E-commerce Churn: Definition and Prediction – The Best Modelling Approach

By

Syed Muhammad Ameer Ghaznavi

Fall-2019-MS-CSE

Supervisor

Dr Mehak Rafiq

A thesis submitted in partial fulfilment of the Master of Computational Sciences and Engineering degree requirements

In

School of Interdisciplinary Engineering and Sciences,

National University of Sciences and Technology,

Islamabad, Pakistan.

March 2023

E-commerce Churn: Definition and Prediction – The Best Modelling Approach

By

Syed Muhammad Ameer Ghaznavi

Fall-2019-MS-CSE

Supervisor

Dr Mehak Rafiq

A thesis submitted in partial fulfilment of the Master of Computational Sciences and Engineering degree requirements

In

School of Interdisciplinary Engineering and Sciences,

National University of Sciences and Technology,

Islamabad, Pakistan.

March 2023

Dedication

Foremost to Almighty Allah for giving me the willpower and strength to complete my dissertation and to my family for their endless love, support and encouragement throughout my pursuit of education. I hope this achievement will fulfil the dream they envisioned for me.

Certificate of Originality

I hereby declare that this submission is my own work and that, to the best of my knowledge, it contains no materials previously published or written by another person nor material that has been accepted for the award of any degree or diploma at the Department of Computational Sciences at or any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution to the research made by others, including those with whom I have worked or elsewhere, is explicitly acknowledged in the thesis. I also declare that, except for assistance from others in the project's deliverable or in style, presentation, and linguistics, the content of this research work is the product of my own work.

Author Name: Syed Muhammad Ameer Ghaznavi

Signature: _____

Acknowledgements

First, I want to express my thanks and praise to Almighty Allah, who has enabled me to understand, think and complete this case report work in His great mercy and benevolence. My heartiest praise goes to the holiest man in the galaxies, Almighty's beloved, Hazrat Muhammad (Peace Be upon Him), humanity's reformer.

I want to express my sincere gratitude and thanks to my supervisor Dr Mehak Rafiq for showing me support and guidance for my PostGrad study; she helped me through all the study time and writing of this report. I also want to thank the member of my thesis committee, Dr Muhammad Tariq Saeed and Dr Shahzad Rasool for their valuable feedback and suggestions throughout the research process. I would also like to thank my teachers for their continuous subsistence and understanding when undertaking my coursework. Your prayers for me were what sustained me this far.

I would like to thank my family for their contributions, which did not let me down until now, for their encouragement, love, care and sincerity, without which I would not be able to fulfil my education at its best.

Finally, I would like to conclude by extending a special thank to Muhammad Sohaib, Mohsin Ahmad, Farhan Bashir, Haseeb Sultan, Mashaal Shah, Huzaifa Arshad, Zunaira Said and all the other data analytics lab members for their feedback sessions and support necessary to complete this research. The time I spent with them will always be cherished.

Table of Contents

List of Figures

List of Tables

Abbreviation

AI	Artificial Intelligence
ML	Machine Learning
E-Commerce	Electronic Commerce
B2B	Business-to-Business
B2C	Business-to-Consumer
C2C	Consumer-to-Consumer
C2B	Consumer-to-Business
Covid-19	Coronavirus Disease of 2019
CLV	Customer Lifetime Value
LR	Logistic Regression
DT	Decision Tree
RF	Random Forest
SVM	Support Vector Machine
CNN	Convolutional Neural Network
URL	Uniform Resource Locator
ATM	Automated Teller Machines
SMOTE	Synthetic Minority Oversampling Technique
MRMR	Minimum Redundancy, Maximum Relevance
RFM	Recency Frequency Monetary
UTC	Universal Time Coordinates
Xg-Boost	Extreme Gradient Boosting
AUC	Area under the ROC curve
ROC	Receiver Operating Characteristics

Abstract

The vast accessibility and advancement of the internet have made it an essential component of modern companies and organizations. Particularly in recent times, with the emergence of the COVID-19 pandemic, the adoption of online platforms and digital solutions has become increasingly prevalent among businesses to connect with their customers. Ecommerce refers to the buying and selling of products or services over the internet. Online firms interact with clients under non-contractual terms, making it difficult to track customer retention. One of the major challenges encountered by e-commerce is churn, which refers to the situation when a customer stop buying a product or service for a prolonged period. The churn rate in e-commerce is closely linked to a company's revenue, as retaining customers leads to higher margins compared to randomly acquiring new customers. It is estimated that the cost of acquiring new customers is four to five times that of retaining existing customers. The foremost objective of this research is to determine the most effective approach for identifying potential customer churn in the e-commerce industry. To carry out the analysis, an unlabelled dataset obtained from an e-commerce store is used to obtain insights regarding customer purchasing pattern. The data undergoes various stages of preprocessing and during this process, new features are derived from the original dataset. To label the customer data, three distinct churn indicator techniques has been applied. These techniques include a comparison of the average purchase duration of customers, the implementation of the RFM (recency, frequency, and monetary) method, and the application of a K-means unsupervised learning algorithm. Ultimately, a comparative analysis of several machine learning classification algorithms is performed to develop an accurate churn prediction model. This study constructed nine models by employing the Random Forests, Support Vector Machine, and Extreme Gradient Boosting algorithms in conjunction with three defining criteria. These models were then evaluated based on a range of performance metrics, including precision, recall, f1-score, accuracy, and auroc. The models attained their highest accuracy when trained on data that had been labelled using the RFM method, with accuracies of 86% and 82%, respectively. Additionally, the memory and time consumption of the models were assessed, and it was discovered that the support vector machine classifier used the least amount of memory, while the extreme gradient boosting approach demonstrated the most time-efficient performance.

Chapter 1

Introduction

This chapter briefly describes the motivation behind selecting and exploring the problem, majorly e-commerce, its different types, and the rise and acceptance of online shopping. It also summarizes the core problem, customer churn, its effect on e-commerce industries and its importance.

1.1 E-Commerce

E-Commerce means electronic commerce; whenever individuals and companies are buying or selling products/services online, they are involved in e-commerce [1]. The internet is a global means of communication with far more reach than the traditional business model. E-commerce, or electronic trade, has flourished due to the Internet's rapid growth [2]. Before being accepted by the larger population, people used to purchase accessories by going to physical stores. Some issues customers may face include the distance travelled to obtain the product, locating the desired item, and little shop operating time. The advent of Covid-19 [1] shows that businesses can close due to uncertain conditions leading to the dire need for most businesses to have some e-commerce available. There are many advantages of e-commerce over traditional trading. Some of them to mention here are; the accessibility to buy the product at any time, availability of buying options nationally and internationally, the more extensive reach of the audience connected to social media platforms, and customers' suggestions and opinions about the product in the review section [3]. In addition, various mediators or players exist in conventional business, like the distributor and wholesaler.

Figure 1: Stages of the traditional business model contain multiple intermediaries like Manufacturer, Distributor, Wholesaler and Retailer

In Figure 1, it can be seen how a product is moved from different vendors before it reaches the customer. Whereas in online business, there is direct selling, all those third-party companies or vendors are skipped, resulting in cost reduction. Once a company has their manufacturing unit, they are moved to the warehouse for storage once the goods are ready to sell. When a customer places an order, the shipper takes the package from the warehouse and is delivered it to the client's address. With time, the method of conducting business is also evolving. E-commerce is gaining friction due to widespread internet use, and start-ups are increasingly turning to this medium as a better and more distinct business model.

1.2 Types of Ecommerce:

E-commerce is a broader term for doing online business. It can be divided into further different segments like the business-to-business model, business-to-consumer, consumer-to-business and consumer-to-consumer model. Although each has its challenges and benefits, some companies coexist in more than one category.

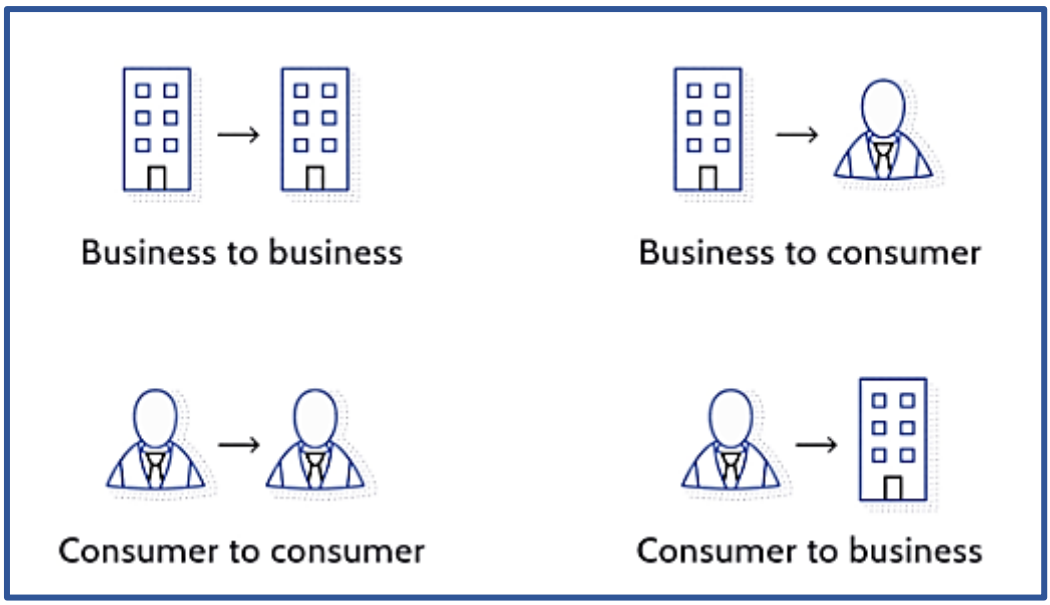


Figure 2: E-commerce comprises four primary business paradigms depending on the parties involved in the transaction [4]

Different models serve differently depending on the goods [5]. If someone makes their items, they might focus on wholesaling to cover production costs and break even faster. If someone is a distributor of someone else’s products, they must put more money into indirect marketing and customer acquisition effort. This research focuses on the business-to-customer model, but it may be applied to other segments of e-commerce with some modification.

1.2.1 Business to Business – B2B

The B2B framework refers to trading directly between two companies . At times the buyer is a consumer, but usually, he resells to the consumer. It includes big orders and cost reduction as, in most cases, work is done through automation that excludes the chances of error. For example, Intel is selling its processor to Dell. Wholesalers, larger retailers, and a variety of other organizations such as schools, non-profits, and others work with B2B vendors.

1.2.2 Business to Consumer – B2C

The B2C business model refers to selling goods to individuals directly. It is the most commonly used model, and most customers are familiar with it. For instance, when a customer buys in an online store like electronics, garments, or household is done through a B2C transaction [6]. It includes not only goods but also services. A famous example of a business-to-customer platform is Amazon. If a customer wants to purchase a product using the B2C model, the first step is creating an account and providing personal information that firms will use for marketing and promotion.

1.2.3 Consumer to Business – C2B

This business model allows individuals to sell items and services to companies [3]. In this framework, the customer may have the product that attracts a company, for instance, people like bloggers or those with a significant following on social media platforms for which consumers can post reviews. A consumer can share and review the product and get paid by the company. For example, smartphone companies may hire a famous TV actor to promote mobile phones among their fan following, and the actor gets paid. Other examples are the freelancing websites through which consumers can work for companies and organizations.

1.2.4 Consumer to Consumer – C2C

In this type of e-commerce exchange of products takes place among customers using a third-party online platform. Consumers are both sellers and buyers instead of businesses [5].

One of the pioneers in this category is eBay which has a unique feature of online auction where the highest bidder wins the item. The advantage of C2C over other segments of e-commerce models is that customers can deal with each other directly and also have

the leverage to bargain. Amazon lies in two categories: the Business to customer and the customer-to-customer model.

1.3 E-commerce Boom

It is not surprising that e-commerce continues to prosper and expand. Several reasons have boosted the e-commerce industry.

1.3.1 Advance Technology

Firms and businesses use the internet as a vital element of their corporate operations. Over time, technology is advancing, and more people are connected to the internet than ever, enabling e-commerce companies to draw in consumers [1]. Customers are looking for new and more convenient ways to meet their sales demands as their lifestyles become busier and technology advances. Several other causes are accelerating the e-commerce revolution in addition to technology.

1.3.2 Accessibility

Online retailers are no longer bound by traditional store hours. Instead, sales can take place at any time. Due to the global rise of e-commerce, sellers can now sell their items in multiple countries without building a physical outlet. Making businesses more accessible at any time attracts new clients [7]. Purchasing online can benefit some people with disabilities who may struggle with physical accessibility in businesses. Access to a store or customer service staff might also help those who do not have time to go out of town.

1.3.3 Multiple Options

Furthermore, consumers desire more options. They appreciate having a variety of colours, sizes, styles, and specification options. While traditional retailers are limited in their product selection, e-commerce allows for all these possibilities. Items are kept in central warehouses, which are far less expensive than store outlets. Customers can browse the entire inventory and purchase as per their wish.

1.3.4 Social Media Integration

Social media continues to play an increasingly important part in modern life. It is evident to see why e-commerce and social networking apps are collaborating.

People are increasingly looking to their phones for answers to everyday difficulties. For example, Instagram is one of the most popular social media platforms impacting retail [8].

1.3.5 Global Pandemic

The global epidemic of COVID-19 has had a significant impact on e-commerce. [1] With store closures due to global lockdowns and the rapid growth of internet shopping, merchants had no alternative but to react to new market dynamics quickly. Small companies are emerging, and existing companies are constantly expanding. The transition of making an online presence is more frequent under such situations than before. Furthermore, the social distancing protocol has led to an increase in global e-commerce.

1.3.6 User-Friendly Interaction

E-commerce's user-friendly approach to everyday tasks is one of the main reasons for its growth. A website or app makes it easy to grocery shop, buy new clothes, and even get prescriptions. Consumers do not have to travel, go through crowded areas, or fight traffic. Products that customers are interested in should be readily visible with simple navigation.

1.4 Customer Churn

After understanding what e-commerce is and how it grows with time, it is essential to know the challenges online businesses encounter. One of the problems is customer churn (a word used in the business world). Customer churn is a situation that occurs when a customer discontinues or stops using a product or service for a prolonged time. Such a customer is considered a churn or lost customer [9].

The churn rate and the company's overall performance are linked. According to the research, keeping clients results in higher margins than randomly recruiting new ones [2]. There are several reasons based on which customers may leave. For example, the website's interface is not easy to use, product quality is going down, service is not up to the mark, the price of the product is, and customer needs concerning what is offered.

The churn rate in e-commerce is significant, and the dataset is unbalanced. The customer is the real asset of any company; after acquiring customers to make them stay for a more extended period, it is essential to satisfy them. Unfortunately, in many instances, things go wrong, leading to customer churn.

1.4.1 Churn in Different Sectors

Most of the studies conducted about churn are in Telecom, Banking and Management [7]. Although, there are businesses, such as gaming and other subscription-based applications, where people are focused on early detection of churn to mitigate the impact.

For example, in the case of games, if a user is not active for a particular period or not making any purchases, it is an alarming sign of being lost. The movie streaming subscription model, where customers are not paying monthly, or yearly charges, is considered churn. Understanding the behaviour and choices of customers or users in any industry is critical to keeping them for a long time. In different industries, different methods or attributes are utilized to predict churn.

Today's businesses are more influenced by their client age; Generation Z and millennials have more access to and knowledge of a wide range of technological products, which is an essential factor for firms looking to improve their digital communication [10].

1.4.2 Churn in Ecommerce

When it comes to e-commerce, customer churn can be divided into two categories one is complete churn and second is low-value churn. Complete churn occurs when customers no longer use a company website and shop from other companies. Low-value churn occurs when a customer's monthly or annual consumption drops dramatically [2]. The acquisition is critical for newer brands or retailers wanting to expand their customer base; therefore, intelligent customer retention techniques are needed. Customer retention is the capacity to keep customers returning once they have made their first purchase. Covid-19 has also boosted the digital economy. According to the Google search engine [11], the demand for terms like computers and smartphones peaked in March and April 2020, which shows that more individuals are drawn to digital media, which was once a want but is now a necessity due to the times. Therefore, online shoppers have high expectations for the services they receive because of the massive demand for this shopping.

The reality is that consumers' learning curves for digital acceptance differ greatly depending on their age group. The youth sector, especially Millennials and Generation Z, accept the digital realm far more quickly and efficiently than their elder counterparts.

Nevertheless, at the same time, the group is vulnerable to changes in service quality [12].

So, it is imperative to satisfy the customers because any error or failure during the transaction could result in the client being lost. So, this information about customers who might go missing is crucial for a company.

1.4.3 Importance of Churn Prediction

For any business, it is critical to analyse consumers' purchasing patterns and determine whether a client has churned or is taking a break. Once this has been determined, this may help determine marketing techniques to enhance the possibility of customer loyalty [13].

Knowing where a customer is likely to go and providing incentives to keep them staying can save a firm much cost; however, in case of incorrect prediction of customers, the model may lose the customer section, which is intended to be saved.

1.5 Problem Statement and Solution

E-commerce stores interact with consumers in a non-contractual context, and acquiring a new customer is also expensive. Existing clients usually consume more services and may generate more client referrals. As the client is the company's most valuable asset, it is critical that they are pleased and remain with the firm for an extended time.

1.5.1 Problem Statement

Customer attrition or churn is one of the most serious concerns that an e-commerce firm encounters, which is the loss of customers (move to a competitor or voluntary churn) and especially in the case where data is not labelled, it is more difficult to determine whether a client is still alive or has perished.

1.5.2 Solution

By assessing customer behaviour and preferences, our system will first set the churn criterion through which customers will be labelled using three different methods and then predict which customers are likely to be churned or loyal using machine learning algorithms.

1.6 Research Objective

The objectives of our study are as follows.

- Examining the customer base, then establishing several approaches to categorize customers as non-churn or churn.
- Data is passed through several machine learning and AI-based models to classify consumers.
- To Evaluate and compare models on performance metrics like accuracy, f1-score, and other factors such as storage and time.

The first target was achieved by performing an in-depth examination of the data using the most effective data analytics tools and techniques (data pre-processing). The secondary goal was achieved by applying and discussing the most highlighted aspects of churn modelling based on the results of machine learning techniques.

Chapter 2

Literature Review

This chapter will examine contributions made by other churn prediction researchers by offering an overview of their work. It begins with a brief history of churn prediction, followed by some examples from other businesses, before narrowing it down to e-commerce.

2.1 Churn-Related Work

As described above, churn is when a customer is inactive and not involved in any transaction for a prolonged period. Over the last decade, research has been conducted to identify churn clients. It is one of the biggest challenges that most companies and businesses are facing in the era of the internet and e-commerce. Kesiraju and Deeplakshmi explained churn from another perspective [14]; in case of non-payment of bills or customer involvement in any fraudulent activity, these types of customers are removed by the company itself, which is called involuntary churn. On the contrary, when the customer decides to change to leave a company, such activity is known as voluntary churn.

Different researchers have developed and implemented machine learning models that can predict client behaviour and patterns in advance. Most of the previous studies for churn are predictions carried out in the telecom and banking sectors. However, user churn is being researched in various fields as it applies to most industries. Before diving into details, it is vital to understand the distinction between the two types of churns: contractual and non-contractual churn [7].

When a customer loses interest in the service and does not renew their contract, it is referred to as contractual churn. This is when the customer first pays for the service and can be done monthly, quarterly, and yearly depending on the company standards. On the other hand, non-contractual churn is when customers are not reqsuired to pay an investment fee and have the leverage to move to competitors at any time. This research study concentrates on non-contractual churn, that is, e-commerce.

2.2 Churn Analysis and Customer Behavior

The inactivity period of a customer is different according to each research field. In the past, customer churn used to be caused explicitly by contract cancellations, but with modern services such as the internet and retail, customer churn is more common due to low client investment costs [15]–[17]. Before declaring a customer lost, it is important to investigate customer behaviour towards the company's product or service. E. Lee, B. Kim, et al. explained in their study that when the duration of inactivity or behavioural change exceeds a certain threshold, the customer is considered a churned customer [18]. Difference service features have their criteria for determining this time window. T. Huang et al. (2019) investigated log data to determine the churn duration of mobile games, and the results revealed that more than 95% of users did not return after being gone for three days [19]. Therefore, they decided on a three-day churn period.

Table 1: Ten-week activity log of the customers using a time window method of three weeks to observe customer transactions to label whether a customer exists or has lost [18]

	Observation Period (Log data)									
	Before window			Churn Determination Window				After window		
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
User A							Churned			
User B							Churned			
User C							Churned			
User D							Survived			
User E							Survived			
User F							Survived			

Table 1 above is an example of a customer activity log, where three weeks is taken as a time window. The active customers from week four to week six are considered to survive, and other clients are labelled churned [18]. So, for example, users A, B and C are active before the time window, and then after the time window, Customers B and C still return; they are considered churn or lost.

The insurance and financial industries also predicted churn using the customer lifetime value metric, which is used to predict customer value for future cash flow. Researchers have used different parameters to calculate the expected future value of a customer [20]. The basic formula contains the price at the time, the direct cost of serving customers, the discount rate, the acquisition cost and the probability of customer repeat. By

analysing the client time for each product, it is shown that the chance of churn varies depending on the tendency of consumers who selected financial products, based on the assumption that the customer group was different according to the financial product attribute. A churner is someone whose Customer Lifetime Value (CLV) value decreases with time [21].

Furthermore, a competition to construct a prediction model was launched in the music streaming service industry, and churn research was also undertaken in the Internet service and newspaper subscription fields. Customer attrition is consistent with the contract renewal time for the newspaper subscription and music streaming service, both fixed-rate services. In addition, online dating, shopping, Q&A services, and social network-based services have all been studied for churn prediction.

Althoff and Leskovec conducted a study using data from DonorsChoose.org, a non-profit organization that funds educational projects. After a year of no donations, the author considered a donor a churner. They devised a complicated model with four categories of attributes defining time, donor, project, and the project's teacher. The most important attributes are those describing donors. A Logistic Regression algorithm was used to predict the donors at risk of being churned. The key attributes that are more responsible for finding churn are donor-related features [22].

In 2011, Pinar et al. utilized a naive Bayes classifier to forecast a telecom company's client attrition. According to their findings, customers' average call duration was substantially connected with customer attrition [23]. Decision Trees, the Support Vector Machine, and artificial neural networks are just a few of the machine learning-based prediction methods [24]. Decision Trees (DT), for example, are commonly employed in practical customer churn prediction.

Agrawal et al. demonstrate the telecom customer churn problem, retrieve necessary variables from the original data using a multi-layered neural network, and suggest an artificial neural network-based churn prediction model. The results showed that this strategy had a prediction accuracy of 80%, and it also mentioned the parameters that are related to churn. [25].

Mishra, Reddy et al. [26] discovered a novel approach for customer churn prediction using a convolutional neural network. They performed their analysis on telecom data comprised of 3333 customers. Area, email, message, call duration, day call, and daily

charge are some of the attributes out of 20 features. The basic concept of a Convolutional Neural Network (CNN) has been described in four steps: from convolution, non-linearity, and pooling and at the end, classification is performed by taking input from the hidden layer and displaying it in the output layer. The model contains three layers one input layer, one hidden and one output layer [26].

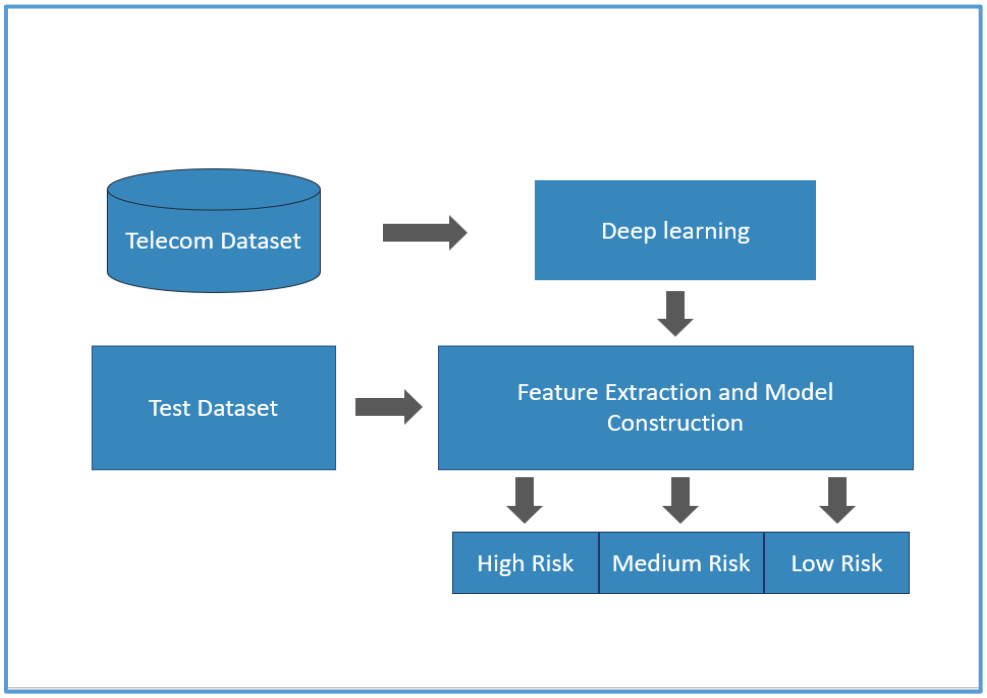


Figure 3: Flow diagram of customer churn prediction using a Convolutional Neural Network to identify the risk of switching to a competitor [26]

Figure 3 shows the general flow diagram of how the model works and how important features are extracted. The model's output is divided into three categories based on the risk of being churned.

Kim et al. predicted customer churn by analyzing social media sales data generated by influencers and customers. Influencers promote and sell items on Instagram as e-commerce does by uploading postings and uniform resource locator (URL) links to their profile. Previous studies that used e-commerce to predict customer churning were referred to determine the churning point. Churners are customers who do not purchase from an influencer more than once. If a purchase is made more than twice, they are considered loyal. The DT algorithm is applied to predict the churn clients and attain 90% accuracy based on F-measure [27].

Bharathi, Pramod et al. performed a survey to predict bank churn, specifically for India's youth sector. They used an online questionnaire for two reasons: the largest youth

population and the highest mobile penetration. The questions are related to customer demography and the relationship with the bank. [34].

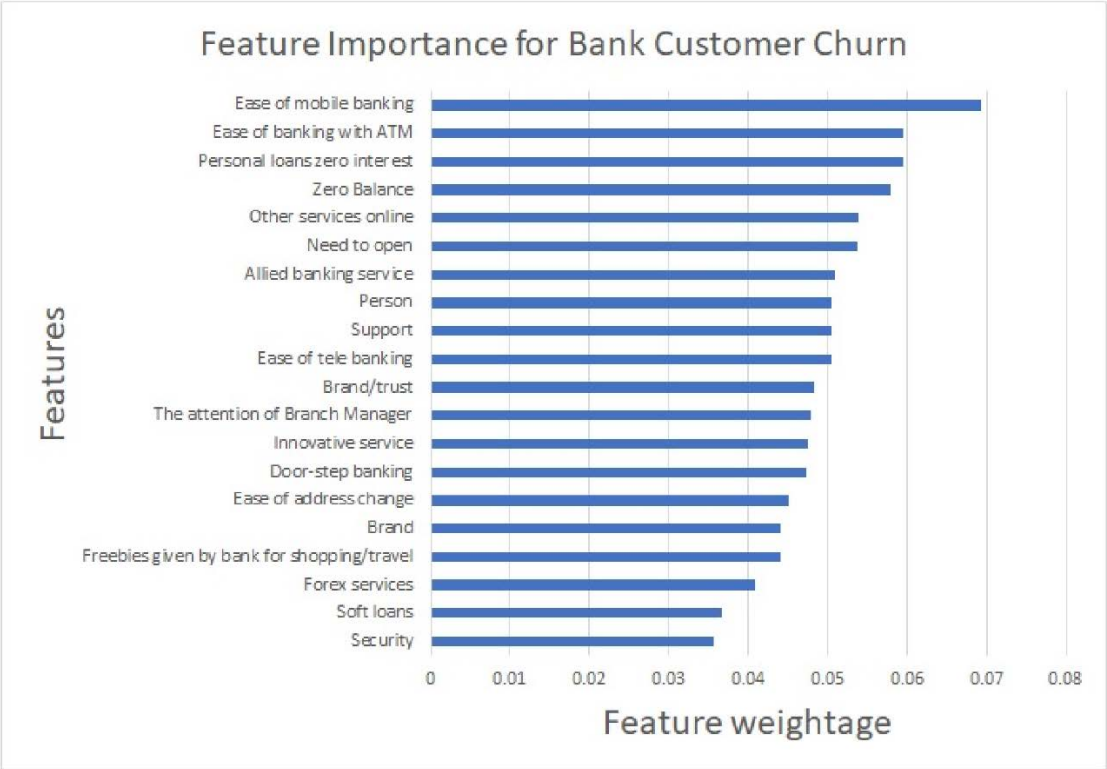


Figure 4: Numbers of features that are selected based on their importance which is shown in percentages on the x-axis for customer churn in the Banking industry [34]

Some critical factors responsible for considering customers as lost or existing are the absence of mobile banking, zero-interest personal loan, access to Automated Teller Machines (ATM), and customer care and support [34]. Figure 4 shows valuable features for bank customer churn for the country's younger population. After analyzing the customer responses, data is gone through various machine learning algorithms. The extra tree classifier comes up with the highest accuracy of 92%.

As one of the classic industrial sectors, the financial business has constantly been evolving for the past ten years. Banks now have an extensive database of their clients, which they use to acquire a strong position over their competitors, particularly in developing countries [28]. In addition, the banking industry's pandemic-driven digital revolution has generated a significant demand for customer churn analyses, particularly for younger clients.

Before Covid-19, consumer experience in retail banking mainly focused on client satisfaction, including loyalty, increased acceptance of services, and pricing. After

Covid-19, companies that want to be successful should focus on consumer experience like transparency, openness, and reliability, backed up by a stable and reliable digital environment [29], [30].

Most churn prediction research focuses on comparing different algorithms to reduce classification errors. In addition, the focus is on boosting prediction accuracy by using a new pre-processing method and evaluating other algorithms**.** Few studies have attempted to anticipate churn in developing sectors as opposed to standardized industries like telecommunications and finance.

2.3 Churn Prediction in Ecommerce

In different studies, the term “churn customer” is defined differently. According to Raeisi & Sajedi (2020), the definitions of churn customers might vary depending on each organisation's services. For example, they may be considered churning clients if they have not utilized the service in a month, three months, or even a year [31].

Customers now have several purchase options as a result of improvements in e-commerce. However, by making it easy for customers to share information, look for products, and move from one online shopping mall to another, e-commerce has increased the risk of churn. As a result of the variable data and difficulty in establishing the churning point, only a few studies on predicting customer attrition in e-commerce have been investigated.

Yanfang & Chen (2017) classified clients as churn if they did not use an e-commerce platform or an online shopping mall in the previous three months [9]. If there was no purchase history within the data collecting period, Zhuang (2018) characterized it as a churn [32]. B2B e-commerce customer loss was forecasted by Gordini et al., and the results demonstrated that the support vector machine worked well in processing noisy, unbalanced, and non-linear B2B e-commerce data [33].

The imbalanced data distribution is one of the primary challenges in churn prediction. Because attrition and retention of clients are not equal in e-commerce, the classifier will be biased towards the majority class. In the preprocessing step of the data, different under-sampling and oversampling approaches are utilised to achieve this. There is research that considers the data’s imbalance. To balance the data, Xiaojun et al. [34] implemented the Improved Synthetic Minority Oversampling Technique, also known

as SMOTE and Hanif et al. [35] employed random oversampling, respectively, then passed it to a classification algorithm for prediction. It can also be applied to other fields.

The selection of features that have an impact on model performance is another crucial component of data. Filter-based, wrapper and intrinsic approaches are all employed in supervised learning. The univariate link between each predictor and response variable is examined in filter-based techniques. In the case of a wrapper, multiple models are built with different input variables to discover the optimum combination that strengthens the model. The intrinsic technique automatically selects the best features. Hanif et al. [35] suggested a feature selection method that used Random Forest, Gradient Boosting, and MRMR (Minimum Redundancy, Maximum Relevance). Lalwani et al. [36] conducted another investigation in which a gravitational search method was used to pick features.

Other research recommends using an unsupervised learning strategy for churn prediction in conjunction with the Recency Frequency and Monetary method (RFM), a marketing model for segmenting clients. For example, a similar approach was utilized in this study [37], which used the K- Mean clustering algorithm to identify churn among retail consumers of a gift store in the United Kingdom.

Several research studies have been undertaken that compare machine learning models such as Logistic Regression (LR), Support Vector Machine (SVM), DT, K-Nearest Neighbor (KNN), and others. Different performance indicators are used to assess the model’s performance. Tingting et al. [38] suggested a customer churn prediction approach that is applicable for imbalance data sets, based on a cluster stratified sampling logistic regression model. An experiment is conducted using realistic public data sets. The model's performance was assessed using the ROC curves and AUC values, which were 90% and 91%, respectively.

According to Manohar et al. [39], churn prediction is a challenging research subject in e-commerce. Supervised learning algorithms, specifically classification algorithms, are used in this work. However, each algorithm has its own set of benefits. Therefore, instead of adopting a single algorithm, a combination of methods such as Bayesian Classifier, Support Vector Machine, and Random Forest are utilized to identify churn and develop an effective model. The results suggest that employing these algorithms collectively provides better results than using them individually.

Zhang et al. [2] conducted a survey and designed a questionnaire to find out the essential factors that make customer churn; the study provided five primary trust factors: website security and reliability, enduring and steady business, realistic product advertisement, corporate strength, and product cost performance and utilized a variety of data mining approaches to understanding client attrition, including Decision Trees, Clustering analysis, and Neural Networks, for customer churn model. In addition, the perception of online purchasing is influenced by educational level, gender, age and income, according to a published study [6]. While evaluating the influencing elements for online purchases, we reduce the original 20 variables to only four. Among these elements, the marketer's commitment to service quality is essential to consumers' online trust.

The research results of the present study [6] also show the correlations between consumers' perceptions of the elements that influence their desire to buy online, specifically, the customers' perceptions of confidentiality and protection, user interface, quality of services, and user experience aspects. The findings of this study [6] favour the conclusions of the prior studies that the most important criteria are the quality of service and web security, which affect and establish consumers' trust in online purchasing.

Several cluster algorithms are examined in [40] utilizing patient attrition data to assess their performance. An evaluation of several clustering algorithms was carried out. It was discovered that the K-means clustering performs better for customer churn analysis but requires effort in preprocessing phase to make data in proper shape before passing it to the model. When the number of clusters increases, K-medoids require additional repetition, and the distance is measured using the Euclidean distance function. The number of iterations and groups in fuzzy c means cluster must be specified before giving it to the model. When the data set is extensive, hierarchical clustering does not group all items in a single step; it requires more time and iterations to group things.

Another paper presents a multiple kernel support vector machines (MK-SVMs)-based customer churn prediction model that combines three information retrieval tasks, namely feature selection, class projection, and decision rule extraction, into a single model. A multiple iteration of two convex optimization problems is designed for feature extraction and classification prediction at the same time. Support vectors are employed to extract decision rules based on the identified characteristics. The effectiveness of this

approach is assessed using open CRM data. Research findings demonstrate that using a rebalanced strategy, MK-SVMs exhibit promising performance on the strongly skewed dataset, and the extracted rules achieve high coverage and low false alarm with a limited set of preconditions [41].

In summary of previous work, researchers used a variety of forecasting approaches to undertake in-depth research on churn using labelled data in the telecom, banking, and B2B e-commerce industries. These studies have also examined the benefits of various methodologies, adding to the knowledge of contractual customer churn prediction. The loss of customers by B2C e-commerce businesses is alarming because clients' purchasing habits are diversified, and their shopping intentions and preferences are tailored to them. As a result, this study examines the loss of non-contractual consumers of B2C e-commerce businesses utilizing customer data features.

2.4 Literature Gap

Most studies' literature is based on labelled data, with insufficient information on whether a customer is considered churn or loyal. Although there are studies labelling customers as churn, how this study differs from others is as follows:

Firstly no one has performed churn prediction on this E-commerce dataset which is used in this research problem. This study proposed three distinct ways to consider or define customers as churn. Lastly, three different machine learning models are used for predicting customers at risk of churn. Altogether there are nine models, and the best one chosen is based on performance metrics. Each model also analyzes time and memory consumption.

Chapter 3

Research Methodology

This chapter gives a general understanding of the project's overall approaches and methods for achieving the research's objectives after knowing what e-commerce is, the importance of churn analysis and past work related to it. Data acquisition is the first step in the technique, which continues until results are attained, and a clearly defined model is created.

3.1 Methodology Overview

The project is motivated to help e-commerce businesses that lose clients due to a lack of information. When there are a bunch of customers, it is easy to manage and satisfy them. However, as a company grows, it becomes increasingly difficult to satisfy its customers. The general flow and steps taken during this research are shown in Figure 5.

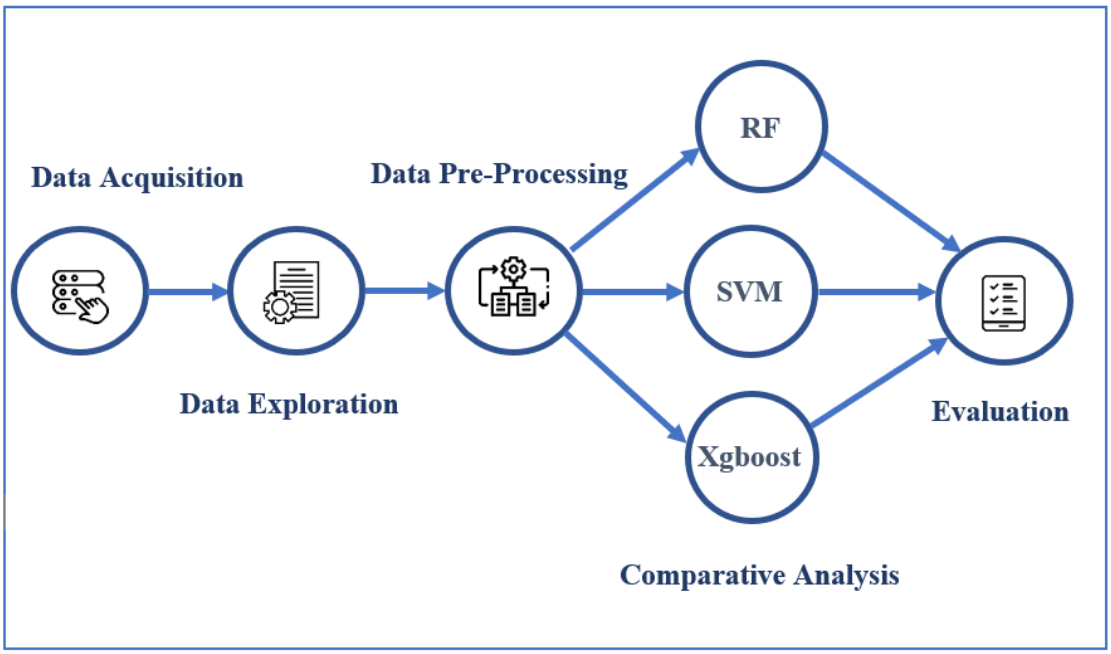


Figure 5: Methodology overview of the churn modelling in E-commerce; data is transitioned from various stages, starting from data selection and pre-processing in a way that is acceptable for the model training and validation

The first step in the methodology is acquiring the data and getting insight into the features. Then data is moved to the preprocessing phase, where it is manipulated to get it into the desired shape intended for the churn modelling. Finally, trained models are evaluated on different performance metrics.

In the latter section, each methodological step is explained in detail. Customer segmentation and predicting churn are two critical steps in this study. According to the research, online shoppers may engage in one of two behaviours: either a non-churn, or existing customer, a churn, or a consumer who will be lost in the future. Predicting customer attrition is, therefore, a binary classification problem. Furthermore, shopping time and behavioural characteristics may significantly influence determining client loss.

3.2 Data Acquisition

The dataset utilized in this study is a secondary dataset taken from an online repository called Kaggle that contains datasets from various topics. It is an open source and was initially gathered in the Middle East and North Africa. The file contains behavioural data for seven consecutive months from October 2019 to April 2020 from a large multi-category online e-commerce company.

3.2.1 Data Characteristics

The dataset is made up of approximately 1,162,048 customer purchase transactional records. Every record corresponds to a particular event. All events are related to products and customers. Many-to-many relationships exist between customers and products. By that, it means multiple customers can buy the same product, and each can buy multiple products. Table 2 mentions all the features or attributes of the dataset with its description.

Table 2: Brief explanation of the features that are present in the E-commerce dataset; information about the customer transaction; the time of the purchase, product-related details and price

Features	Description
event_time	The time when the event occurred
event_type	Types of events: purchase
product_id	Identification number of the product
category_id	Category identification number of product
category_code	Meaningful name of the category of items
Brand	Names of the brand
Price	The present price of the product
user_id	Customer Identification Number
user_session	Temporary code assigned

The dataset is not labelled, indicating that it does not have customer status information. Therefore, one initial task is finding features and patterns to help label customers.

3.3 Data Exploration

The data that is under consideration is time-series data. In this study, each customer's transaction is stored over time. The timestamp of the data is kept in Universal Time Coordinates (UTC), the global time zone used to control clocks and times.

The dataset that is used in this research work has three data types: Integer, float, and object. Both the integer and float data types are numerical. While float is a number with a decimal place, an integer represents whole numbers without decimal points. Object datatype can hold any python data, not only strings, but it can also hold a dictionary, list, and combination of mixed data. All the packages and modules that have been used during this project are discussed in the next section.

3.3.1 Python Libraries

Python is a prevalent-programming language, particularly in data science. Numerous libraries are open to the public. In addition, it is being developed and maintained by a larger community. Therefore, if anyone encounters a problem, there is a greater chance of finding a solution [42], [43]. Multiple libraries have been employed in this study to analyze and predict client behaviour based on purchases. Libraries with their applications are mentioned below in Table 3.

Table 3: Detail of python libraries that are used in this project and also mentioning the purpose of utilizing the packages

Library	Usage
Pandas	Data analysis begins with importing the data files, followed by data cleaning, integrating several datasets into one, statistical analysis, and much more.
Numpy	It makes it possible to work with multi-dimensional arrays effectively and is the foundation upon which Pandas, Matplotlib, and Scikit-Learn are based.

Matplotlib	It is a tool that is used for data visualization and plotting.
Seaborn	Visualisation package that is built on top of Matplotlib to generate aesthetically appealing graphs
Plotly	Plotly facilitates various languages and offers a high level of customisation and interactivity.
Datetime	This module supplies classes that are used for accessing and manipulating data and time
Tracemalloc	To determine how much memory a specific code block consumes or the entire application.
Sklearn	It is used to anticipate customer churn because it has established a wide range of machine-learning algorithms, pre-processing techniques, performance indicators, and many other things.
Imblearn	It is used to solve the issue of class imbalance, under-sampling method is applied

3.3.2 Basic Analysis and General Trends

The dataset comprises seven separate files, each representing a month from October 2019 to April 2020. All the data files are in the comma-separated values (CSV) format, which enables the saving of data in tabular form and is commonly used for storing and manipulating data.

Transactional information is stored in the dataset about all the customers concerning the brand and product; the e-commerce store contains 4081 brands and 139 unique product categories, and each product and customer is provided with a unique identifier for future analysis of the customer or product buying pattern. Therefore, it is an excellent

technique to begin the analysis from general to specific characteristics per the research problem's requirement. For that purpose, data is combined into a single CSV file. As data is in time series, monthly customer transactions, product categories and customer purchase count are extracted from raw data.

Monthly transactional data shows how many customers purchased each month. Predominantly December and February are the top selling months. The product category is also valuable in knowing which customers are most interested and which products need more attention to keep them running. Another important trait for a company is to focus on the number of transactions the customers make through the life cycle before they are no longer interested in buying from the same company.

3.4 Data Preprocessing

After gathering data from multiple sources, it cannot simply be sent to the machine learning program. The first thing that must be done is data preprocessing. It is a critical step in transforming the original dataset into helpful information for further investigation before passing it to machine learning algorithms. It may take up to 80% of the efforts in the whole data mining process [44]. It begins with data cleaning, reduction, transformation and scaling, as shown in Figure 6.

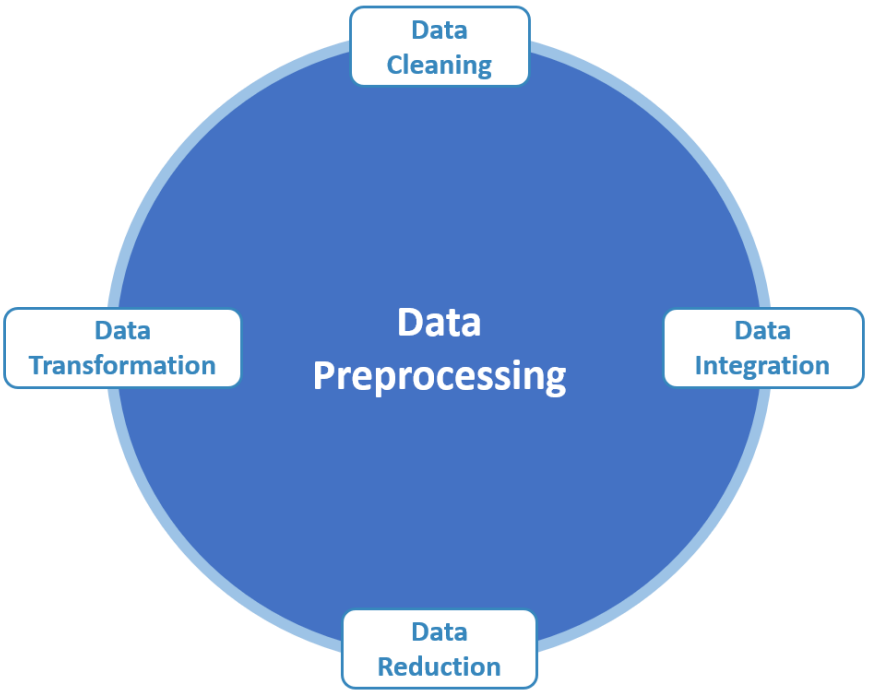


Figure 6: Different phases of data manipulation during data preprocessing; this includes four steps Data Cleaning, Data Integration, Data Reduction, Data Transformation

3.4.1 Data Cleaning

Data may include a variety of irregularities, including duplicate data points, missing values, and incorrect formatting. Data cleaning is a process to address all of these problems. However, when ML models are deployed, data inaccuracies may lead to inaccurate predictions, affecting crucial decision-making systems.

3.4.1.1 Handle Missing Values

Missing values cannot be ignored as they can negatively influence the model results. There are several ways to deal with such data. The straightforward option is to discard such observations (mindfully opt for this choice). Then these empty cells can be filled based on other observations. It has to be done carefully, as data should not lose integrity. Fortunately, the dataset was huge in terms of customers. Only two variables exist, brand and category code, whose values were missing. The rows containing missing values were removed as the dataset contain huge number of customer, so the removal of such rows was not a concern.

3.4.1.2 Duplicate Data Removal

Additionally, duplicate data is present; this data should be removed before the data is divided into training and testing sets. A testing set of fundamental unseen data is provided to evaluate the model's predictive capability. However, if the testing set contains duplicate transactions or has the same attribute values as present in the training dataset. The goal of segregating the dataset into two sets is not achieved.

3.4.1.3 Outliers Detection

Outliers are records of observations that may have data points that are excessively low or incredibly high compared to the majority of the data points. An outlier in one study can be accepted, while it may not be compromised in another. For example, every instance is vital for health and medicine-related data, while in marketing data, such observation can be disregarded.

3.4.2 Data Reduction

The amount of data generated by web application and various sensors are growing exponentially, which can strain the machine learning model. Most of the time, the operation needed to be performed is on targeted data; instead of applying to the actual data. Hence, this complicates the procedure and extends the time needed to produce the

desired outcomes. However, the irrelevant and extraneous data it includes can prevent the algorithms from acquiring the correct information [45].

3.4.3 Data Integration

With the passage of time and advancement in technology, there are now multiple sources through which data is generated. Data needs to be combined in a single file for analysis. With the help of high computational capabilities, it becomes easier to analyze large amounts of data. However, this puts more responsibility on the analyst to utilize and understand it well.

3.4.4 Feature Engineering

Feature engineering or feature discovery is a process in which new variables are retrieved from the source data. The objective is to accelerate and simplify data transformation to increase the machine learning algorithm performance.

3.4.4.1 Timestamp Extension

Time is a good measure, especially when dealing with transactional data. It helps extract information regarding the specific period involved in purchasing (e.g. cultural events, holidays, and natural disasters). For more detailed insight regarding customer purchases, new variables like the day of the month customer purchased something, similarly weekly, monthly and yearly information is obtained.

3.4.4.2 Average Purchase Duration

The study's main emphasis is on recurrent or recurring customers. Each consumer might have unique buying patterns. As a result, just one number is required for each consumer to represent their total number of purchases. For that reason, each customer's average duration between purchases is monitored and recorded in a new data frame column.

3.4.4.3 Customer Frequency

Customer frequency is the number of orders placed by each customer over a specific time period or the number of visits to the store where a purchase was made. This feature is crucial for e-commerce to track consumer buying trends.

3.4.4.4 Customer Monetary

The amount of money a customer has already spent is recorded in monetary value. It is the overall value that each customer brings to the business. The possibility that someone who is spending more money now will continue to do so in the future as well.

3.4.4.5 Customer Recency

It can be defined as when the customer made the last or most recent transaction. The benefit of customer recency is that if the customer has purchased recently, then there is a probability that the customer will buy in the future. Most often, it is stored in terms of days. Other than days (depending on the type of goods), it can be measured in hours, weeks or even years.

3.4.5 Correlation Analysis

After defining the variables, statistical methods are employed to determine the relationship between them and how they are related. Said, it may determine how much a change in one variable results from a change in another. For example, quality and price are two variables. If by increasing one variable, another variable also increases, then it is a positive relationship, And if by increasing one variable, the other variable decreases, then it is a negative relationship. There exist two kinds of variables, independent and dependent. Having two independent variables with strong relationships does not help in selecting the variables. The relationship between the dependent and independent variables is more important.

3.4.6 Churn Defining Criteria

The dataset that is acquired for this research work is not labelled. This means the information regarding its customer retention or customer status is missing. This study defines three distinct ways for considering customers as churn or non-churn, which are as follows:

- Average Purchase Time
- RFM analysis
- K-mean Clustering

3.4.6.1 Average Purchase Time

Calculating the time between a customer's purchase orders serves as the initial step in defining customer labels. A repeat client must make at least two purchases over a period of days in order to qualify. When determining whether a customer churns or not, the average time it takes for a customer to make a purchase is first compared to the sum of all the customer's average purchase times.

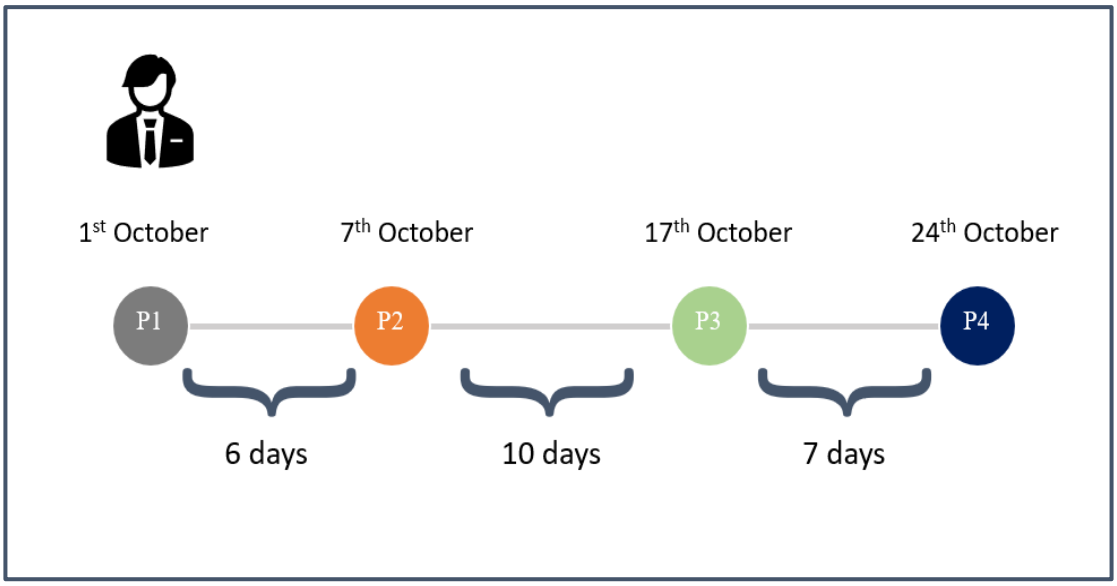


Figure 7: Labelling customers using the average purchase time method; Customer may have several transactions, and the time taken for another purchase may vary; the average time is taken as a measure to have concise information for every customer's time of coming back

With figure 7, the first method for defining customer status is by average time. This can be explained with the help of an example, assuming a customer who has made four orders from the company. P1, P2, P3 and P4 are the days customers buy products. Then there are days when they do not make any purchases. Like there is a difference of six days between the first and second purchase.

Similarly, there is a gap of ten days between the second and third purchases. Lastly, a seven days gap between the third and fourth purchase. The average time for the next order is approximately eight days.

3.4.6.2 RFM Analysis

RFM analysis is a method that is used by marketing officials to rank customers quantifiably and then group them according to their behaviour. Three parameters are involved in this process: recency, frequency and monetary, which are explained in the previous chapter. Concisely, recency is the last purchase time; frequency is the number of purchases, and monetary is the total spending amount of each customer. The flow of this approach is presented graphically in below Figure 8. This technique calculates a score for each of these factors from one to three separately, with three being the highest and then finally adding them all together. Lastly, customers are grouped into two classes based on their final combined score.

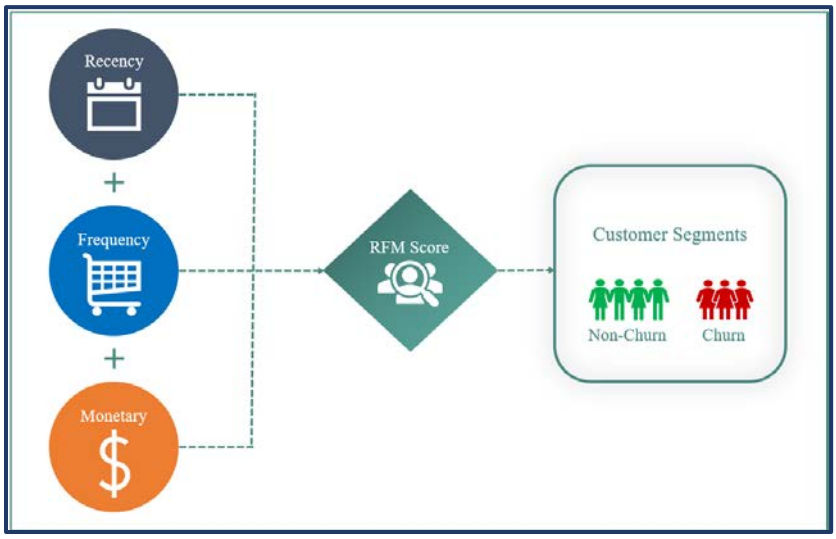


Figure 8: The RFM technique is applied to label the customer sales data; three different variables are engineered from the original data on the basis of which their scores are calculated to segment the customers into different classes

3.4.6.3 K-mean Clustering

K-mean is an unsupervised algorithm that groups customers based on similarity in the feature set. In our case, two types of customers needed to be clustered together, so a value of 2 is set for k, representing the number of clusters to be formed. Then the model will select two data points randomly as centroids and finds the distance for every other data point with these centroids. Customer data values are compared with the centroid values and are grouped with the centroid having a smaller distance. Once all the data points are grouped within their closest cluster. The centroids are recomputed, and again customers will group into newly formed clusters. The data points are reassigned based on the distance parameter. This process is repeated until the newly assigned centroid is the same as the previous one or a limit number of iterations has been reached.

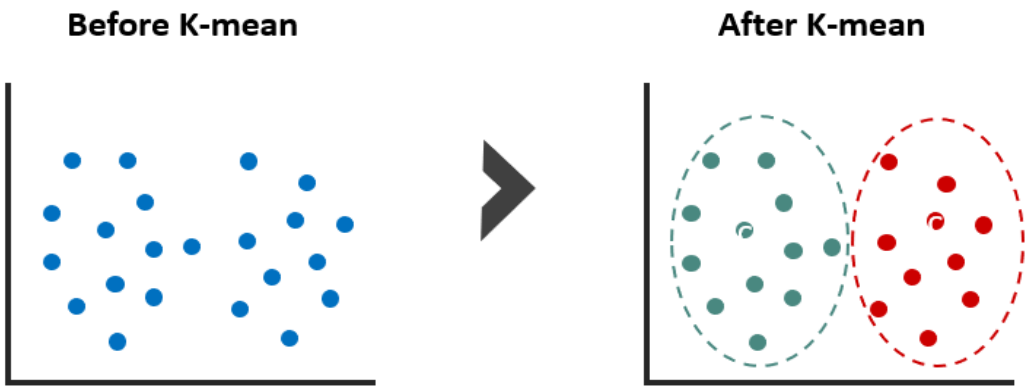


Figure 9: K-means clustering, an unsupervised learning algorithm, is used to label customer data in order to obtain target variables that were not initially included in the dataset.

Figure 9 above describes the left plot containing all the data points when clustering is not administered. The plot on the right shows two clusters after K-mean is employed.

3.5 Churn Modelling

Predictive modelling is the second primary step in our research after defining the customer. At each customer level, it suggests the number of chances that individual customers will be leaving in the future based on previous customer purchases. Three machine learning models are utilized for the prediction of churn customers, namely support vector machine, random forest, and extreme gradient boosting. Each algorithm will be explained later in detail.

3.5.1 Support Vector Machine

Support Vector Machine or SVM belongs to the family of supervised learning algorithms in which the dataset is labelled for algorithm training to predict the outcome or classify data points. The basic concept is to create a hyperplane that splits the data into various classes depending on the target variable.

A Hyperplane is an approximation or simply a separating line between data points, as in below Figure 10. To get the best separator, it finds the data points closer to the separator from both sides, which are called support vectors. Then the next step is to find the distance between the support vectors and the separator line, which is referred to as the margin. So, the best hyperplane will be one which has maximum margins.

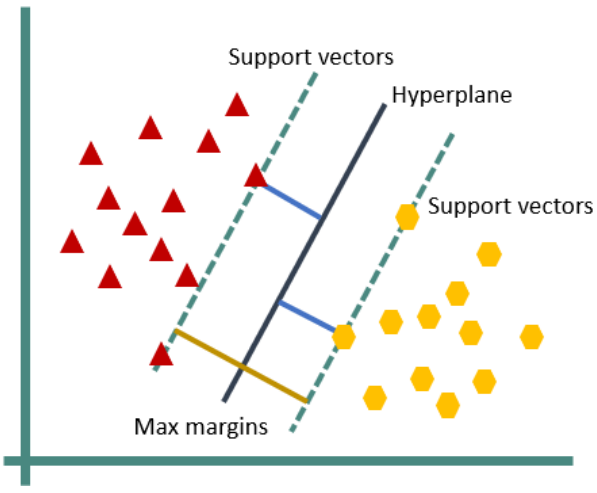


Figure 10: Classification of data using the Support Vector Machine; in the diagram, the hyperplane represents the decision boundary between the classes; support vectors are the points that are closest to the margins, which need to be maximum on either side of the Hyperplane for more accurate classification

3.5.2 Random Forest:

Accurate group data is precious for business applications, for example, to predict whether a customer will buy a product. Other examples are fraud detection, product categorization and disease prediction. Random forest is the advanced decision tree version consisting of two elements, nodes and branches. At each node, features are evaluated to find the best data splitting.

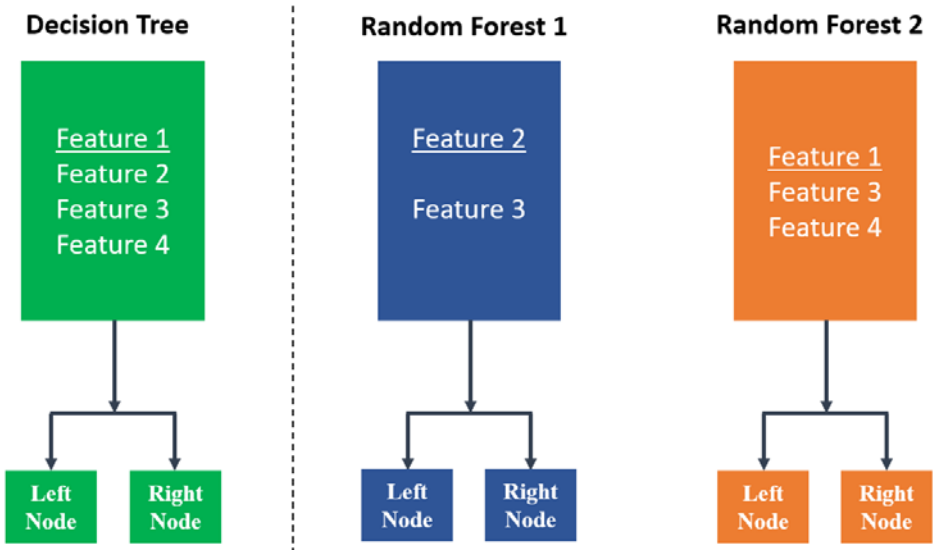


Figure 11: Visualization of Decision Tree with Random Forest classifier; a single tree is constructed in a Decision Tree model where all the features are used depending on the feature importance; in Random Forest, several trees are generated with different features selected at random

This process continues until it reaches the leaf node that holds a class label. Multiple decision trees are created in the Random Forest algorithm and work as a committee. Each member of the committee or decision tree predicts the class label, and the one having the majority of votes becomes the model's prediction. Instead of relying on a single tree, Figure 11 demonstrates how random forest makes the final prediction based on multiple trees using the majority voting method.

3.5.3 Extreme Gradient Boosting:

It is a part of ensemble learning. For better predictive performance, multiple models are combined, referred to as ensemble learning. Bagging and Boosting are the two types of ensemble learning. Xg-boost lies under the category of boosting.

Like Random Forest, Xg-boost is constructed from the Decision Trees. However, in the case of random forest, multiple trees are constructed independently and then combined with the result by using the averaging or voting method. On the contrary, extreme

gradient boosting also builds multiple trees, but one after the other, by correcting the errors of the previous model, which is the critical property of the algorithm.

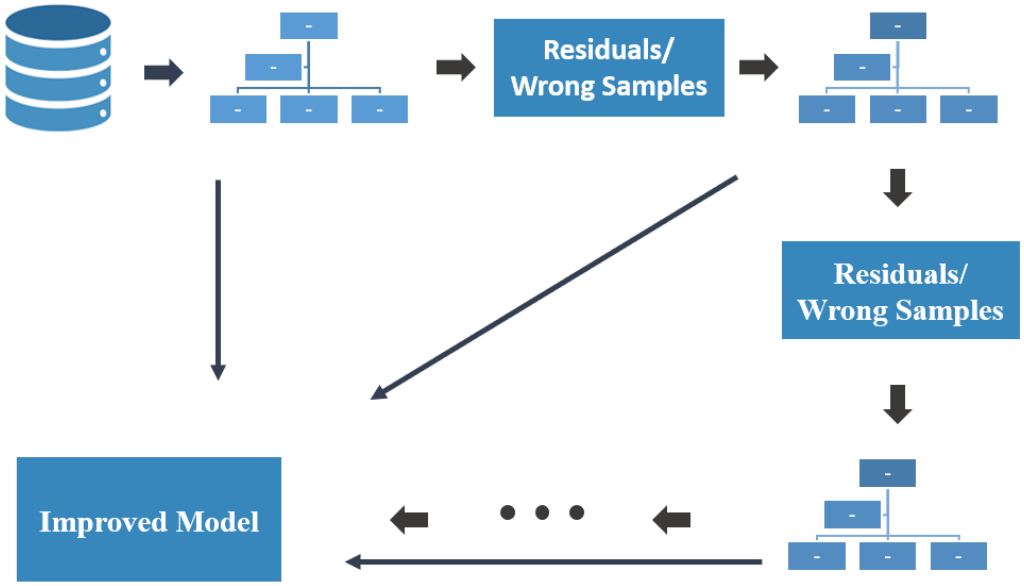


Figure 12: Workflow of the Extreme Gradient Boosting classifier; each tree is generated, and the wrong samples are trained repeatedly until model performance is improved

The workflow of the Xg-boost algorithm is demonstrated in Figure 12 above. The stopping criteria for this algorithm are either the maximum depth has been reached or the additional split does not impact the model's accuracy.

3.6 Model Evaluation

After applying the machine learning models to the data, different evaluation measures can be used to check how accurate or correct the model is. The problem under consideration for this research belongs to the supervised learning category, classification. The measures used in this methodology are accuracy, precision, recall, F1-score and AUC.

3.6.1 Accuracy

Accuracy is a metric that is widely used in industries. It can be calculated as the ratio of the correctly predicted observation to the total number of predictions made by the model. It summarizes the performance of the classification problem. Accuracy should be used when the dataset is genuinely balanced, meaning data has an equal proportion of the class labels.

$$Accuracy = \frac{No\ of\ correct\ prediction}{No\ of\ total\ prediction}$$

3.6.2 Confusion Matrix

A Confusion matrix is a powerful technique that is used for classification problems. It not only considers the predictions that are right but also predictions that are misclassified. It makes it easy to see if the model is getting confused with target class labels. Further explanation is provided with the help of Figure 13.

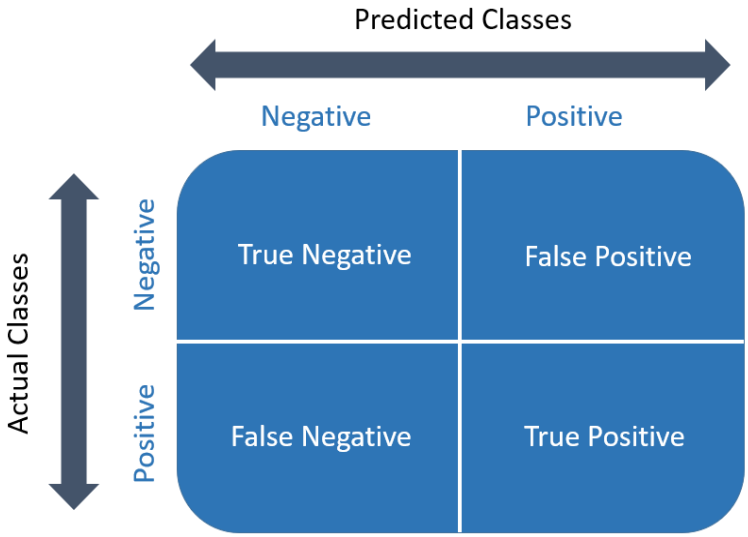


Figure 13: The confusion matrix is used to depict algorithm performance; negative and positive class labels are written on both the top and left sides of the matrix, the label on the top is referred to as the predicted label, and the label on the left is referred to as the actual class label

An example of disease prediction, there are two possible outcomes, either yes or no. When a patient is predicted to have a disease, and it has the disease, it is referred to as a true positive. A false positive is when a patient is predicted to have a disease and does not have it. Then there is a false negative, those observations which are predicted to have no disease, and in reality, they have it. Finally, the last section of the confusion matrix is a true negative, in which the model has anticipated the patient has no disease, and it has no disease.

3.6.2.1 Precision

It is defined as the total optimistic predictions, how much are correctly predicted as positive. Precision should be used when false positives are essential. One does not want

them or as want few as possible. It checks the quality of the prediction, which the predictor claims to be positive.

For example, if two classes exist of humans and animals, the goal is to place humans in a safe place (positive class). The cost of false positives is higher, placing animals in the safe zone. Here precision is more valuable.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

3.6.2.2 Recall

The proportion of observations is predicted as positive, and they are actually positive. It is more valuable when false negatives are crucial. It measures the quality based on the mistakes our model made. In the case of disease prediction, the cost of a false negative predicts a patient with no disease, which is deadly for the patient as the disease worsens. So, in that case, more focus would be on recall.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

In real-world problems, most of the time, one is either interested in precision or recall, depending on the cost or damage caused by the false positive or false negative.

3.6.2.3 F1-score

F1-score is the harmonic mean of both the precision and recall metric. After understanding the importance of the above terms, it has been realized that a tradeoff exists between precision and recall. When both are equally important, the f1-score measure is used. If one has a low value, then the resultant f1 score has a low value.

$$F1_{score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

3.6.2.4 AUROC Curve

AUROC stands for the area under the receiver operator characteristic. It is a graph that evaluates the performance of the classification model. The graph is plotted between two

values, true positive rate and false positive rate. The ratio of positive class examples that are predicted correctly is referred to as a true positive rate or recall.

On the other hand, the ratios of negative class examples are incorrectly predicted. It is an easy way to summarize the model's overall performance. A model with a higher AUC score is the best. The figure that shows the AUROC curve with two thresholds is presented below:

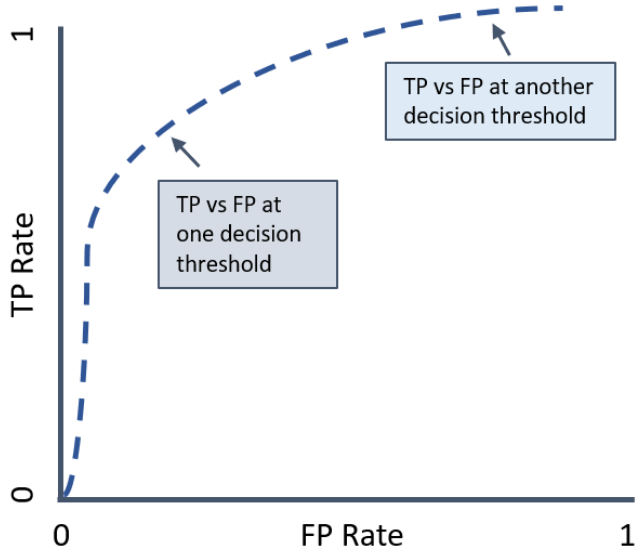


Figure 14: The Area under the Receiver Operating Characteristic curve with various thresholds is displayed in the figure; a graph is plotted between true positive and true negative rates; the greater the area is under the ROC curve illustrates an increase in the performance

Chapter 4

Result and Discussion

This chapter presents the results of the methodology approach used to address the research subject of customer churn prediction. A thorough discussion and personal insight of the researcher is also extended behind the steps taken for the analysis of customer purchase pattern. The primary objective of this project is to define or label the customers in two groups and apply a comparative analysis of the machine learning algorithm to predict churn. This work uses three machine learning models and three alternative ways for categorizing data as the dataset is not labelled initially. To find the optimal solution that is not only good in terms of accurate results but also that is less time-consuming and memory-efficient, the algorithm's time and memory requirements are also taken into account.

4.1 Customer Purchase Inclination

The data used for this research problem comprises customer transactional records with their timestamp. For a data science specialist, the foremost thing after acquiring data is to have a general understanding of customers' existing data, like when and what they are purchasing and the effect of external factors, before moving forward to the more specific or desired information.

4.1.1 Monthly Customer Transactions

Any business depends on its consumers; thus, it is critical to understand the clients who are purchasing. Some customers just make one purchase, while others do so repeatedly. Prior to becoming more precise, it is crucial to have a general purchase pattern based on months.

E-commerce is multiplying yearly, and COVID-19 has forced businesses and corporations to establish an online presence [46]. It is evident from Figure 15 that there is an upward trend in sales. The company’s two most successful sales months were December 2019 and February 2020, when covid-19 first began to take off globally.

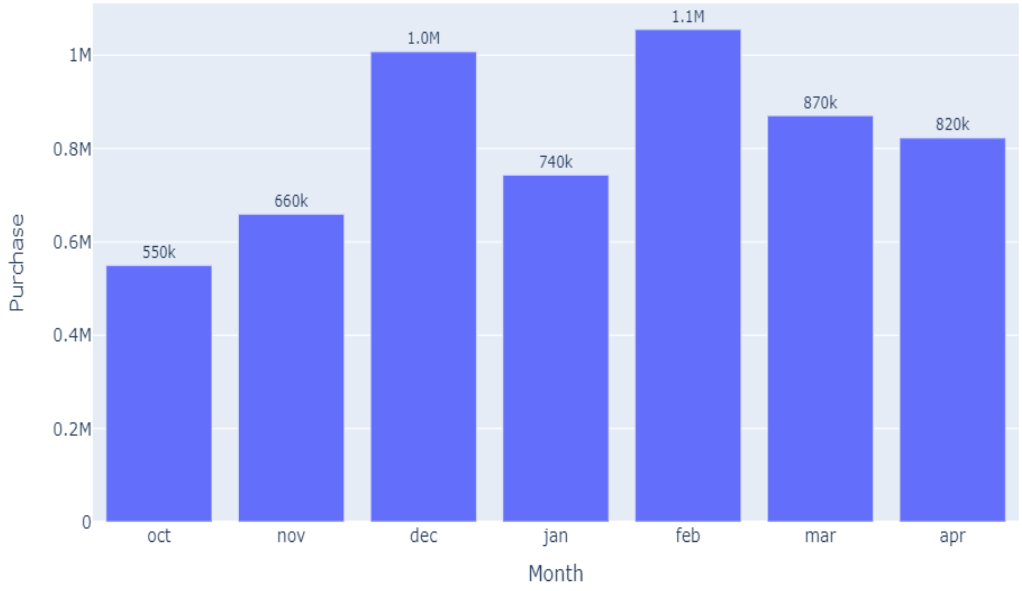


Figure 15: Customer who has made more than one purchase throughout their active period over the series of months; their purchasing frequency is shown in the bar plot

4.1.2 Product categories

The dataset contains a wide range of products; in total, it has 96037 unique products. There is a hierarchy of categories. For example (construction - tools - light), the primary category is construction, which is further broken into subcategories, and the final product is light.

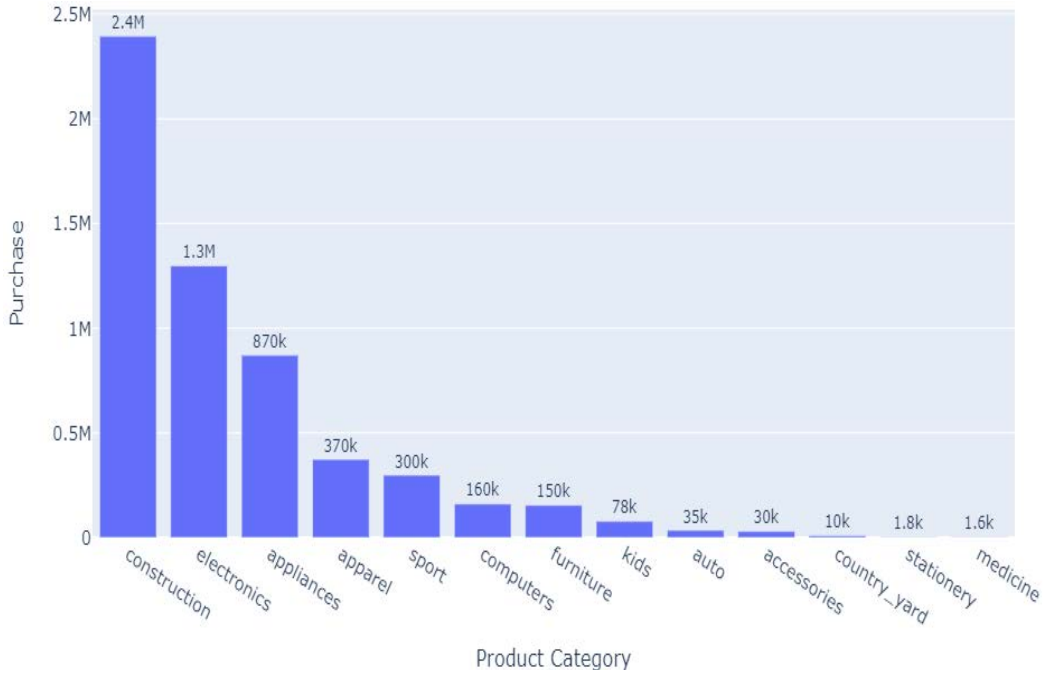


Figure 16: The plot displays client purchases by product category type, allowing the firm to discover which product categories are popular and which product types have the lowest sales

There are 13 broader product categories in the data. How many customers purchased each sort of product category is shown above (Figure 16). The benefit of having this information is that businesses can determine which product categories are in high demand and which require more attention to make them successful.

4.1.3 Customer Purchase Frequency

The companies that are successful and produce a lot of revenue mostly have recurring customer bases. The earliest strategy to grow businesses in e-commerce was to acquire new customers and spend handsome amounts of money on advertising and marketing. However, with time, experts analyzed that acquiring new is much more costly than retaining existing customers [47], [48]. So, to make customers stay with the company for a prolonged time, it is essential to retain existing customers before they leave. According to McIlroy and Barnett [3], 80% of the company revenue comes from 20% of the customers, which tells how important each customer is for the business. In terms of the analysis, the first step is identifying repeat customers. It is defined as customers who made at least two purchases.

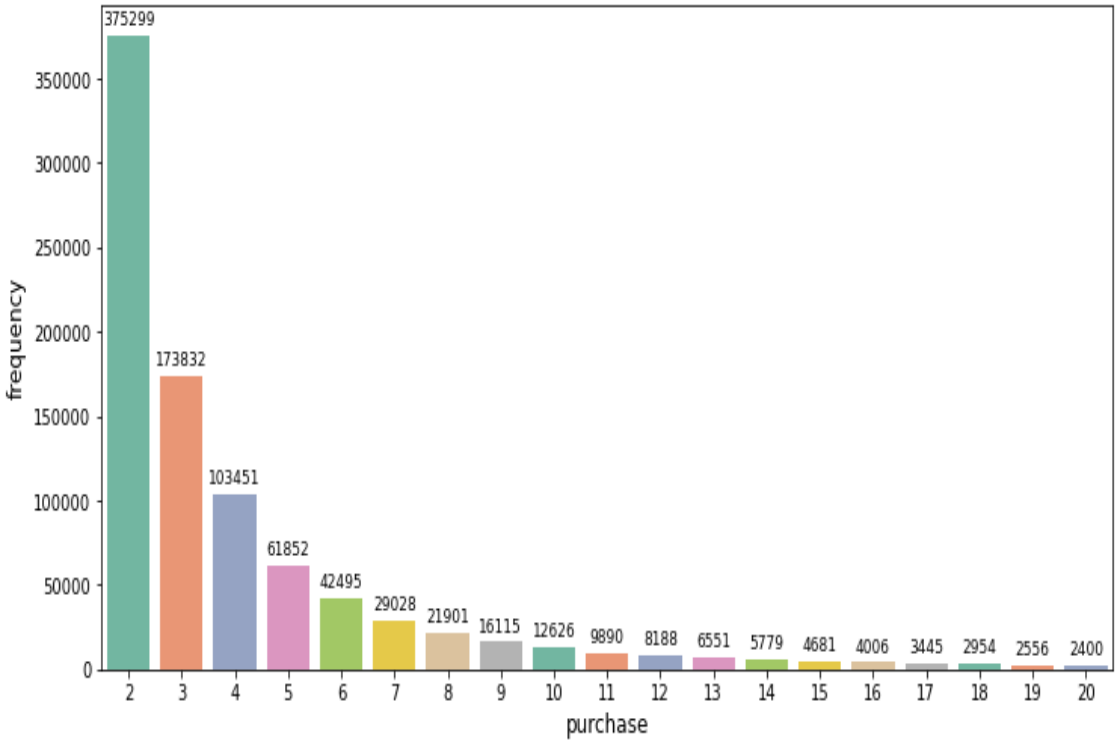
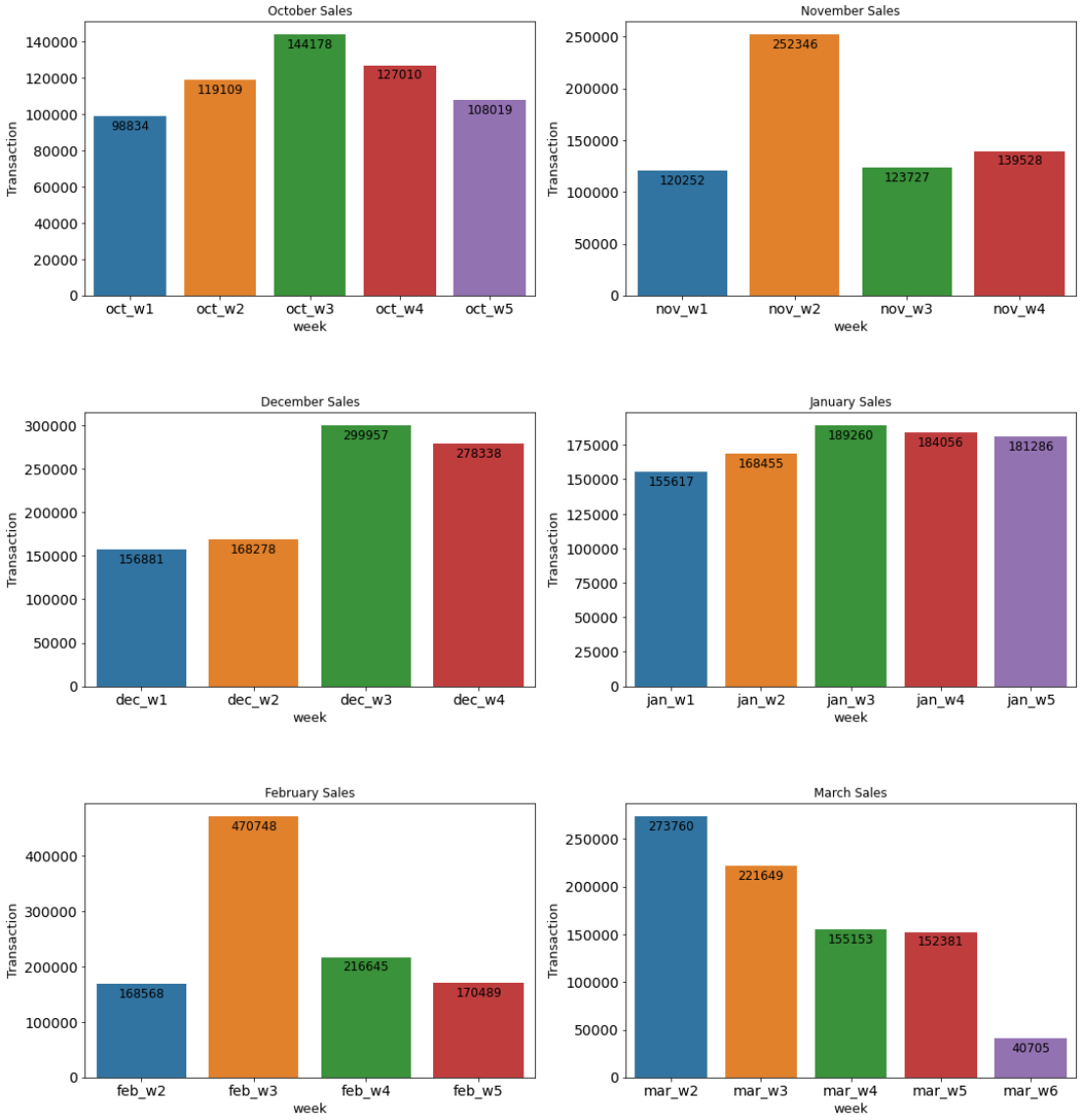


Figure 17: Returning customers and their buying volume; clients with multiple sales on various days over a period of months are represented in the bar graph

In Figure 17 above, the x-axis has the purchase count ranging from 2-20, and the y-axis represents the number of customers in each group. For example, 42495 customers made six purchases. Though there are customers who made more than 20 purchases but purchase count is fixed to 20 to make it clear and visually understandable. The buying trend is generally the same as it appeared in Figure 17.

4.1.4 Weekly Transactional Record

Weekly analysis of client records is another approach to looking at the data. The marketing team can benefit significantly from such information. It demonstrates the time or week of the month when the maximum number of customers are interested in purchase. Figure 18 illustrates the graphs for each month; the trends show that weekly, the second and third weeks of the month are more active in sales.



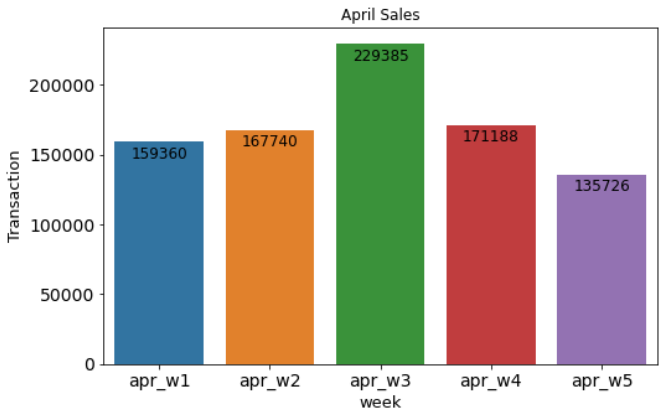


Figure 18: Summarizing weekly purchase data for seven months to identify the flow of customers at the beginning, middle and end of the month

4.2 Data Discovery and Preparation

For machine learning projects, most of the time is spent preprocessing the data. Because the integrity of the data set is compromised by errors, duplicates, missing values, and inconsistencies, one must address these problems for a more accurate result. For example, consider using a flawed dataset to train a machine learning system to handle customer purchases. The system's likelihood of exhibiting biases and aberrations that negatively impact the user experience is high.

4.2.1 Missing Values

The data contains two columns with null values: category code and brand. Both these features are related to products. So, this information cannot be imputed because this will land the product in another category and brand. The good thing is that the data has a substantial transactional record. So, it does not matter if some of the records are removed.

4.2.2 Data Redundancy

Each customer’s data is grouped to have more generalized information based on the feature set, which results in some duplicate information. For instance, two or more customers may have the same number of purchase count, money spent and last purchase time. So, when data is divided into training and testing groups, and both sets have some records that are the same, then splitting the data is not accomplished. Ideally, the data in the test set should be unseen to examine the model capability on the new data. Such kind of rows has been removed from the dataset.

4.2.3 Outlier Removal

There exist observations or records that are far dissimilar from the rest of the customers; they are called outliers. They can increase the variation in the dataset, which decreases the statistical power. In this scenario, less than 1% of clients with a purchasing history of more than 46 purchases. Similarly, the same ratio of customers exists that spend more than 20,218. The result becomes more significant by excluding outliers.

4.2.4 Integrate Datasets

Our dataset contains seven files separated based on months from October 2019 to April 2020. The data is merged into a single file. Because at times, instead of being specific, it is essential to integrate data to have a unified view which is then analyzed to make data-driven decisions for the business.

4.2.5 Feature Extraction

Feature engineering or feature extraction is a way in which researchers use domain knowledge to extract and transform the raw data into new attributes that will help in predictive modelling. Frequency, Monetary or spend money, recency, average purchase time, maximum purchase time and total duration. Table 4 describes the necessary information regarding the customers.

Table 4: Key information about the customer purchase pattern that is engineered during the pre-processing phase, together with their statistics such as mean, standard deviation, lowest value, maximum value and percentile

	Frequency	spend money	recency	average duration	max duration	total duration
mean	6.1648	2127.165	74.0696	20.8019	37.5052	52.0028
Std	12.068	5978.086	54.5063	28.4845	38.1072	50.3513
Min	2.0000	1.17000	1.0000	0.0000	1.0000	1.0000
25%	2.0000	401.430	27.0000	3.0000	7.0000	9.0000
50%	3.0000	866.280	62.0000	10.000	25.000	35.000
75%	6.0000	2008.092	116.000	27.000	57.0000	84.000
90%	11.000	4463.685	154.000	55.000	93.0000	130.000
95%	18.000	7422.106	176.000	80.000	117.000	154.000
99%	46.000	20218.680	201.000	140.000	159.000	189.000
max	1975.00	790098.290	212.000	211.000	212.000	212.000

This explains the key features in detail, starting with the mean average value and standard deviation, which shows the variation in the data values. It also presents the minimum and maximum numerical values in the respected feature column. Then there are percentiles defined as the portion of scores lower than a given threshold. For instance, in the frequency column, the second last row of the table shows that 99% of the customers are buying 46 or fewer products.

4.2.6 Repeat Customer Sample

In supervised learning based on historical data, the model learns the functional relationship between the features based on which the model performs prediction. In this study, the customers must have completed some prior transactions to examine their behavioural patterns. Therefore, from the overall data, those customers who have made at least two purchases over several days are extracted for further research. There is a total of 686548 recurring customers.

4.2.7 Correlation Analysis

The feature and information that was initially not present in the dataset, before passing it to the churn model, their relationship with each other is analyzed to get insight. Figure 19 presents the correlation matrix as a result of feature engineering.

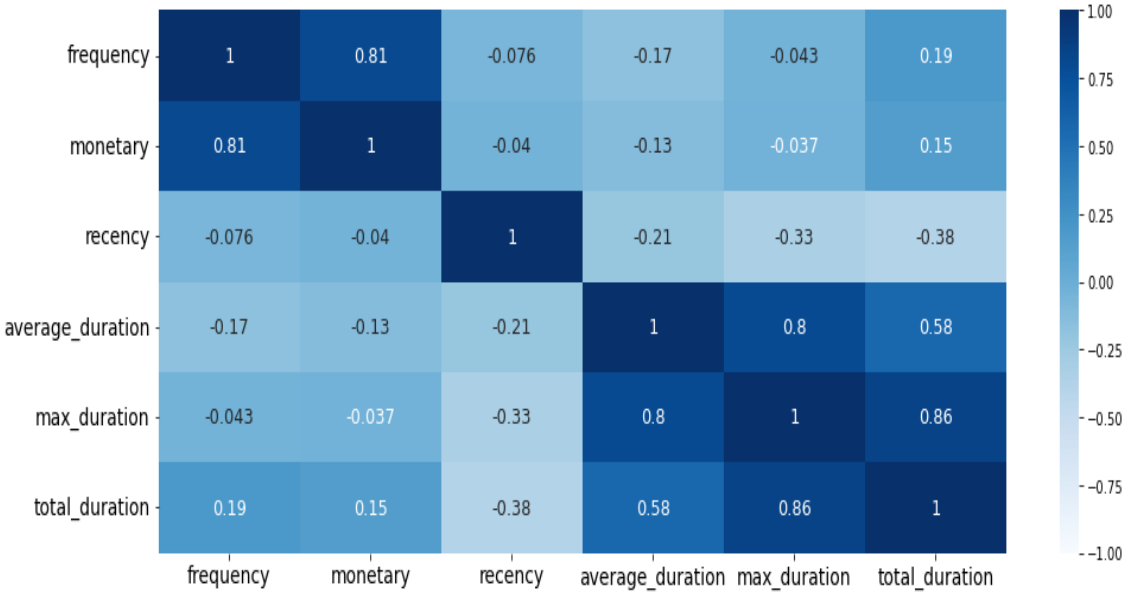


Figure 19: Correlation matrix of the important features for churn modelling; allows to see the relationship among the variables; change in one variable may have an impact on the other variable that can be positive, negative or have no relation

4.2.8 Cohort Assessment

Cohort analysis is the process of analyzing customers over time to look at how their behaviour evolves. In the study, customer churn is the behaviour of interest. Few customers make transactions every month. These customers are the top priority customers. The total count for such customers is 1540. Table 5 demonstrates how many customers stopped buying products in terms of percentage based on the month. The first cell of every month in which customers come has a 0% churn rate, gradually repeat count of customers decreases and the churn rate increases. For example, the second cell of Oct 19 shows that 77% of the customers who joined on October 19 did not return the next month, i.e. November 19.

Table 5: Examining customer churn behaviour by obtaining total time duration; calculating the proportion of consumers who stopped purchasing items based on the month

Cohort	0	1	2	3	4	5	6
Oct, 19	0%	77%	90%	95%	97%	98%	99%
Nov, 19		0%	74%	91%	95%	97%	99%
Dec, 19			0%	80%	92%	96%	99%
Jan, 20				0%	75%	90%	97%
Feb, 20					0%	75%	93%
Mar, 20						0%	82%
Apr, 20							0%

4.3 Churn Target Variable

The dataset this study begins with was not labelled. Therefore, the target variable information is lacking (whether a customer is a churn or non-churn). After going through the literature, three different methods are defined for dividing customer data into binary classification problems.

4.3.1 Average Time Interpretation

The first method used for labelling customers is on the basis of average purchase duration in which two values are computed, one is the overall average purchase time of all the customers, and the second average purchase duration of each customer is calculated. The average purchase duration of all the customers is 21 days, which is a

benchmark to consider whether a customer is lost or still exists. If a customer has an average purchase time less than the overall average time of all the customers, that is

Twenty-one days, then it is considered as non-churn or loyal customers; otherwise, it is considered as churn or lost clients. After labelling customers, we obtained non-churn and churn percentages of 62% and 38%, respectively, as shown in Figure 20.

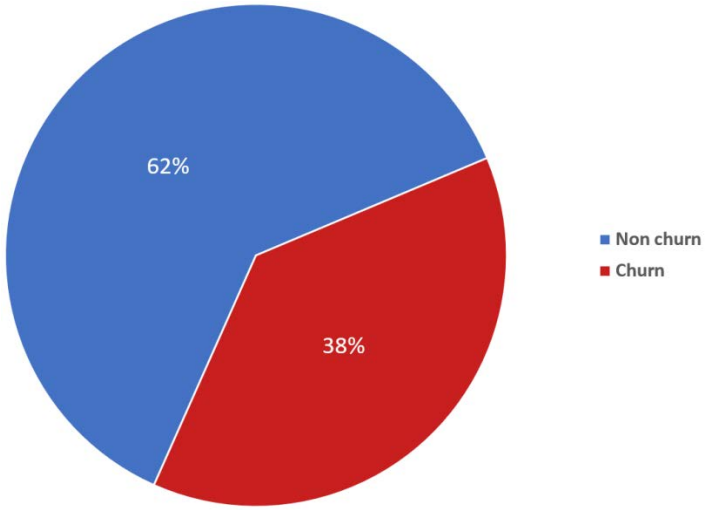


Figure 20: Percentage of customers when data was labelled using the average time methodology; the majority of the consumers are considered non-churn, with around two-fifths portion of the customers, are considered churn

4.3.2 RFM Estimation

Recency, Frequency, and Monetary, or RFM, is a marketing technique used to rank or categorize customers efficiently. First, these RFM features have to be computed from the dataset. The following step is to specify the ranges between the values and, separately, to compute the scores as in Table 6. The best or most excellent score for each of the abovementioned variables is three on a scale of 1 to 3.

Table 6: Estimation of Recency, Frequency, and Monetary attributes, as well as their ranges; each feature has three intervals which are formed after analysing customer buying patterns

	1	2	3
Frequency	[1 - 5]	[6 - 100]	[101 - 1975]
Recency	[142 - 212]	[71 - 141]	[1 - 70]
Monetary	[1.169 - 521]	[522 - 1485]	[1486 - 790098]

The frequency numbers range from 1 to 1977 and represent the total number of purchases made by each client. There are three intervals: (1 to 5), (6 to 100), and (101 to 1975). The data values for the frequency feature are not normally distributed, so these

ranges are customized. Here, the data is divided into three groups. As 70% of the consumer's purchase count is below or equal to five, these customers are considered low-value customers and will receive a score of 1. Similarly, 29% of customers with a purchase count in the range of 6–100 will receive a score of 2, and customers with more than 100 transactions will get a score of 3.

The last time a transaction was made by the customer is known as recency, which is described in terms of days. The intervals for recency are (1–70), (71–141), and (142–212). The maximum value for recency is 212 days, which is simply divided into three groups with a 70 day interval for each group. Customers with scores ranging from 1 to 70 are given a maximum of 3, and customers with scores ranging from 71 to 141 are considered average performers. Finally, consumers whose last purchase time is above 141 days are at risk of churn, and such individuals are given the score of 1.

The entire amount of money the customer spends is stored in a variable called "monetary". Additionally, the ranges are 1.169–521, 5.22–1485, and 14.86–790098. The record for the monetary feature is not normally distributed. So, the quantile discretization function is applied, which has divided the data into three equal sized bins. The highest score is given to customers who spend more money. In Table 7, a few data values are taken from the dataset, and their scores are calculated based on the ranges mentioned above for each extracted feature.

Table 7: RFM methodology scores are obtained by adding individual Recency, Frequency, and Monetary scores, which are used to specify target variables into two classes

Frequency	Recency	Monetary	F_score	R_score	M_score	Total score
50	10	200	2	3	1	6
106	83	3453	3	2	2	7
3	150	50	1	1	1	3
120	18	10000	3	3	3	9

After calculating the total score in the range of 3 to 9, the customers whose total score is in the range of 5 to 9 are considered as non-churn and customers whose score is 4 or 3 are labelled as churn. According to the scoring system, a customer is regarded to be non-churn if they have at least a highest score of 3 in one of the three features or must perform moderately in two features, which is scoring 2 in two parameters out of the

three features. The remaining customers are classified as churn. The number of customers classified as non-churn or existing customer is 59%, and churn is 41%.

4.3.3 K-mean Clustering

The last method used for categorising and labelling customers is the k-mean algorithm. As with k-mean clustering, the algorithm should be mentioned in advance how many clusters it needs to produce. It is a distance-based algorithm. Data is labelled based on the distance from the centroid. As the value of k is 2, two random centroids are formed, and their distance is computed against each customer. This process continues until the maximum iteration is completed or the centroid does not change its position. Some of the scatter plots of customers are presented below after labelling the customer data into churn and non-churn as shown in the Figure 21.

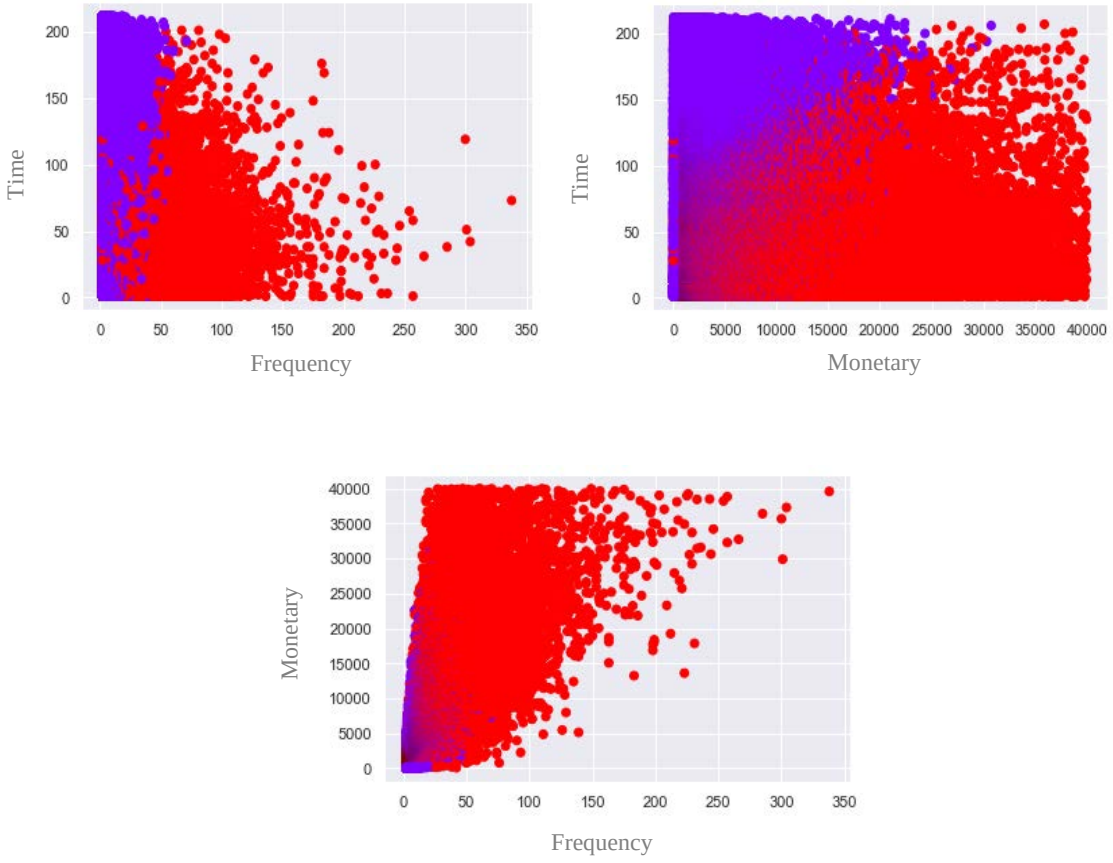


Figure 21: Visualization of scatterplot after applying K-means clustering algorithm to label the consumer data; to figure out the active and non-active customers

4.4 Model Selection

Three models, namely Random Forest, Support Vector Machine and Extreme Gradient Boosting algorithms, are selected to predict customers who are at risk of being lost in the future. Each one has its importance, depending on the type of data set.

The Support Vector Machine is not affected if the data is distributed regularly. Hyperplanes and margins are used to best fit the data and classify more. The aim is to have a maximum margin between the two classes to help test the data for prediction. One of the other benefits of using SVM is that it can work on nonlinear problems as well. Also, outliers have less influence on the training of the model.

According to the literature, tree-based algorithms work well for the churn prediction problem [49]. Random Forest is based on Ensemble Learning, which is why it is not dependent on the output of a single model. Multiple independent decision trees are formed by random selection of the features. Then based on majority voting, the final results are aggregated. It will allow for checking multiple variable combinations.

Then there is another category of ensemble learning that is called boosting. An Extreme gradient boosting algorithm is also a tree-based algorithm in which multiple trees are constructed which are not independent. Trees are generated sequentially, i.e. one after the other, in order to minimize the mistakes or errors that occurred by the previously generated tree, instead of combining the result at the end of the algorithm like in Random forest. Results are aggregated along the way to compute the final outcome.

4.5 Model Result and Evaluation

The machine learning models that will be employed have already been chosen as of right now. This section contains all the findings from the nine models we used to identify at-risk clients. The outcomes are presented in the following order: random forest with its three defining approaches, support vector machine findings, and extreme gradient boosting.

4.5.1 Random Forest using Average Time

In this method target variable is labelled based on a comparison with the average time of the overall customers, A Random Forest is applied, and the following result is generated. The confusion for this method is shown below in figure 22; critical

performance metrics that are precision, recall and f1-score are computed. All these metrics and confusion matrices are defined in the previous chapter.

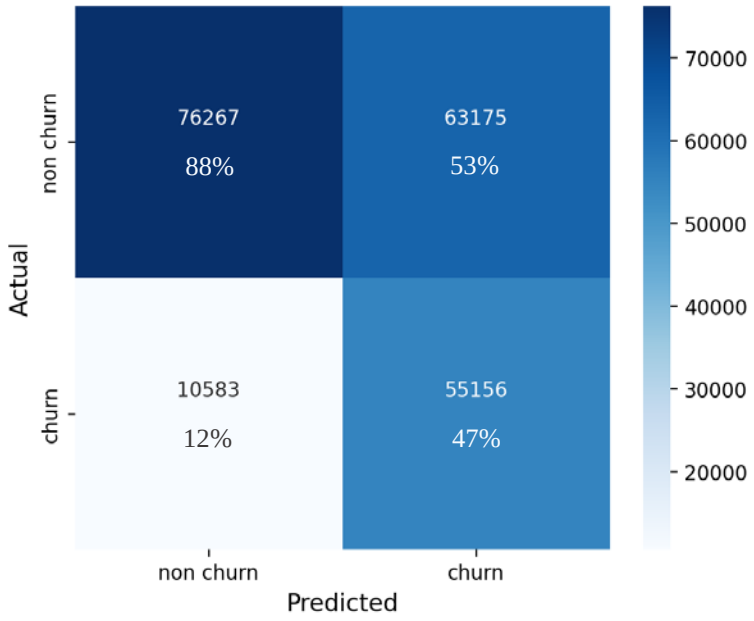


Figure 22: Confusion matrix of the Random Forest using the Average time method to determine the positive and negative class outcomes in detail

The result in Table 8 is presented below, in which it can be seen that the precision of the churn class is not very well. On the other hand, the recall result for the churn class is quite good, but it is very poor for the non-churn class. These percentages ultimately influenced the F1 score. The model managed to get an average performance of 64% by using a single feature average time.

Table 8: Performance metrics of the Random Forest using the Average time are observed using Precision, Recall and their combination that gives the F1 score

	Precision	Recall	F1-score	Support
Non-Churn	0.88	0.55	0.67	139442
Churn	0.47	0.84	0.60	65739
Accuracy			0.64	205181
Macro avg	0.67	0.69	0.64	205181
Weighted avg	0.75	0.64	0.65	205181

4.5.2 Random Forest using RFM:

The second method for defining the target variable is the Recency, Frequency and Monetary (RFM) method. The score is calculated based on these variables. Below is the Figure showing the confusion matrix for this approach.

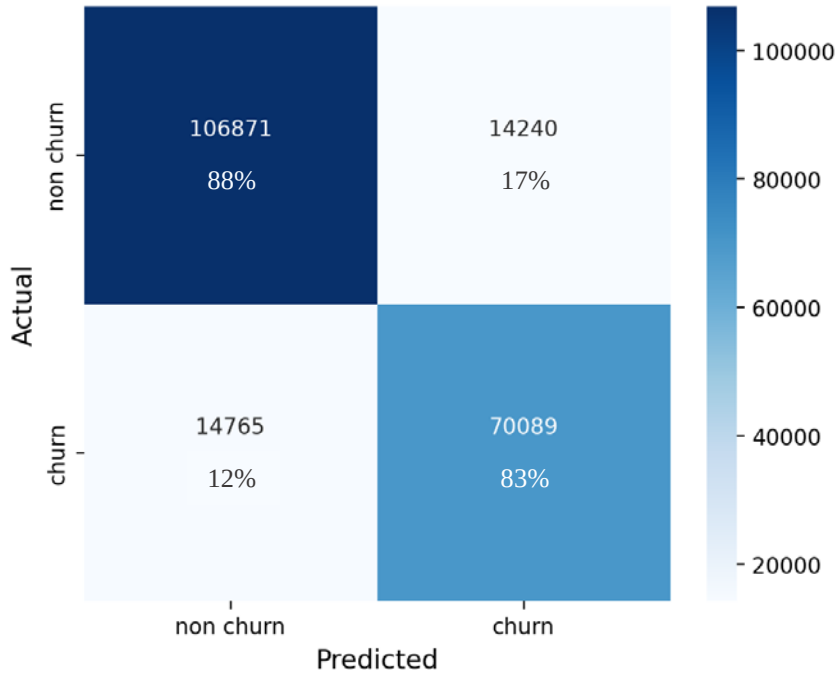


Figure 23: Random forest model misclassification error when data is labelled using the RFM approach is visualized using the Confusion matrix in terms of false negative and false positive

Looking at this approach, we have got a decent number for precision, recall and f1-score and the good thing is that these are all equal to or above 83%, which is visible in the table below. It outperforms the result of the previous method of the Random Forest using average time.

Table 9: The RFM approach is used for random forest assessment; new data is evaluated on the trained model, which provides more than or equal to 83% accurate prediction

	Precision	Recall	F1-score	Support
Non-Churn	0.88	0.88	0.88	121111
Churn	0.83	0.83	0.83	84854
Accuracy			0.86	205965
Macro avg	0.85	0.85	0.85	205965
Weighted avg	0.86	0.86	0.86	205965

4.5.3 Random Forest using K-means Clustering:

The last method used for Random forest is the unsupervised technique called k-mean clustering for the labelling of customers. From Figure 24, it is clear that 77% of the customers are represented in the matrix's bottom-suited cell, which contains true positives or churn, which makes it obvious that the k-mean technique classified more customers as leaving than staying.

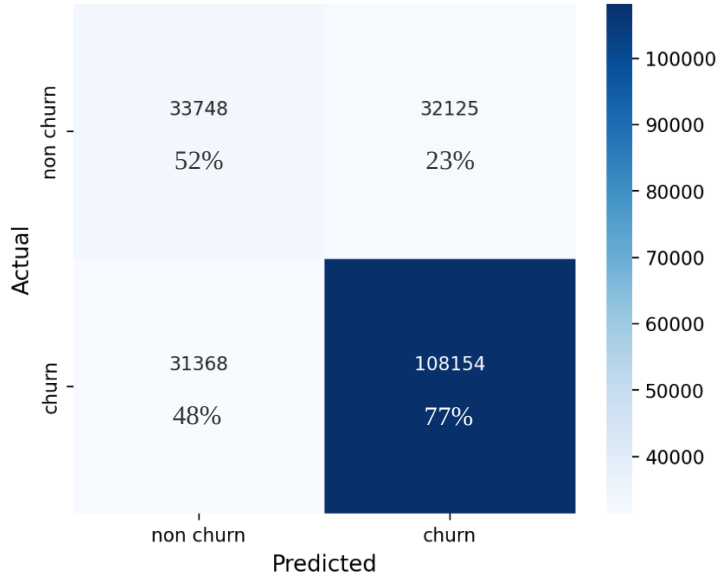


Figure 24: Random forest contingency table utilising K-means clustering with two dimensions of actual and predicted observation to validate model effectiveness

The precision and recall for the churn class are 77% and 78%, respectively, and the results are not all that significant, as indicated in Table 10. The result is around 50% accurate for the non-churn class.

Table 10: Performance evaluation of the Random Forest using the K-means clustering; around half of the customer records for non-churn are not correctly anticipated, and for churn class, the result is also less than 80%

	Precision	Recall	F1-score	Support
Non-Churn	0.52	0.51	0.52	65873
Churn	0.77	0.78	0.77	139522
Accuracy			0.69	205395
Macro avg	0.64	0.64	0.64	205395
Weighted avg	0.69	0.69	0.69	205395

After applying random forest with three different techniques of average time, RFM and k-mean clustering, the result shows clearly with a high margin that the RFM method is better than the two other techniques.

4.5.4 Support Vector Machine using Average Time

The three methods for labelling customers will be employed in a manner similar to that of the Random Forest, but this time, another algorithm called a Support Vector Machine is used. In figure 25 below, a confusion matrix is used to depict true negative, false positive, false negative, and true negative.

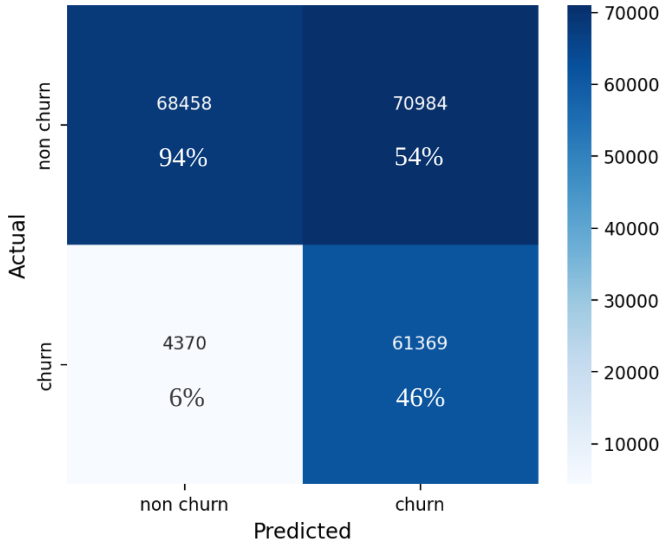


Figure 25: Confusion matrix of the Support Vector Machine using the Average time method to observe the influence of two classes as well as their erroneous prediction

It's advantageous to have a high recall (Table 11), indicating how many of the examples of customers expected to churn are classified as such. However, recall for non-churn is less than 50% in the other situation when comparing projected vs actual observation, even though churn recall is relatively good.

Table 11: Results of the Support Vector Machine using the Average time technique; Precision for the churn class and Recall for the non-churn class are both less than 50%

	Precision	Recall	F1-score	Support
Non-Churn	0.94	0.49	0.65	139442
Churn	0.46	0.93	0.62	65739
Accuracy			0.63	205181
Macro avg	0.70	0.71	0.63	205181
Weighted avg	0.79	0.63	0.64	205181

4.5.5 Support vector machine using RFM

In below Figure 26, the number of false positives and false negatives tells about examples or customers that are misclassified, that is, 17% and 27%, respectively. Out of the three methods used with the Support Vector Machine, when data is labelled using the RFM technique model works better.

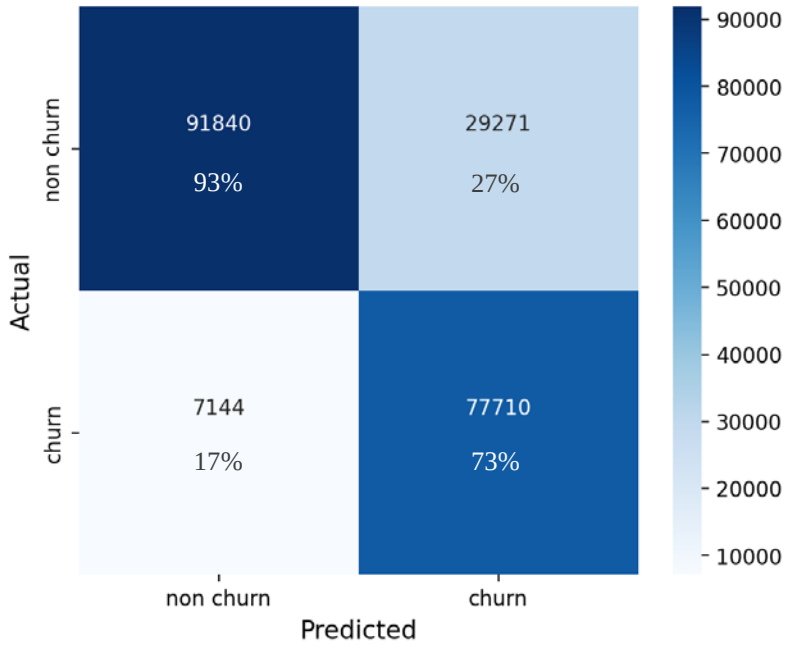


Figure 26: Measuring the performance of the Support Vector Machine algorithm with a contingency table where data is categorized using the RFM scoring methodology

The Recall is above 90% (Table 12), which is the highest so far for the positive class churn. Precision for non-churn is high, but recall is on average. Overall, this strategy produces positive results. Accuracy and f1-score are both above 80%, which is a good, respectable outcome.

Table 12: The results of the Support Vector Machine adopting the RFM approach are sufficiently considerable, with F1 scores for the positive and negative classes above 80% and Recall for the churn class exceeding 90%

	Precision	Recall	F1-score	Support
Non-Churn	0.93	0.76	0.83	121111
Churn	0.73	0.92	0.81	85854
Accuracy			0.82	205965
Macro avg	0.83	0.84	0.82	205965
Weighted avg	0.84	0.82	0.82	205965

4.5.6 Support Vector Machine using K-means Clustering

According to Figure 27, there are about 53% of the non-churn consumers are miscategorized as churn and customers who are incorrectly predicted as non-churn is 7%. In the event of churn, 47% of the observations are forecast as churning or at-risk clients.

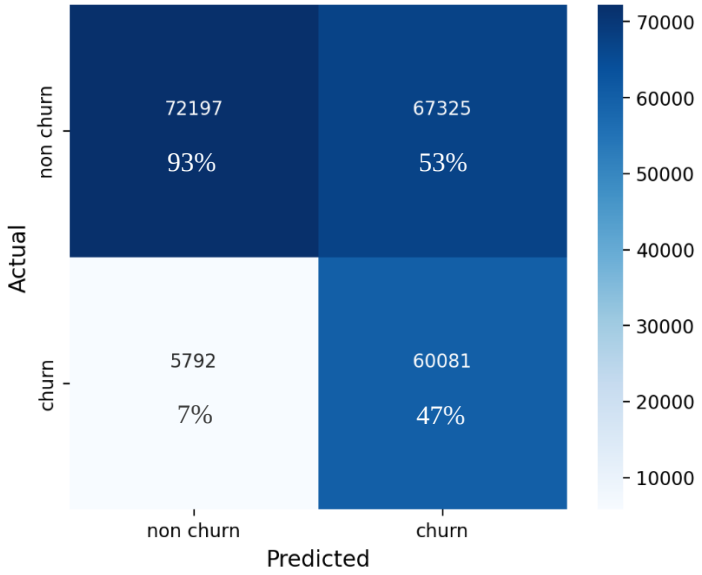


Figure 27: Confusion matrix of the Support Vector Machine employing K-means clustering; customers considered non-churn by the trained model are inadequately identified, resulting in lower overall performance

Although there is a very high recall for the positive class but for the negative class, recall is just 52%, the precision is less than 50% for the churn class performance-wise, but the overall scores are not siginificant. Table 13 mentions the performance metrics of SVM using K-means clustering.

Table 13: The outcome of the Support Vector Machine yields a high Recall and a substantial F1 score when employed with the K-means clustering strategy; however, the negative class exhibits a negligible Recall

	Precision	Recall	F1-score	Support
Non-Churn	0.93	0.52	0.66	65873
Churn	0.47	0.91	0.62	139522
Accuracy			0.64	205395
Macro Avg	0.70	0.71	0.64	205395
Weighted Avg	0.78	0.64	0.65	205395

Among the three methods that are used in combination with the Support Vector Machine, RFM is the best method, then comes k-mean and lastly, average time, which is the least-performing method.

4.5.7 Extreme Gradient Boosting using Average Time

The third approach for predicting non-returning customers is Extreme Gradient Boosting or Xg-boost. Customers who were expected to churn but didn't, according to the data. The number of customers who are misclassified as non-churners when they are actually churners, their ratio is 7%, and 53% of the consumers are wrongly predicited as churn. (For reference, see Figure 28).

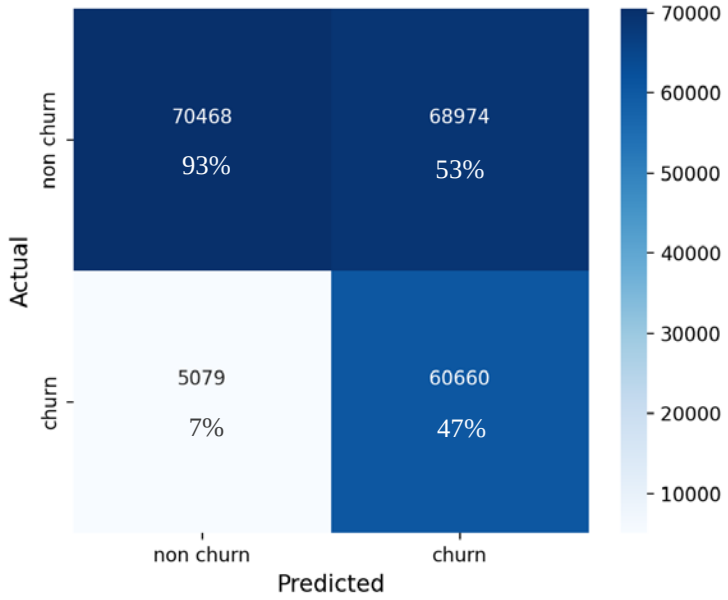


Figure 28: The average time approach for labelling the target variable and Extreme Gradient boosting for the classification of customer data produces an unsatisfactory result when actual and anticipated outcomes are compared

The recall metric is high for churn, and for non-churn, precision is high. Other values are not that remarkable. More information is required for labelling; only average time will not be enough. The overall results are presented below in Table 14.

Table 14: The findings of the Extreme Gradient Boosting using the Average Time Method show that the positive class has a low Precision and a high Recall, but the negative class has a lowered accuracy when anticipated observation is compared to real data

	Precision	Recall	F1-score	Support
Non-Churn	0.93	0.51	0.66	139442
Churn	0.47	0.92	0.62	65739
Accuracy			0.64	205181
Macro avg	0.70	0.71	0.64	205181
Weighted avg	0.78	0.64	0.64	205181

4.5.8 Extreme Gradient Boosting using RFM

When RFM was combined with random forest and support vector machines, it performed incredibly well. It also worked really well with Xg-boost. Miss-categorization is quite rare for churn class. Initially anticipated to churn, but predicted as non-churn is 5% of the customers as presented in Figure 29.

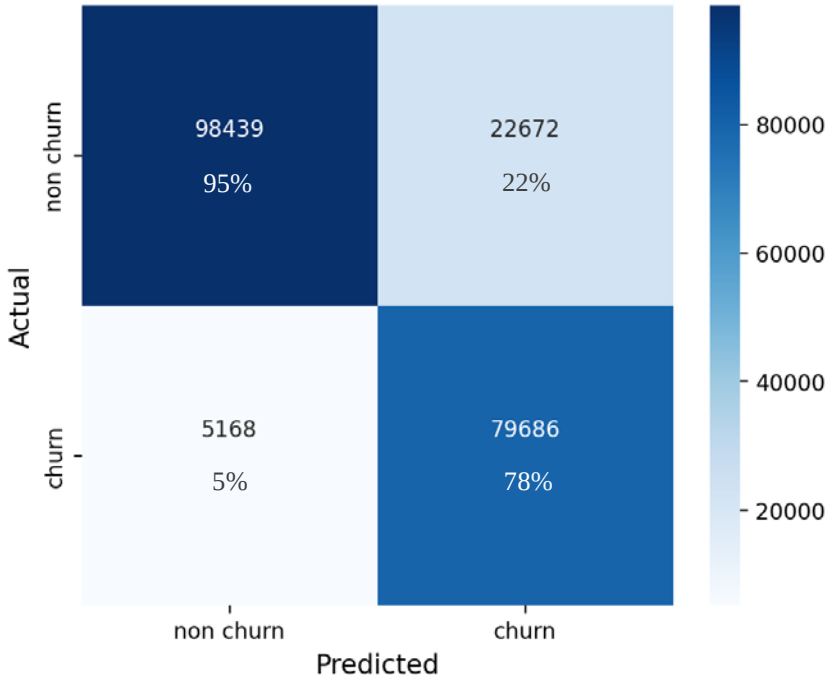


Figure 29: Confusion matrix of the Extreme Gradient Boosting using the RFM method indicates that the observational errors are minimal, at most 5% for the positive class and 22% for the negative class

From Table 15, the result of Xg-boost are satisfying and can be used for the prediction of churn customers. Both accuracy and f1 score gives (85% to 88%) accurate results.

Table 15: The F1 score for both types of consumers, whether they are existing or at the chance of being churned, is over 84%, and both classes perform the same when data is labelled using the RFM approach, and the model is trained by Extreme gradient boosting

	Precision	Recall	F1-score	Support
Non-Churn	0.95	0.81	0.88	121111
Churn	0.78	0.94	0.85	84854
Accuracy			0.86	205965
Macro avg	0.86	0.88	0.86	205965
Weighted avg	0.88	0.86	0.87	205965

4.5.9 Extreme Gradient Boosting using K-means Clustering

K-mean is the third method that is used with extreme gradient boosting. The number of examples were actually churn and are misclassified as non-churn is 33%, as in Figure 30. The three methods that are used with k-mean extreme gradient boosting have better results.

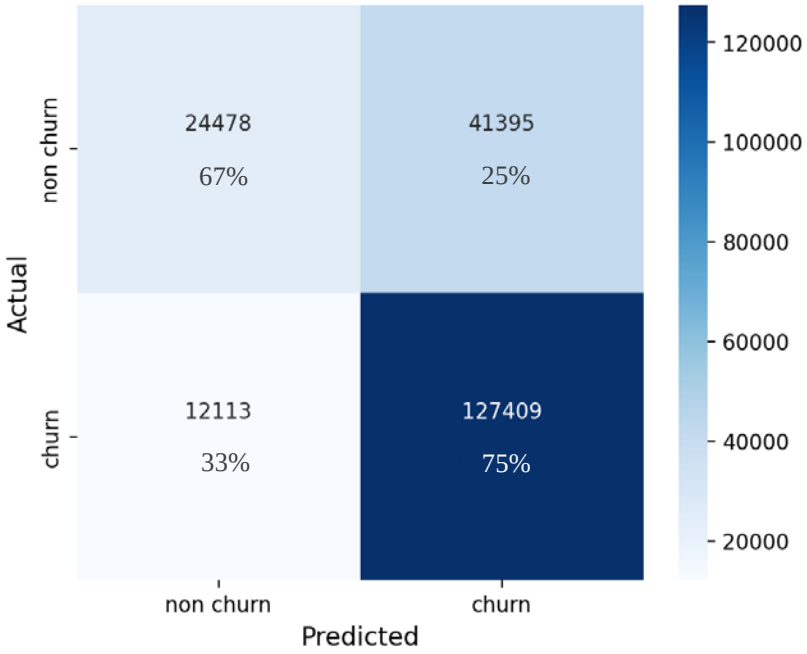


Figure 30: Contingency table of the Extreme Gradient Boosting when customers are described by applying the K-means clustering algorithm, the results are significant for the churning customers but not for existing clients

By looking at table 16, it can be seen that while the model performs poorly for the non-churn label, the churn class is predicted well with somewhat reasonable precision and recall.

Table 16: The performance of the non-churn class is negatively affected; this model is not the best choice when evaluating the Extreme Gradient Boosting model using the K-means clustering approach

	Precision	Recall	F1-score	Support
Non-Churn	0.67	0.37	0.48	65873
Churn	0.75	0.91	0.83	139522
Accuracy			0.74	205395
Macro avg	0.71	0.64	0.65	205395
Weighted avg	0.73	0.74	0.71	205395

4.6 AUROC Validation

The AUROC curve, also known as the area under the receiver operating characteristic curve, was used to verify the models' efficiency. AUROC helps to see the model's performance visually. High true positive rates and low false positive rates are the desired outcomes. It shows the ratio of true positives to false positives at various thresholds.

4.6.1 Validating Random Forest Result

ROC curve and Area under the curve score or AUC of random forest with three variations of average time, rfm and k-mean are compared in Figure 31 below. The highest score for AUC is achieved, or the best performing method is when Random Forest is used with the RFM.

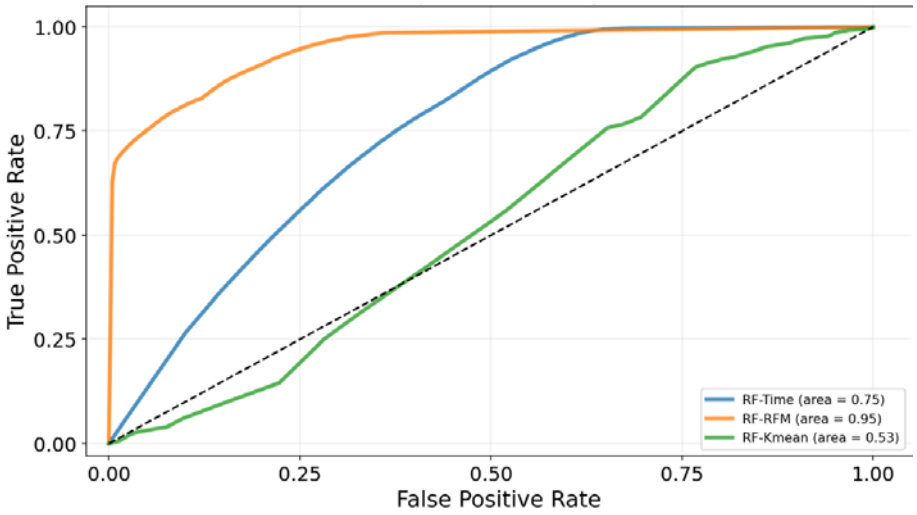


Figure 31: Area under the roc curve plot for the Random Forest algorithm presented in the graph determines that when the customer data are specified by utilising the RFM strategy, it achieves maximum performance

4.6.2 Validating Support Vector Machine

The best ROC curve is the one which is close to the y-axis and has the maximum area under the curve. Considering Figure 32, when the Support Vector Machine is used with the average time method and k-mean, the measure of separability is 76% and 78%, respectively. When SVM is used with RFM, the area under the curve is 92%, which is the best among the three methods used with a Support vector machine.

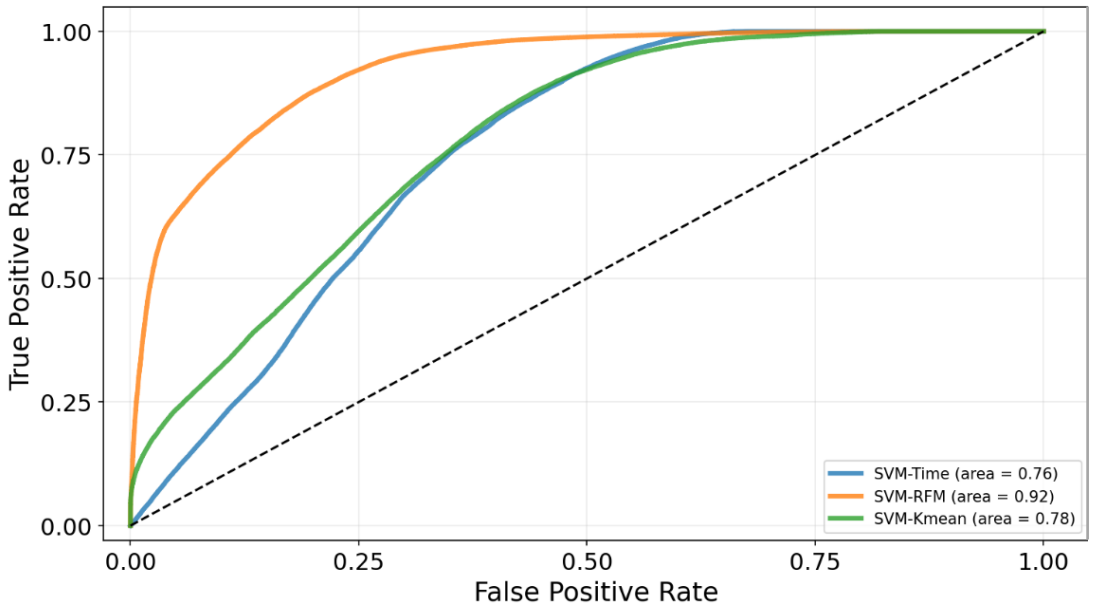


Figure 32: Support vector machine algorithm is applied to the customer data after labelling data through distinct methods that, are the Average time approach, RFM technique, and by using an unsupervised learning algorithm, the Area under the roc curve is higher when the customers are defined by using RFM method

4.6.3 Validating Extreme Gradient Boosting Result

In each of the three methodologies and algorithms utilised for customer labelling, Extreme Gradient Boosting, or Xg-boost, typically produces better results. RFM outperforms the k-mean technique and average time in terms of performance.

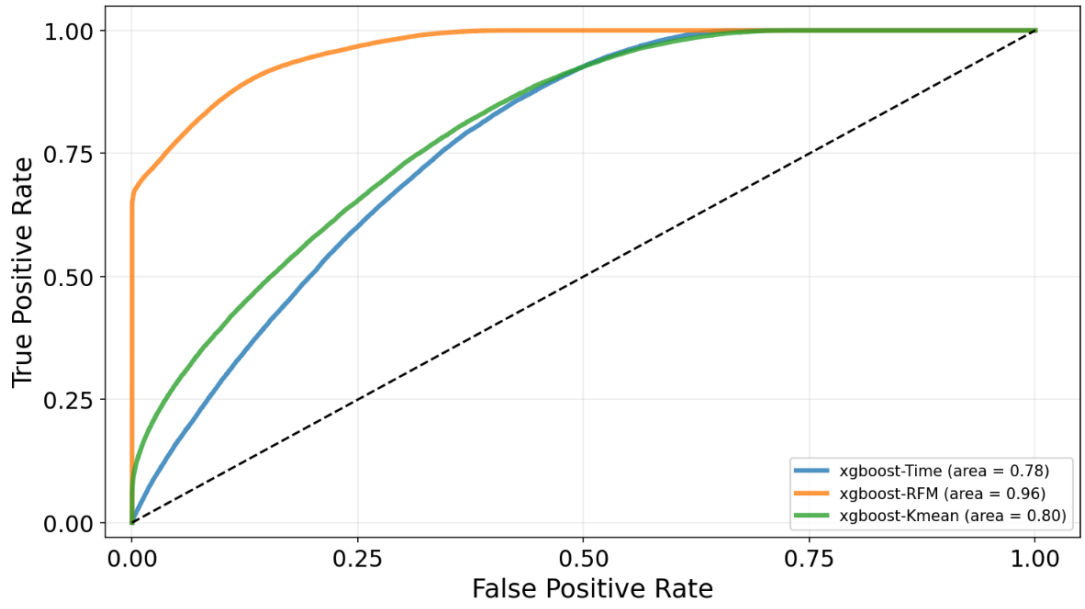


Figure 33: Area under the curve plot for Extreme gradient boosting algorithms is visualized, and it is evident from the graph that Recency, Frequency and Monetary methodology is the best method among the three that are employed for labelling the customer data to attain accurate prediction

4.6.4 Best Models

In the former section, the ROC curve and AUC score are plotted in such a way that each algorithm with its three variations for labelling customers is demonstrated separately. Interestingly the method which is more accurate out of the three is RFM. The three best algorithms are plotted in Figure 34. XG-Boost is the best algorithm to predict churn.

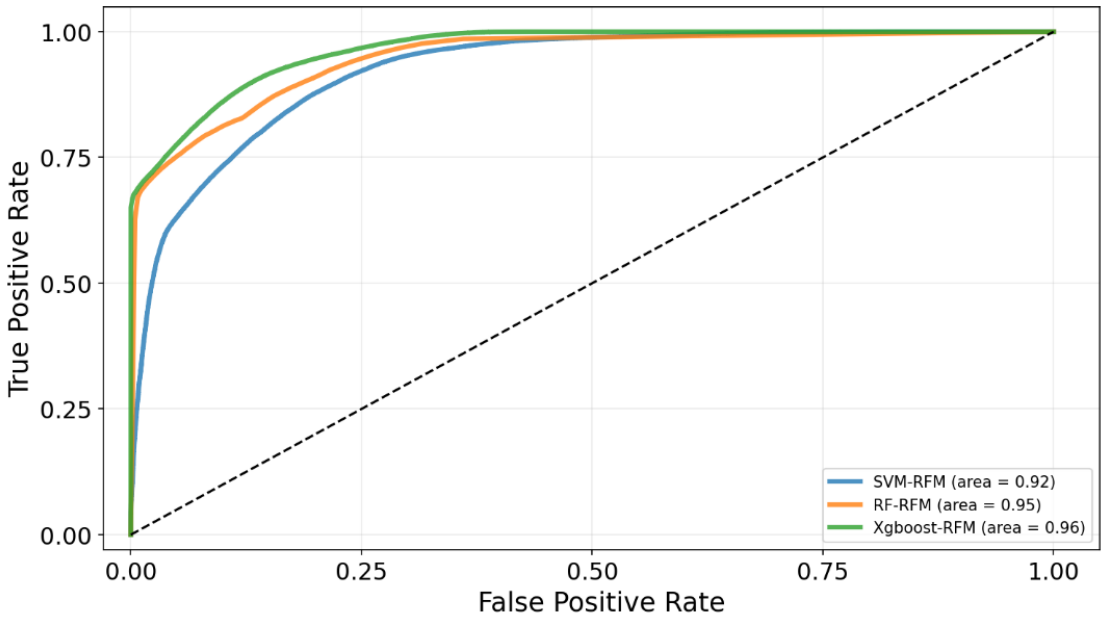


Figure 34: Area under the roc curve plot for the top three performing models are displayed in the graph for The Extreme Gradient Boosting, Random Forests and Support Vector Machine; all these algorithms achieve significant performance when the RFM Technique defines the client data

4.6.5 Accuracy and AUROC

Both metrics are used for evaluating the performance of the classifier. Accuracy takes predicted classes, and ROC takes predicted scores as input. Accuracy is computed at a single threshold, while AUC is calculated for all threshold values. AUROC analysis shows how well a positive class can be separated from the other class. In this case, there are different values for accuracy and AUROC. For example, when Xg-boost is used with the RFM, accuracy is 86%, and AUROC is 96%. AUROC performs well when data is imbalanced, while accuracy does not perform well if the data is not balanced. There is often the case when a classifier works excellently on one class, which results in a high AUC score, as it is the ratio of the true positive rate with the false positive rate.

4.7 Model Memory and Time Consumption

Models are assessed using a variety of performance metrics, including area under the ROC curve, precision, recall, and f1-score. In addition, the amount of memory and time used by each algorithm is also measured. Table 17 summarizes the information.

Regarding execution time, Xg-boost takes much less time—between 12 and 17 seconds—and uses a maximum of 163 megabytes of memory. According to our findings, Random forest is the second-best technique. The model uses a maximum of 93 megabytes of memory and takes 27–59 seconds to execute. The longest running time was 59 minutes, 42 minutes, and 1 hour 59 minutes for SVM, i.e. 15 megabytes of RAM is used when RFM and SVM are combined.

Table 17: The memory and temporal performance of all the models are presented in the table to determine the maximum execution time and also the memory space it requires to identify the best, average and worst performing algorithms

	Machine Learning Model	Memory Consumption (in MB)	Execution Time
1	Random Forest using the Average Time	28.365879	0:00:26.517163
2	Random Forest using the RFM method	93.024988	0:00:46.185095
3	Random Forest using the K-mean	42.902219	0:00:58.635014
4	Support Vector Machine using the Average Time	12.19267	0:59:32.263626
5	Support Vector Machine using the RFM method	15.557115	0:42:15.197559
6	Support Vector Machine using the K-mean	89.803629	1:10:26.102443
7	XG-Boost using the Average Time	7.908419	0:00:12.625136
8	XG-Boost using the RFM method	163.264952	0:00:17.122215
9	XG-Boost using the K-mean	4.800199	0:00:16.206894

Conclusion

E-commerce is multiplying, and one of the significant problems that e-commerce companies and businesses face is customer churn. It is a situation when a customer would not return after one or more bad experiences with the company. This can affect the company's overall performance in terms of reputation and revenue. Another advantage of working with this problem was that acquiring a new customer costs 4-5 times more than working on retained customers. An interesting fact is that 80% of the revenue comes from the 20% returning customers, which suggests the importance of each customer.

Secondary data was taken from the online repository to address the customer churn problem. Data passes through different stages in preprocessing, including data cleaning, data reduction, data combining and scaling. The dataset was not labelled, so three methods were defined for labelling customers as either churn or non-churn. Average time, RFM and k-means clustering were used for that purpose. Random forest, support vector machines and extreme gradient boosting are the algorithms used to predict customers at risk of being churned in the future.

Nine models were trained to find the functional relationship. Among these, the best models that performed well include random forest with RFM, support vector machine used with RFM and XG-boost used with RFM based on accuracy, precision, recall and AUROC. In addition, time and memory consumed by each algorithm were analyzed, and XG-boost was found to be on the top when these two were a concern.

References

[1] P. Jílková and P. Králová, ‘Digital Consumer Behaviour and eCommerce Trends during the COVID-19 Crisis’, *J. Mark. Res.*, 2018, doi: 10.1007/s11294-021-09817-4.

[2] H. L. Wu, W. W. Zhang, and Y. Y. Zhang, ‘An empirical study of customer churn in ecommerce based on data mining’, *2010 Int. Conf. Manag. Serv. Sci. MASS 2010*, 2010, doi: 10.1109/ICMSS.2010.5576627.

[3] G. Taher, ‘E-Commerce: Advantages and Limitations’, *Int. J. Acad. Res. Account. Financ. Manag. Sci.*, vol. 11, no. 1, pp. 153–165, 2021, doi: 10.6007/IJARAFMS/v11-i1/8987.

[4] B. Galhotra, ‘Evolution of E-commerce in India: A Review and Its Future Scope’, *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 226–231, Feb. 2019, doi: 10.1109/COMITCON.2019.8862252.

[5] L. N. K. Leonard and K. Jones, ‘Consumer-to-consumer e-commerce research in information systems journals’, *J. Internet Commer.*, vol. 9, no. 3–4, pp. 186–207, Jul. 2010, doi: 10.1080/15332861.2010.529052.

[6] A. K. Singh and A. Ajmani, ‘Future of B2C E-commerce (Buyers Perspective) in India: An Empirical Analysis’, *https://doi.org/10.1177/2319510X16688986*, vol. 12, no. 3–4, pp. 203–216, May 2017, doi: 10.1177/2319510X16688986.

[7] J. Ahn, J. Hwang, D. Kim, H. Choi, and S. Kang, ‘A Survey on Churn Analysis in Various Business Domains’, *IEEE Access*, vol. 8, pp. 220816–220839, 2020, doi: 10.1109/ACCESS.2020.3042657.

[8] M. Bharti, ‘E-business Through Social Media: An Instagram Page Attribute-Conversion Model in Context of Fashion Apparel Industry’, *Glob. Bus. Rev.*, Sep. 2021, doi: 10.1177/09721509211038832.

[9] Q. Yanfang and L. Chen, ‘Research on E-commerce user churn prediction based on logistic regression’, *2017 IEEE 2nd Inf. Technol. Networking, Electron. Autom. Control Conf.*, vol. 2018-January, pp. 87–91, Feb. 2017, doi:

10.1109/ITNEC.2017.8284914.

[10] A. Velez-Calle, M. Mariam, M. A. Gonzalez-Perez, A. Jimenez, J. Eisenberg, and S. M. Santamaria-Alvarez, ‘When technological savviness overcomes cultural differences: millennials in global virtual teams’, *Crit. Perspect. Int. Bus.*, vol. 16, no. 3, pp. 279–303, May 2020, doi: 10.1108/CPOIB-01-2018-0012.

[11] ‘OECD Employment Outlook 2019’, Apr. 2019, doi: 10.1787/9EE00155-EN.

[12] S. Vijayakumar Bharathi, D. Pramod, and R. Raman, ‘An Ensemble Model for Predicting Retail Banking Churn in the Youth Segment of Customers’, *Data 2022, Vol. 7, Page 61*, vol. 7, no. 5, p. 61, May 2022, doi: 10.3390/DATA7050061.

[13] A. K. Ahmad, A. Jafar, and K. Aljoumaa, ‘Customer churn prediction in telecom using machine learning in big data platform’, *J. Big Data*, vol. 6, no. 1, pp. 1–24, Dec. 2019, doi: 10.1186/S40537-019-0191-6/TABLES/4.

[14] R. G. K. Vln and P. Deeplakshmi, ‘Dynamic Churn Prediction using Machine Learning Algorithms - Predict your customer through customer behaviour’, *2021 Int. Conf. Comput. Commun. Informatics*, Jan. 2021, doi: 10.1109/ICCCI50826.2021.9402369.

[15] R. Valero-Fernandez, D. J. Collins, K. P. Lam, C. Rigby, and J. Bailey, ‘Towards Accurate Predictions of Customer Purchasing Patterns’, *IEEE CIT 2017 - 17th IEEE Int. Conf. Comput. Inf. Technol.*, pp. 157–161, Sep. 2017, doi: 10.1109/CIT.2017.58.

[16] A. T. Jahromi, M. M. Sepehri, B. Teimourpour, and S. Choobdar, ‘Modeling customer churn in a non-contractual setting: the case of telecommunications service providers’, *https://doi.org/10.1080/0965254X.2010.529158*, vol. 18, no. 7, pp. 587–598, Dec. 2010, doi: 10.1080/0965254X.2010.529158.

[17] M. Hassouna, A. Tarhini, T. Elyas, and M. S. A. Trab, ‘Customer Churn in Mobile Markets: A Comparison of Techniques’, *Int. Bus. Res.*, vol. 8, no. 6, p. p224, May 2015, doi: 10.5539/IBR.V8N6P224.

[18] E. Lee, B. Kim, S. Kang, B. Kang, Y. Jang, and H. K. Kim, ‘Profit optimizing churn prediction for long-term loyal customers in online games’, *IEEE Trans. Games*, vol. 12, no. 1, pp. 41–53, Mar. 2020, doi: 10.1109/TG.2018.2871215.

[19] W. Yang *et al.*, ‘Mining player in-game time spending regularity for churn prediction in free online games’, *IEEE Conf. Comput. Intell. Games, CIG*, vol. 2019-August, Aug. 2019, doi: 10.1109/CIG.2019.8848033.

[20] S. Gupta *et al.*, ‘Modeling customer lifetime value’, *J. Serv. Res.*, vol. 9, no. 2, pp. 139–155, Nov. 2006, doi: 10.1177/1094670506293810.

[21] N. Glady, B. Baesens, and C. Croux, ‘Modeling churn using customer lifetime value’, *Eur. J. Oper. Res.*, vol. 197, no. 1, pp. 402–411, Aug. 2009, doi: 10.1016/j.ejor.2008.06.027.

[22] T. Althoff and J. Leskovec, ‘Donor retention in online crowdfunding communities: A case study of DonorsChoose.org’, *WWW 2015 - Proc. 24th Int. Conf. World Wide Web*, pp. 34–44, May 2015, doi: 10.1145/2736277.2741120.

[23] P. Kisioglu and Y. I. Topcu, ‘Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey’, *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7151–7157, Jun. 2011, doi: 10.1016/J.ESWA.2010.12.045.

[24] S. A. Neslin, S. Gupta, W. Kamakura, L. U. Junxiang, and C. H. Mason, ‘Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models’, *https://doi.org/10.1509/jmkr.43.2.204*, vol. 43, no. 2, pp. 204–211, Oct. 2018, doi: 10.1509/JMKR.43.2.204.

[25] S. Agrawal, A. Das, A. Gaikwad, and S. Dhage, ‘Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning’, *2018 Int. Conf. Smart Comput. Electron. Enterp. ICSCEE 2018*, Nov. 2018, doi: 10.1109/ICSCEE.2018.8538420.

[26] A. Mishra and U. S. Reddy, ‘A Novel Approach for Churn Prediction Using Deep Learning’, *2017 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2017*, Nov. 2018, doi: 10.1109/ICCIC.2017.8524551.

[27] S. Kim and H. Lee, ‘Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees’, *8th Int. Conf. Inf. Technol. Quant. Manag. - Dev. Glob. Digit. Econ. after COVID-19*, vol. 199, pp. 1332–1339, 2021, doi: 10.1016/J.PROCS.2022.01.169.

[28] ‘Digital transformation in financial services provision: A Digital transformation in financial services provision: A Nigerian perspective to the adoption of chatbot. Journal of Enterprising Communities: People and Places in the Global Economy Nigerian perspective to the adoption of chatbot’, doi: 10.1108/JEC-06-2020-0126.

[29] S. Dinesh and Y. Muniraju, ‘Scalability of E-commerce in the Covid-19 era’, *Int. J. Res. ISSN*, vol. 9, no. 1, p. 123, 2021, doi: 10.29121/granthaalayah.v9.i1.2021.3032.

[30] M. Hussain and A. Papastathopoulos, ‘Organizational readiness for digital financial innovation and financial resilience’, *Int. J. Prod. Econ.*, vol. 243, p. 108326, Jan. 2022, doi: 10.1016/J.IJPE.2021.108326.

[31] S. Raeisi and H. Sajedi, ‘E-Commerce Customer Churn Prediction by Gradient Boosted Trees’, *2020 10h Int. Conf. Comput. Knowl. Eng. ICCKE 2020*, pp. 55–59, Oct. 2020, doi: 10.1109/ICCKE50421.2020.9303661.

[32] Y. ZHUANG and Y. ZHUANG, ‘Research on E-commerce Customer Churn Prediction Based on Improved Value Model and XG-Boost Algorithm’, *Manag. Sci. Eng.*, vol. 12, no. 3, pp. 51–56, Sep. 2018, doi: 10.3968/10816.

[33] N. Gordini and V. Veglio, ‘Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry’, *Ind. Mark. Manag.*, vol. 62, pp. 100–107, Apr. 2017, doi: 10.1016/j.indmarman.2016.08.003.

[34] X. Wu and S. Meng, ‘E-commerce customer churn prediction based on improved SMOTE and AdaBoost’, *2016 13th Int. Conf. Serv. Syst. Serv. Manag. ICSSSM 2016*, Aug. 2016, doi: 10.1109/ICSSSM.2016.7538581.

[35] A. Hanif and N. Azhar, ‘Resolving Class Imbalance and Feature Selection in Customer Churn Dataset’, *Proc. - 2017 Int. Conf. Front. Inf. Technol. FIT 2017*, vol. 2017-January, pp. 82–86, Jan. 2018, doi: 10.1109/FIT.2017.00022.

[36] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, ‘Customer churn prediction system: a machine learning approach’, *Computing*, vol. 104, no. 2, pp. 271–294, Feb. 2022, doi: 10.1007/S00607-021-00908-Y/FIGURES/7.

[37] W. Essayem, F. A. Bachtiar, and D. Priharsari, ‘Customer Clustering Based on RFM Features Using K-Means Algorithm’, *Proc. - 2022 IEEE Int. Conf. Cybern. Comput. Intell. Cybern. 2022*, pp. 23–27, 2022, doi: 10.1109/CYBERNETICSCOM55287.2022.9865572.

[38] P. Li, S. Li, T. Bi, and Y. Liu, ‘Telecom customer churn prediction method based on cluster stratified sampling logistic regression’, *IET Conf. Publ.*, vol. 2014, no. CP660, pp. 282–287, 2014, doi: 10.1049/CP.2014.1576.

[39] E. Manohar, P. Jenifer, M. S. Nisha, and B. Benita, ‘A collective data mining approach to predict customer behaviour’, *Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mob. Networks, ICICV 2021*, pp. 1310–1316, Feb. 2021, doi: 10.1109/ICICV50876.2021.9388558.

[40] I. Franciska and B. Swaminathan, ‘Churn prediction analysis using various clustering algorithms in KNIME analytics platform’, *2017 Third Int. Conf. Sensing, Signal Process. Secur.*, pp. 166–170, Oct. 2017, doi: 10.1109/SSPS.2017.8071585.

[41] Z. Chen and Z. Fan, ‘Building comprehensible customer churn prediction models: A multiple kernel support vector machines approach’, *8th Int. Conf. Serv. Syst. Serv. Manag. - Proc. ICSSSM’11*, 2011, doi: 10.1109/ICSSSM.2011.5959439.

[42] I. Stancin and A. Jovic, ‘An overview and comparison of free Python libraries for data mining and big data analysis’, *2019 42nd Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2019 - Proc.*, pp. 977–982, May 2019, doi: 10.23919/MIPRO.2019.8757088.

[43] A. Nagpal and G. Gabrani, ‘Python for Data Analytics, Scientific and Technical Applications’, *2019 Amity Int. Conf. Artif. Intell.*, pp. 140–145, Apr. 2019, doi: 10.1109/AICAI.2019.8701341.

[44] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, ‘A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data’, *Front. Energy Res.*, vol. 9, p. 77, Mar. 2021, doi: 10.3389/FENRG.2021.652801/BIBTEX.

[45] V. ÇETİN and O. YILDIZ, ‘A comprehensive review on data preprocessing techniques in data analysis’, *Pamukkale Üniversitesi Mühendislik Bilim. Derg.*, vol. 28, no. 2, pp. 299–312, 2022, doi: 10.5505/PAJES.2021.62687.

[46] B. Galhotra and A. Dewan, ‘Impact of Covid-19 on digital platforms and change in E-commerce shopping trends’, *Proc. 4th Int. Conf. IoT Soc. Mobile, Anal. Cloud, ISMAC 2020*, pp. 861–866, Oct. 2020, doi: 10.1109/I-SMAC49090.2020.9243379.

[47] R. A. de Lima Lemos, T. C. Silva, and B. M. Tabak, ‘Propension to customer churn in a financial institution: a machine learning approach’, *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11751–11768, Jul. 2022, doi: 10.1007/S00521-022-07067-X/FIGURES/10.

[48] N. N. Y. Vo, S. Liu, X. Li, and G. Xu, ‘Leveraging unstructured call log data for customer churn prediction’, *Knowledge-Based Syst.*, vol. 212, pp. 1–15, Jan. 2021, doi: 10.1016/J.KNOSYS.2020.106586.

[49] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, ‘A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector’, *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.