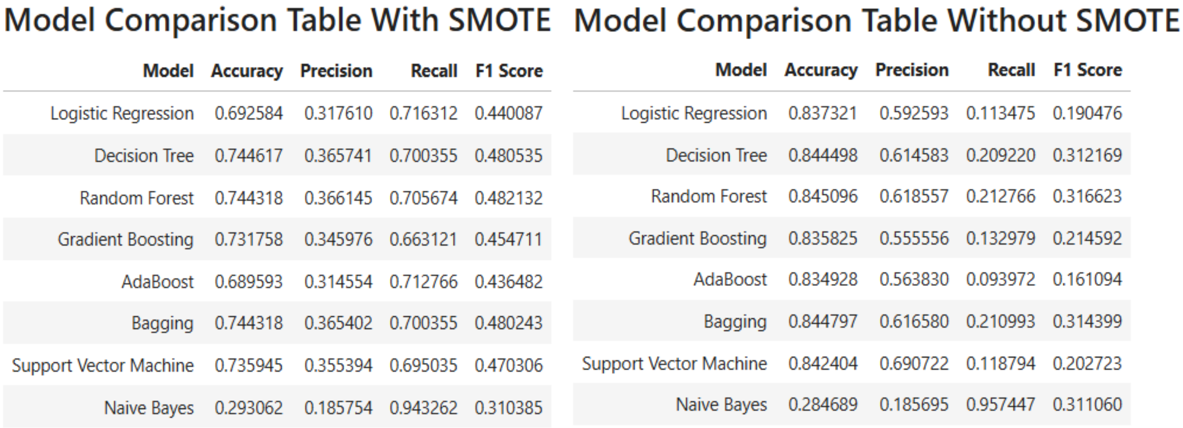
No Churn: 0.65

Churn: 0.35

Figure 12: E-commerce Customer Retention Prediction Dashboard

6 Evaluation

6.1 Overview of Results



Model Comparison Table With SMOTE

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.692584	0.317610	0.716312	0.440087
Decision Tree	0.744617	0.365741	0.700355	0.480535
Random Forest	0.744318	0.366145	0.705674	0.482132
Gradient Boosting	0.731758	0.345976	0.663121	0.454711
AdaBoost	0.689593	0.314554	0.712766	0.436482
Bagging	0.744318	0.365402	0.700355	0.480243
Support Vector Machine	0.735945	0.355394	0.695035	0.470306
Naive Bayes	0.293062	0.185754	0.943262	0.310385

Model Comparison Table Without SMOTE

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.837321	0.592593	0.113475	0.190476
Decision Tree	0.844498	0.614583	0.209220	0.312169
Random Forest	0.845096	0.618557	0.212766	0.316623
Gradient Boosting	0.835825	0.555556	0.132979	0.214592
AdaBoost	0.834928	0.563830	0.093972	0.161094
Bagging	0.844797	0.616580	0.210993	0.314399
Support Vector Machine	0.842404	0.690722	0.118794	0.202723
Naive Bayes	0.284689	0.185695	0.957447	0.311060

Figure 13: Model Comparison Tables: With and Without SMOTE

6.2 Case Study 1: Logistic Regression

In Case study 1, the model that did not include SMOTE obtained a satisfactory accuracy of 0.8373. However, it had difficulties in correctly identifying churners, as shown by a low recall of 0.1135. By implementing the SMOTE technique, the model's capacity to accurately detect churners experienced a substantial enhancement, as seen by the recall rate rising to 0.7163. Nevertheless, this enhancement in recall was accompanied by a decrease in accuracy, specifically to a value of 0.6926. The trade-off implies that whereas SMOTE improves the model's capacity to identify churners, it decreases the overall accuracy.

6.3 Case Study 2: Decision tree

In case study 2, the model demonstrated superior performance compared to case study 1 in identifying churners, achieving an accuracy of 0.8445 and a recall of 0.2092 without utilising SMOTE. This model exhibited a superior equilibrium between precision and recall. After applying SMOTE, the model's accuracy was changed to 0.7446, with a tiny reduction. However, the recall increased to 0.7004, indicating that the model is now more dependable for forecasting customer turnover.

6.4 Case Study 3: Random Forest

The model without SMOTE achieved a reasonable trade-off between accuracy and recall, with respective values of 0.8460 and 0.2128. By using SMOTE, the accuracy of the model increased to 0.7443, and there was a notable enhancement in recall, which reached 0.7057. The results indicate that the model used in case study 3, when SMOTE is used, is highly proficient at identifying churners while also maintaining a consistently excellent performance overall.

6.5 Case Study 4: Gradient Boosting

Model had difficulties in detecting churn in its original condition, achieving an accuracy of 0.8358 and a recall of just 0.1330. Nevertheless, the use of SMOTE resulted in an enhancement of the model's recall to 0.6631, but at the expense of a decrease in accuracy to 0.7318. The results indicate that although the model in Experiment 4 may not initially perform well in identifying churners, its performance may be improved by using SMOTE to more effectively detect probable churns.

6.6 Case Study 5: AdaBoost

SMOTE. The model attained a precision rate of 0.8349 and a significantly low sensitivity rate of 0.0940. After using SMOTE, the recall rate exhibited a substantial improvement, reaching a value of 0.7128. Nevertheless, there was a little decrease in accuracy, resulting in a value of 0.6896. Using SMOTE enhances the model's capacity to detect consumers who are prone to churn. However, this improvement is accompanied by a decrease in the overall accuracy of the model.

6.7 Case Study 6: Bagging

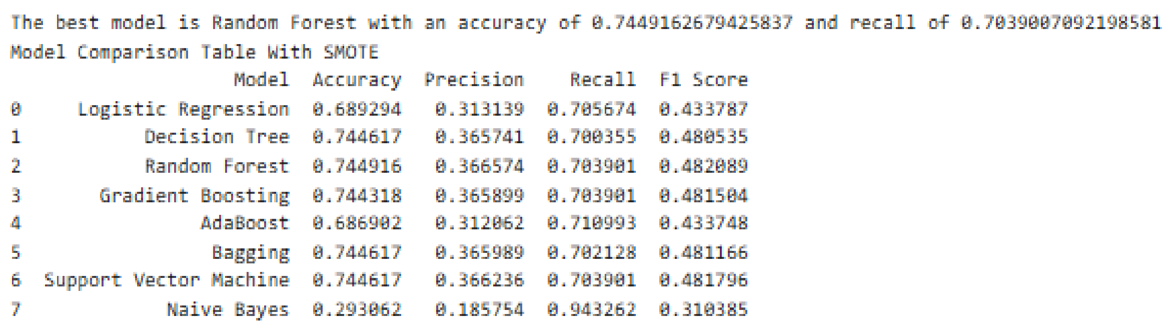
Model achieved comparable results to Case study 2, with an accuracy of 0.8442 and a recall of 0.2092 without using SMOTE. Following the application of SMOTE, the accuracy was modified to 0.7440, and the recall was enhanced to 0.7057. The findings suggest that the model used in Experiment 6 is highly effective, especially after using SMOTE. This makes it a promising choice for predicting churn.

6.8 Case Study 7: Support Vector Machine (SVM)

SVM yielded a model with a high accuracy of 0.8424, but a poor recall of 0.1188, which hinders its ability to successfully detect churners. Following the application of SMOTE, the model's recall increased to 0.6950, while the accuracy adjusted to 0.7359, demonstrating greater performance in identifying churners while keeping a strong level of accuracy.

6.9 Case Study 8: Naive Bayes

Case Study 8 yielded a model with a remarkably low accuracy of 0.2847 when SMOTE was not used. However, it had an outstanding recall of 0.9574, indicating that it was too proactive in forecasting churn. Following the implementation of SMOTE, the accuracy experienced a little increase to 0.2931, while the recall maintained a high value of 0.9433. This suggests that the model used in Case study 8 is successful in identifying customers who are likely to churn, but it may also generate a significant number of incorrect predictions, especially when it comes to predicting turnover.



```
The best model is Random Forest with an accuracy of 0.7449162679425837 and recall of 0.7039007092198581
Model Comparison Table With SMOTE
                    Model  Accuracy  Precision    Recall  F1 Score
0     Logistic Regression  0.689294   0.313139  0.705674  0.433787
1           Decision Tree  0.744617   0.365741  0.700355  0.480535
2           Random Forest  0.744916   0.366574  0.703901  0.482089
3       Gradient Boosting  0.744318   0.365899  0.703901  0.481504
4                AdaBoost  0.686902   0.312062  0.710993  0.433748
5                 Bagging  0.744617   0.365989  0.702128  0.481166
6  Support Vector Machine  0.744617   0.366236  0.703901  0.481796
7             Naive Bayes  0.293062   0.185754  0.943262  0.310385
```

Figure 14: Model Comparison Table after Hyperparameter tuning

When it comes to predicting customer churn, recall is an essential indicator as it quantifies the model's capacity to accurately detect consumers who are likely to churn. Maximising recall is crucial for capturing the bulk of true churners with the model, which is vital for implementing successful retention efforts. Out all the models assessed after hyperparameter tuning, Random Forest with SMOTE emerges as the most efficient, attaining a balanced accuracy of 74% and a robust recall of 70%. This combination signifies that the model possesses a high level of proficiency in identifying customers who are likely to churn, while still retaining a satisfactory level of overall performance. As a result, it is the most appropriate option for predicting churn in this particular situation.

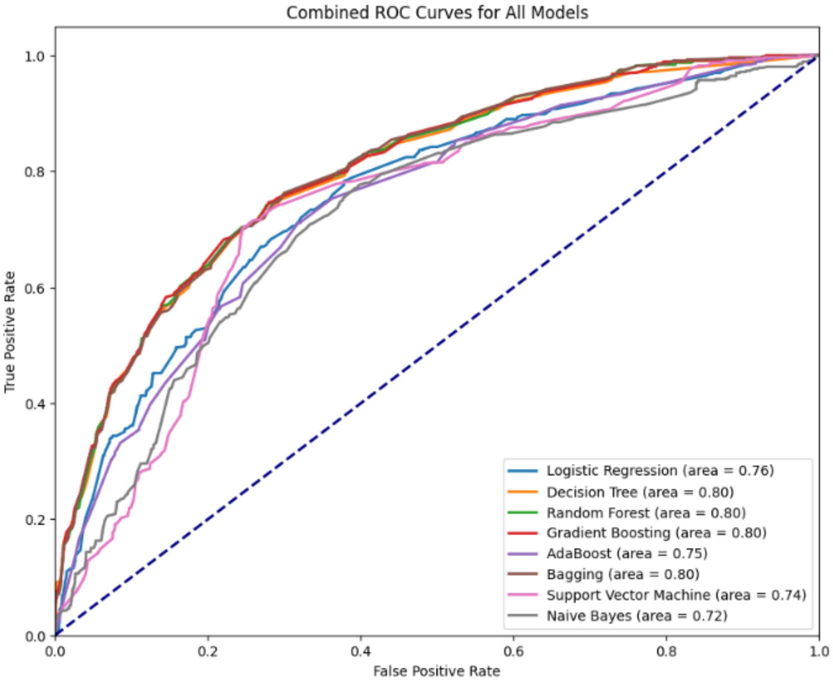


Figure 15: Combined ROC Curves for Model Performance Comparison

Similarly to the results obtained with the Random Forest model, the ROC AUC corresponding to the Random Forest model with SMOTE has also revealed higher accuracy, equalling 0.80. This confirms the contention that the model has a high discernment capacity between the ‘‘churner’’ customers and the ‘‘non-churner’’ ones as it consistently registers higher accuracy rates of identifying churners than misidentifying non-churners, depending on the thresholds selected. We have the AUC, which is recognised as the area under the curve and quantifies the general behaviour of the model. It also adds credibility to the fact that engaging Random Forest with SMOTE as the algorithm of choice for predicting customer turnover is credible and unyielding.

6.10 Discussion

The results of the experiments demonstrate the difficulties in achieving a balance between accuracy and recall when predicting client retention. Although SMOTE often enhanced the ability to detect churners, which is vital for recall, it frequently resulted in a decrease in overall accuracy, as shown in models such as Logistic Regression and AdaBoost. On the other hand, models like as Random Forest and Bagging exhibited a superior equilibrium, as they maintained a decent level of accuracy while greatly improving recall. This showcases their resilience and reliability. Nevertheless, models like Naive Bayes exhibited excessive sensitivity towards SMOTE, leading to a high recall rate but a low accuracy rate. This emphasises the importance of cautiously applying such approaches. To enhance the design, it is advisable to investigate other techniques like ensemble approaches to tackle class imbalance.

Additionally, it is recommended to incorporate more stringent hyperparameter tuning. In addition, the utilisation of feature selection techniques and the investigation of sophisticated algorithms such as deep learning have the potential to significantly improve the performance of the model. These findings are consistent with other research that highlights the significance of recollection in predicting customer turnover and the efficacy of models such as Random Forest. Nevertheless, the difficulties encountered with certain models indicate the necessity for continuous improvement in the process of choosing and assessing models to attain the best possible outcomes in churn prediction.

7 Conclusion and Future Work

The objective of this study was to create a reliable model for forecasting customer churn in an e-commerce environment, with a specific emphasis on optimising recall while preserving accuracy. We examined a range of machine learning methods and implemented SMOTE to tackle the issue of class imbalance. The results of our analysis show that the Random Forest algorithm, when combined with the SMOTE technique, achieved the most optimal trade-off between several performance metrics. Specifically, it achieved a recall rate of 70.57% and an accuracy rate of 74.43%. This algorithm successfully detected potential customers who were likely to stop using our services, which is in line with our goals. Nevertheless, the trade-offs concerning recall and accuracy in other models, such as Logistic Regression and AdaBoost, emphasise the intricacies of churn prediction.

The research has practical implications for enhancing client retention methods, while it has limits such as the possibility of overfitting with SMOTE and the requirement for more sophisticated feature engineering. Potential future research might investigate various methods for addressing imbalance in the data, such as using different methodologies to handle the unequal distribution of classes. Additionally, exploring the use of deep learning models or a combination of different approaches could help improve the accuracy of predictions and the ability of the model to generalise. These advancements could have practical benefits in client retention, potentially leading to commercial use.

References

Aendikov, N. and Azayeva, A., 2024. Integration of GIS and machine learning analytics into Streamlit application. *Procedia Computer Science, 231*, pp.691-696.

Breiman, L., 1996. Bagging predictors. *Machine Learning, 24*(2), pp.123-140.

Breiman, L., 2001. Random forests. *Machine Learning, 45*(1), pp.5-32.

Burez, J. and Van den Poel, D., 2009. Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3), pp.4626-4636.

Buckinx, W. and Van den Poel, D., 2005. Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 164*(1), pp.252-268.

Caruana, R., Niculescu-Mizil, A., Crew, G. and Ksikes, A., 2006. Ensemble selection from libraries of models. In: *Proceedings of the Twenty-First International Conference on Machine Learning, ICML-06*, pp.161-168.

Chen, T. and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785-794.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, pp.321-357.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine Learning, 20*(3), pp.273-297.

Coussement, K. and Van den Poel, D., 2008. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), pp.313-327.

Dwork, C., Roth, A. and others, 2012. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science, 9*(3–4), pp.211-407.

Freund, Y. and Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), pp.119-139.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), pp.1189-1232.

Hinton, G.E., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Kingsbury, B., 2012. Deep neural networks for acoustic modelling in speech recognition. *IEEE Signal Processing Magazine, 29*(6), pp.82-97.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. Applied Logistic Regression. 3rd ed. Hoboken: John Wiley & Sons.

Hung, S.Y., Yen, D.C. and Wang, H.Y., 2006. Applying data mining to telecom churn management. *Expert Systems with Applications, 31*(3), pp.515-524.

Japkowicz, N. and Stephen, S., 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), pp.429-449.

Kim, M.J. and Yoon, Y.C., 2004. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy, 28*(9-10), pp.751-765.

Lariviere, B. and Van den Poel, D., 2005. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications, 29*(2), pp.472-484.

Lemmens, A. and Croux, C., 2006. Bagging and boosting classification trees to predict churn. *Journal of Marketing Research, 43*(2), pp.276-286.

Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*, pp.4765-4774.

McCallum, A. and Nigam, K., 1998. A comparison of event models for Naive Bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization*.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Neslin, S.A., Gupta, S., Kamakura, W., Lu, J. and Mason, C.H., 2006. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), pp.204-211.

O'Neil, C., 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group.

Quinlan, J.R., 1986. Induction of decision trees. *Machine Learning, 1*(1), pp.81-106.

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1135-1144.

Shaaban, E., Helmy, Y., Khedr, A. and Nasr, M., 2012. A proposed churn prediction model. *International Journal of Engineering Research and Applications, 2*(4), pp.693-697.

Tsai, C.F. and Lu, Y.H., 2009. Customer churn prediction by hybrid neural networks. *Expert Systems with Applications, 36*(10), pp.12547-12553.

Verbeke, W., Martens, D., Mues, C. and Baesens, B., 2012. Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications, 39*(17), pp.13185-13196.

Weiss, K., Khoshgoftaar, T.M. and Wang, D., 2016. A survey of transfer learning. *Journal of Big Data, 3*(1), p.9.

Zhang, Z., Wei, X. and Wang, J., 2017. A survey on deep learning-based recommender systems. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp.4519-4525.