PREDICTING CUSTOMER CHURN IN E - COMMERCE USING MACHINE LEARNING

MSc Research Project
Data Analytics

AHSAN SHAHZAD
Student ID: x23256460

School of Computing
National College of Ireland

Supervisor: SYED MUSLIM JAMEEL

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	AHSAN SHAHZAD		
Student ID:	x23256460		
Programme:	MSc in Data Analytics	**Year:**	2024-25
Module:	MSc Research Project		
Supervisor:	SYED MUSLIM JAMEEL		
Submission Due Date:	15th September 2025		
Project Title:	PREDICTING CUSTOMER CHURN IN E-COMMERCE USING MACHINE LEARNING		
Word Count:	6249	**Page Count :**	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: AHSAN SHAHZAD

Date: 15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

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

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

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

Table of Contents

List of Tables

List of Figures

1 Chapter 1: Introduction

1.1 1.1 Introduction

Due to e-commerce, businesses can now reach consumers more freely, with more options and with fewer limits on customization (Vanneschi *et al.*, 2018). At the same time, the extremely competitive nature of this field has created major issues, and keeping clients is one of these. It costs more to get new clients than to retain them, so online retailers worry about customer churn a lot. A successful way to retain customers is by knowing and forecasting churn (Shobana *et al.*, 2023). Handling business issues with data-driven approaches has become much more widespread over the past several years and customer churn prediction is just one such example. Machine learning (ML) can inspect a lot of consumer data to track the patterns that signal churn is happening using firms may take advantage of ML for tailored offers, rapid responses, and excellent customer service by using information stored in the past and customers' actions (Shaker Reddy *et al.*, 2024). Conventional churn analysis normally focuses on straightforward rules and looks at old trends, because they could not deal with the detailed and irregular ties in consumer behavior (Jahan and Sanam, 2024).

1.2 1.2 Background

Due to e-commerce, people can shop for goods and services whenever they want which has greatly impacted the global retail industry (Al Rahib *et al.*, 2024). As more activities happen online, companies can collect lots of information about their customers which they use to make valuable decisions (Pondel *et al.*, 2021). Customer churn which happens when customers stop using a company, still worries businesses and reduces e-commerce earnings, recent developments notwithstanding (Yayun, 2018). Customer turnover directly affects a business's financial results, ability to keep customers, and aspirations to expand. It is easier and cheaper to keep a customer than to find a new one, so predicting when clients might drop out is very important (Raeisi and Sajedi, 2020). Surveying customers and looking at past trends is not enough to find out about complicated risks and changing customer behavior. Powerful enough to act as a substitute, machine learning can now observe both structured and unstructured data and use what it learns to foresee results. Machine learning algorithms can discover high-risk clients before they attempt to withdraw from the business by checking their purchase record, shopping activity, previous transactions, and demographics. Using this function, online stores can grow the client's lifetime value by targeting retention efforts, adjusting marketing processes, and more (Xia and He, 2018). It examines plans to enhance the accuracy of predicting customer churn by applying machine learning models.

1.3 1.3 Research Aim and Objectives

1.3.1 1.3.1 Aim

This research aims to develop an advanced machine learning-based churn prediction model for e-commerce businesses by leveraging historical customer data, optimizing feature selection, and evaluating the effectiveness of multiple machine learning algorithms to improve customer retention strategies.

1.3.2 *1.3.2 Objectives*

- To analyze customer churn patterns in e-commerce businesses by preprocessing historical data and identifying key behavioral and transactional indicators.
- To implement and compare multiple ML models, including XGBoost, LightGBM, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks, to determine the most effective algorithm for churn prediction.
- To optimize model performance through feature selection and hyperparameter tuning, ensuring higher predictive accuracy and model efficiency.
- To evaluate the effectiveness of the developed models using precision, recall, F1-score, and AUC-ROC metrics, while enhancing model interpretability for actionable business insights.

1.4 **1.4 Research Question**

- What key factors contribute to customer churn in e-commerce businesses, and how can they be effectively analyzed using machine learning techniques?
- Which machine learning model such as XGBoost, LightGBM, Convolutional Neural Networks (CNN), or Long Short-Term Memory (LSTM) provides the highest accuracy and reliability for predicting customer churn?
- How can feature selection and hyperparameter tuning improve the performance of machine learning models in churn prediction?
- How can the developed churn prediction model be effectively evaluated using precision, recall, F1-score, and AUC-ROC to ensure its practical applicability in customer retention strategies?

1.5 **1.5 Research Significance**

Online retailers need to be able to predict when customers will churn for their long-term well-being and money savings (Hasan *et al.*, 2024). Since there are plenty of choices and much competition nowadays, it is more difficult to hold onto existing clients (Xia and He, 2018). This research is beneficial as it studies if machine learning techniques can foresee which consumers will not stay signed up. The study strives to help firms refine how they retain customers by improving churn predictions using both historical and behavioral information.

If online platforms use this study, they might enjoy lower costs to get new clients, more satisfied customers, and a smaller loss of current customers. By looking at and studying the performance of various ML models, the field of customer relationship management continues to develop.

1.6 1.6 Problem Statement

Businesses in online retail often face big issues with customer turnover. Since churn depends on many factors such as a person's number of purchases, how involved they are, and how satisfied they are with the service, knowing who will churn is a tough task. This research uses machine learning to try to meet the industry demand for a better and more active way to predict client attrition.

2 Chapter 2: Literature Review

2.1 2.1 Introduction

By emphasizing machine learning, this chapter reviews previous published papers on identifying why e-commerce clients might unsubscribe. To make it easier to understand how machine learning approaches used in predicting customer churn and to note the gaps present in the previous studies. This study explains crucial findings, common algorithms, the nature of data, and evaluation methods.

2.2 2.2 Understanding Customer Churn in E-Commerce

As per Ulas *et al.*, (2021), spending on existing customers is more economical than trying to gain new ones. Therefore, companies should analyze why customers leave to adjust their operations and stop such customers from churning later. Telecoms, insurance, and financial services use churn research a lot since customers switching to another provider is easy for them. Very little attention has been paid by researchers to studying and predicting churn in online shopping. A reliable churn prediction model is then presented after this study covers all the factors that can decrease consumer loyalty in e-commerce, both immediately and in the background. Applications are made to show that the model is useful and works well. With actual data, we noticed that the model was able to predict very accurately.

2.3 2.3 Traditional Approaches to Churn Prediction

Customer churn models reveal people who might soon leave the company. In this way, an effective machine learning model helps increase earnings by catching "at-risk" clients before they have problems. Due to this retention advertising reaches more people and reduces expenses for employee exits (De and Prabu, 2022). It evaluates the latest approaches to machine learning models for churn prediction. A review of the literature has been performed using 5 research topics and applying high-quality standards. From the 420 studies published between 2018 and 2021, a total of 38 main studies were picked for the study. It explains well-known churn prediction techniques and suggests looking into other research areas. This means researchers have to study other techniques rather than just doing more experiments, to improve the performance of classifiers in different places (De and Prabu, 2022). Second, traditionally, churn prediction depends mostly on demographics, how often the product is used, and income. Using social network theory in churn models has given good results in recent studies. All over the business world, clients leaving impacts profits and adds to the costs of finding new customers. ML and DL have improved how churn prediction is done since they can analyze fast-changing and complex consumer data in ways that statistical methods cannot. Research studies from telecommunications, retail, banking, healthcare, education, and insurance launched between 2020 and 2024 show the present progress made in predicting churn (Imani *et al.*, 2025). It explores how traditional ML approaches including Decision Trees, Random Forests, and boosting change over time, their strengths and weaknesses, and also investigates advanced DL methods like CNNs, LSTM networks, Transformers, and hybrid models. Beneficial frameworks, ensembles, advanced architectures, attention-based techniques, Explainable AI (XAI), and adaptable learning are the top new developments (Imani *et al.*, 2025). Still, issues of unequal treatment, difficulty explaining DL models, and adjusting to changing consumer lifestyles need to be overcome.

2.4 **2.4 Machine Learning Techniques for Churn Prediction**

2.4.1 ***2.4.1 Logistic Regression***

Handling customer turnover is an important matter in e-commerce and this study creates a mix of logistic regression and an XGBoost method to determine customer turnover. The model is used on actual orders and transactions made on an e-commerce website (Li and Li, 2019). The real data mining process found over twenty top indexes that help in areas such as purchase facts, details about customers, and help after the sale. These indices were used with the hybrid technique to predict whether each sample consumer would churn or not. With the help of the algorithm, the authors were able to spot customer attrition with a high degree of accuracy (more than 85%), the data indicates (Li and Li, 2019). These findings give useful advice to online companies trying to build customer loyalty.

2.4.2 ***2.4.2 Random Forest***

Today, with so many businesses operating on the Internet, managing customer relationships is very important for remaining ahead of competitors. A good way to set up a CRM strategy is to define each customer class by analyzing their Customer Lifetime worth (CLV) which shows their total value as customers (Win and Bo, 2020). The value to customers is recognized by applying a customer lifetime value formula. It allows merchants to figure out how to use their resources. The prediction model counts on 10,000 transactions from Super Store retail all around the world. With CLV-based predictions, the online retailer will know which customers are worth long-term investment for CRM (Win and Bo, 2020). The model is built with Random Forest (RF) and optimized by Random Search for high accuracy in predictions. Using experimental results to test how AdaBoost algorithms function on the same dataset.

2.4.3 ***2.4.3 Support Vector Machine***

An accurate churn estimate allows one to speak with the customer and try to keep them. The main purpose of the new user model in this article is to predict if or when a user might decide to churn (Berger and Kompan, 2019). The model is putting together feature sets that represent how a user interacts with the web app. Analyzing churn with real data from online shops helps us judge how well the model functions (Berger and Kompan, 2019). The suggested model for prediction surpasses the results of baseline model prediction in both domains.

2.4.4 ***2.4.4 Gradient Boosting Machines***

Imani, (2025), experimented with Artificial Neural Networks, Decision Trees, Support Vector Machines, Random Forests, Logistic Regression and new methods of advanced gradient boosting (XGBoost, LightGBM, CatBoost) using a public dataset to predict customer churn in the telecommunications industry. Precision, Recall, F1-score, and ROC AUC were used to measure performance of models. LightGBM outranked Decision Trees, and Logistic Regression with F1-score of 92 percent and ROC AUC of 91 percent. High-fidelity algorithms do a better job of dealing with unbalanced datasets and complex feature interactions and thus are perfectly suited to real world patterns of churn prediction.

2.4.5 ***2.4.5 Neural Networks***

Recency, Frequency, and Monetary (RFM) analysis, K-means clustering, and deep learning models (LSTM, GRU) all predict e-commerce customer attrition in this research study, as opposed to Logistic Regression (LR), Random Forest (RF), and Decision Trees. RFM and K-means were used to classify the customers into six groups on Olist, Instacart, and Online Retail datasets. On large datasets, LSTM and GRU model achieved a precision of 99.9 percent and 99.6 percent, respectively, as compared to classical models (Zaghloul *et al.*, 2025). LR worked well with fit smaller datasets whereas RF and DT worked moderately. Due to the better consideration of temporal relationships, deep learning is also a successful churn prediction and retention methodology.

Telecommunication companies need to make projections of customers who might leave the company as it is more profitable to retain clients rather than having to acquire them. Conventional churn predictive models are insensitive, and thus it becomes hard to understand customer turnover factors (Neeraj *et al.*, 2025). A multi-stage churn prediction model is created using Graph Neural Networks (GNNs) to enhance accuracy and efficiency applied to forecasts of the telecom sector. This technique records the complex relationship and dependence of customers in a graphical form. The method provides intelligible and viable decision-making interpretations through LIME and SHAP. Experimental findings demonstrate that GNN-based model does not lose explainability, a fact that can assist telecom companies with churn management and the improvement of customer retention levels (Neeraj *et al.*, 2025).

2.5 **2.5 Literature Gap**

Different from banking and telecommunications, e-commerce has neglected to research customer turnover prediction. With access to a lot of consumer data, various researchers point out that online retailers lack detailed prediction models. The majority of the time, traditional marketing relies on information like age, gender, and data on what someone has bought. No one has compared machine learning methods on datasets created especially for e-commerce, although Logistic Regression, Random Forest, and Support Vector Machines achieve outstanding accuracy. It is difficult to make models easy to understand and handle continuous changes in user actions, so many people do not use deep learning methods as much as they should. There is not a lot of attention given to online shopping-specific hybrid or ensemble methods in the research world. Therefore, this research evaluates several ML models using real e-commerce data and aims to improve accuracy in predictions which can help keep customers linked to the service.

3 Chapter 3: Research Methodology

3.1 3.1 Introduction

This chapter presents the research methodology followed in predicting consumer attrition in the e-commerce industry using machine learning. To bring a complete representation of the problem, the study employed the mixed-method design with its qualitative and quantitative methods. To identify the existing approaches to the problem of churn prediction, as well as the boundary of the models and the research gaps, the qualitative part encompasses a literature review of the relevant sources. The quantitative part focuses on the application of machine learning models to the issue of customer churn prediction on a real-world dataset available on Kaggle (Mammadov, 2024). The work follows the actions of Knowledge Discovery in Databases (KDD) and involves the selection of the appropriate data that is followed by its cleaning, transformation, mining, and lastly, it is evaluated. Among the common assessment tasks on machine learning algorithms, such as XGBoost, LightGBM, CNN, and LSTM, there are accuracy, precision, recall, F1-score, and AUC-ROC. This approach to the improvement of model creation and performance evaluation through data-driven research and a strong foundation on prior academic concepts is also reinforcing.

3.2 3.2 Knowledge Discovery in Databases (KDD)

The general objective of Knowledge Discovery in Databases (KDD) is the discovery of actionable information within huge data sets (Azevedo, 2019). Data transformation is a complex process as it assists in converting raw data into handy models, conclusions, or trends that can be utilized effectively as part of the decision-making (Shikhli and Hammad, 2018). A typical KDD process consists of five major stages: selection of data that is relevant to a larger data set; pre-processing the data (e.g. cleaning and treating noisy or missing data); transformation of the data (e.g. normalization or aggregation of the data); mining of the data (e.g. applying an algorithm to find patterns or models) and finally, the actual evaluation and interpretations of the results (e.g. making sure results are useful, accurate and relevant) (Ljupco Markuseski *et al.*, 2019). Knowledge discovery and data preparation (KDD) encompasses the entire cycle, that is, starting with data preparation, focusing on knowledge extraction, as opposed to mere data mining.

3.2.1 3.2.1 Data Selection

The Knowledge Discovery in Databases (KDD) approach starts by defining the data that is of interest to the study objective. This study utilized the Kaggle Customer Churn Prediction -All Features dataset as the dataset has a full range of customer information used to predict e-commerce churn (Mammadov, 2024). The dataset contains a churn label, the demographic variables, purchasing history, statistics of the use of the service, and billing-associated variables, such as tenure, monthly payments, and overall expenses.

These characteristics serve the research objective because the demographics and spending patterns of the customers and the duration of their tenure influence churn behavior. Unnecessary characters such as customerID column were deleted to ensure that the predictive value was highlighted. All the characteristics that have been preserved are well selected to give the relevant binary categorization details. This selective screening and selection was in line with the purpose of the study which was to predict turnover, and it met the KDD data selection criteria.

3.2.2 *3.2.2 Data Preprocessing*

Data preprocessing is critical to ensure data quality and eliminate inconsistencies, noise, and missing values. In this study, the dataset underwent several preprocessing steps:

- **Handling missing values:** The dataset contained a small number of missing values in the TotalCharges column. These were imputed using median values to avoid data loss, while non-informative attributes like customerID were removed.
- **Categorical encoding:** Categorical variables (e.g., gender, contract type, internet service) were converted into numerical form using label encoding and one-hot encoding where appropriate, enabling compatibility with machine learning algorithms.
- **Scaling numerical features:** Continuous variables such as MonthlyCharges, TotalCharges, and tenure were normalized to ensure balanced influence on distance-based and neural models.

These steps ensured a clean and reliable dataset, which is crucial for model training.

3.2.3 *3.2.3 Data Transformation*

In the transformation stage, the data is reshaped and normalized to improve model performance. The following transformations were applied:

- **Feature scaling:** Considering continuous variables such as MonthlyCharges, TotalCharges and Tenure as scalable variables using MinMaxScaler and StandardScaler. Such a standardized value is essential to such neural models as LSTM and CNN.
- **Feature recoding and preparation:** The numerical variables were recoded to satisfy the model need, and the categorical features such as PaymentMethod and ContractType would be one hot coded. The data set was shaped into 3D array (samples × timesteps × features) to be processed sequentially by deep learning models using LSTM.

This transformation prepares the dataset for effective data mining.

3.2.4 *3.2.4 Data Mining*

At this stage, various machine learning algorithms were applied to the processed and transformed dataset to extract patterns and build predictive models. Four algorithms were selected based on their proven performance in classification tasks:

- **XGBoost:** XGBoost is applied to tabular information and sophisticated feature interaction efficiency.

- **LightGBM:** A gradient boosting network called LightGBM was considered fast and precise due to its superior performance through grid search cross-validation and 5-fold cross-validation.
- **Convolutional Neural Networks (CNNs):** CNN was applied experimentally to reshaped tabular data to explore whether convolutional layers could capture local feature dependencies.
- **Long Short-Term Memory Networks (LSTM):** The data set is divided into a sequence format (samples x timesteps x features) and experimented with the capacity to identify the quasi-temporal associations such as tenure and billing sequence using Long Short-Memory Networks (LSTM).

These models were trained using an 80/20 train-test split and further validated using k-fold cross-validation to ensure generalizability.

*3.2.5 **3.2.5 Evaluation and Interpretation***

The final stage of the KDD process involves evaluating the models to determine their effectiveness in predicting churn. The following performance metrics were used:

- **Accuracy:** Measures the overall correctness of the model’s predictions.
- **Precision:** Assesses the proportion of correctly identified churners among all those predicted as churners.
- **Recall:** Evaluates the model's ability to identify actual churners.
- **F1-score:** Provides a balance between precision and recall.
- **AUC-ROC:** Gauges the model’s ability to distinguish between churn and non-churn classes.

Tweaking of hyperparameters was done with the aid of GridSearchCV in order to optimize performance and prevent overfitting. This allowed investigating many parameter combinations to find optimal parameters per algorithm. The models were further tested by k-fold cross-validation in order to further ensure that the results obtained could be generalized to a larger situation.

The main focus in the interpretation of the findings was finding out which algorithm will have the most accurate and reliable prediction of churn detection, and the comparison of the performance of models using evaluation measures.

4 Chapter 4: Data Analysis and Findings

4.1 4.1 Introduction

This chapter provides the results and discussion of predicting the turnover of customers using four machine learning models, XGBoost, LightGBM, CNN, and LSTM. Data preprocessing and feature engineering were done to ensure the same dataset was used to train and test all the models to enable a fair comparison. There was GridSearchCV (and 5-fold cross-validation) done, only in XGBoost and LightGBM, to make the performance of the tree-based models better. A set of metrics to measure the model performance included accuracy, precision, recall, F1-score, and AUC-ROC. The consumer attrition is predicted by each of the models, and the findings indicate the extent to which each model predicts the customer attrition better than the other.

4.2 4.2 Exploratory Data Analysis (EDA)

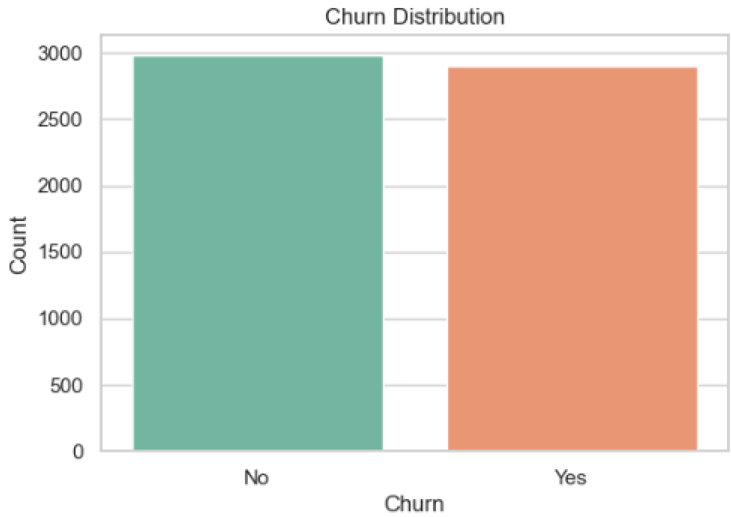

Figure 1: Churn Distribution

The bar chart explains the distribution of customer turnover in the datasets using the ratio of No to customers who stayed and Yes to customers who left. It can be seen in the graphic that the number of churners and non-churners is approximately even, with the sample size of the two studied groups at about 2,900. Such a balance reduces potential majority-class discrimination in the model, a desirable attribute in a binary classification task. The metrics accuracy, recall, and F1-score, which demonstrate the ability of a model to distinguish between churn and retention, are better demonstrated on a balanced sample. Conversely, an uneven dataset may lead to inaccurate results as well as a low level of applicability. It would appear that it is not necessary to use resampling such as SMOTE or undersampling at the preprocessing stage since the distribution is basically equal. Put together, these results speak in favor of the fact that this data is the most appropriate to teach ML models to predict churn with high precision without taking extreme actions in terms of optimizing this imbalance in the classes.

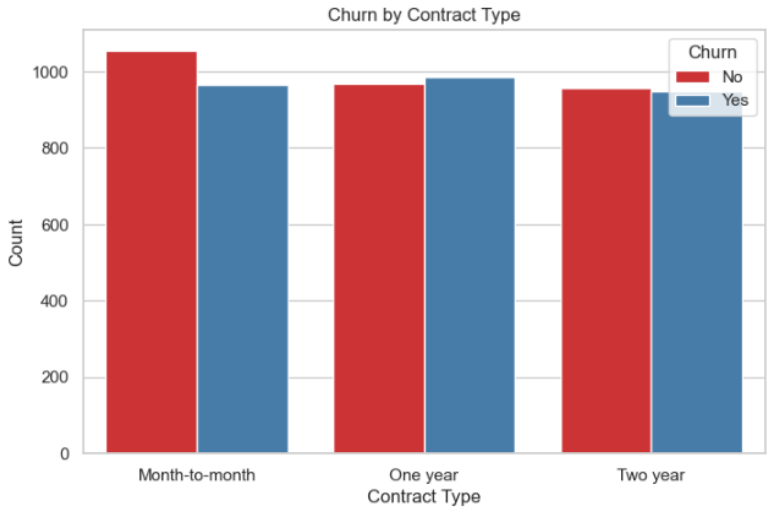

Figure 2: Churn by Contract Type

The month-to-month, one-year, and two-year contracts in terms of the distribution of the customer turnover are given in the bar chart. Based on the data, it can be seen that there are various types of contracts, but the ones that suffer most as far as churn rate is concerned are the month-to-month contracts, where the proportion of churners is even higher. It seems that the contracts are longer and thus provide more stability, lowering the risk of turning the client off, and this may be due to incentives, discounts, or contractual terms. Since the current literature indicates that contract type has been identified as an important forecaster of churn behavior, the data is valuable to firms that are engaged in e-commerce business. To reduce churn, it can be beneficial to introduce a strategy (loyalty program or discounts) that forces the customer to select the long-term contract. The infographic becomes a basis for tenure-based retention programs and outlines the importance of contract-type churn prediction models.

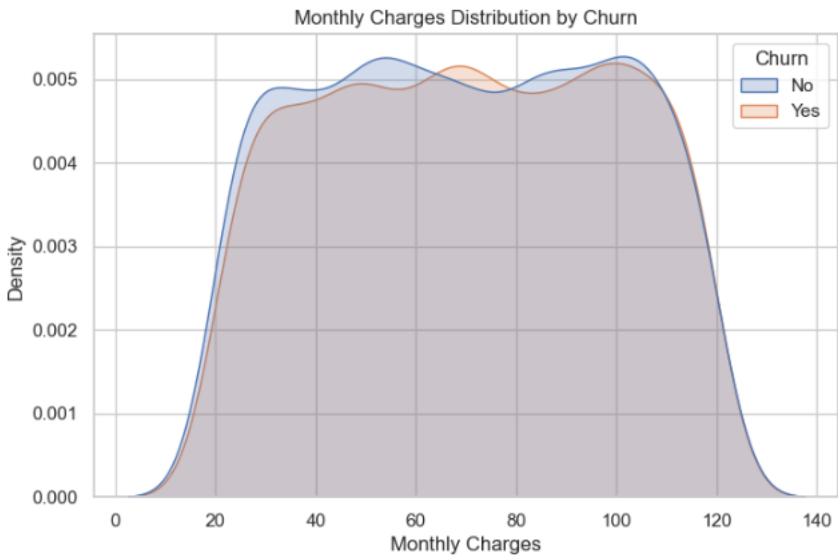

Figure 3: Monthly Charges Distribution by Churn

The density map displays monthly expenses of Yes or Yes (Yes) and the No or No (No) churning and staying clients, respectively. All in all, the distributions of the two groups are extremely similar, a fact that implies that churn occurs across a large spectrum of monthly rates. However, unlike non-churned customers, churned customers are relatively distributed in the mid-high charge range (approximately, $60-$90). Customers who pay medium to higher monthly fees are most likely to churn off, and this could be as a result of their perception that they are not getting the value of what they are paying, or they might not be satisfied with the service delivery at the price they pay. On the other hand, customers having small monthly rates appear to be slightly more susceptible to maintenance. The difference is not very drastic, yet it does indicate that the monthly payments constitute a significant, albeit not vital, element of retention scores among customers. The churn can be avoided by adding price-adjusting or value-added services to more paying segments of customer groups, and the insight implies that the cost of monthly service should be incorporated into the predictive model as a continuous parameter.

Figure 4: Tenure Vs Monthly Charges by Churn

With the churn status as a separate variable, the correlation between the number of months and the monthly cost is illustrated by the scatter plot. An evident pattern will be evident: the churn (orange dots) is more concentrated in terms of a shorter tenure (less than 20 months), which leads to the conclusion that newly acquired customers are more likely to leave the company prematurely. The customers with more than 40 months of stay with the company are likely to remain loyal, and it does not matter what they are paying a month. Customer lifetime turns out to be a strong indicator of retention. Although monthly charges vary to some extent, tenure appears to be a better indicator of churn behavior and ought to carry more weight in the prediction models.

4.3 **4.3 Results**

4.3.1 ***4.3.1 XGBoost***

Table 1: XGBoost Performance (Without GridSearchCV and 5-Fold Cross-Validation)

Metric	Value
Accuracy	0.489
Precision	0.482
Recall	0.501
F1-Score	0.492
AUC-ROC	0.475

The XGBoost model was especially poor when it was not hyperparameter-tuned and cross-validated on all the measures of evaluation. Having the AUC value of 0.475 and being accurate of 48.9 percent, the model is only a little better than blindly guessing, which is worse than that provided by chance. F1-score, recall, and precision are placed between the 0.48 and 0.50 values, which in turn means that the model is not optimal and does not reveal meaningful patterns in the data. These results show that hyperparameter optimization and cross-validation are important in churn prediction tasks, such that the base configuration does not generalize effectively and provide reliable forecasts.

Table 2: XGBoost Performance (With GridSearchCV and 5-Fold Cross-Validation)

Metric	Value
Accuracy	0.4787
Precision	0.4716
Recall	0.4724
F1-Score	0.472
AUC-ROC	0.4652

Even after tuning the hyperparameters of the XGBoost model using GridSearchCV and 5-fold cross-validation, the performance of the model remained well behind the mark. The metrics show minimal or perhaps no increase over the model not tuned. At that level of area under the curve (AUC-ROC) of 0.465 and accuracy level of 47.9%, the model can hardly be said to be significantly better, forming a guess much more than it makes a distinction between churn and non-churn. The observation may indicate that XGBoost is not suitable for this dataset or that the current set of features is not strong enough to make predictions.

*4.3.2 **4.3.2 LightGBM***

Table 3: LightGBM Performance (Without GridSearchCV and 5-Fold Cross-Validation)

Metric	Value
Accuracy	0.5034
Precision	0.4966
Recall	0.5017
F1-Score	0.4991
AUC-ROC	0.4904

Table 4: LightGBM Performance (With GridSearchCV and 5-Fold Cross-Validation)

Metric	Value
Accuracy	0.4855
Precision	0.4424
Recall	0.1655
F1-Score	0.2409

AUC-ROC	0.472

As the performance of the LightGBM model showed, applying GridSearchCV and the cross-validation in 5-folds to adjust the hyperparameters was not beneficial, as compared to the setting of default values. Recall was reduced to 16.5%, meaning a weak ability to correctly identify churners, and the optimized model demonstrated a drop in accuracy and F1-score on the whole. Probably, the tuned model has overfitted the training folds or used suboptimal parameters, whereas the untuned model had balanced yet low numbers. As seen in the results, it is essential to get it tuned well and integrate it with wonderful features. In the future, some modelling methods or improved feature engineering might be needed.

4.3.3 *4.3.3 Convolutional Neural Network (CNN)*

Table 5: CNN Model Performance

Metric	Value
Accuracy	0.4915
Precision	0.4741
Recall	0.2845
F1-Score	0.3556
AUC-ROC	0.4864

The CNN model did not reach an impressive score when tested on the task of churn prediction. Despite the near-random performance of 49.1 in accuracy, the recall is rather low at 28.4%, and this implies that the model did not capture a high volume of consumers who churned. There is also a lack of predictive capacity, as revealed in the F1-score and the AUC-ROC values. Convolutional neural networks (CNNs) are more appropriate when the data is flat and tabular as opposed to spatial or image. These results once again prove the idea that XGBoost or LightGBM are more suitable algorithms to be employed with structured data of such a nature as compared to CNN.

4.3.4 *4.3.4 Long Short-Term Memory (LSTM)*

Table 6: LSTM Model Performance

Metric	Value
Accuracy	0.506
Precision	0.4992
Recall	0.5224
F1-Score	0.5105
AUC-ROC	0.5108

The Long Short-Term Memory (LSTM) model produced the best scores among the models tested, with an F1-score of 0.5105 and AUC-ROC of 0.5108. LSTM model has a decent ability to identify the patterns of churn, with an accuracy of a bit over 50 percent and balanced recall and precision. The success of this fixed tabular data could have been due to the capability of LSTM to capture long-term dependencies, since this technique is typically used on sequential or time series data. However, the gains remain minuscule, pointing to the possibility that to attain improved abilities in prediction, more feature engineering may be necessary.

4.4 4.4 Evaluation

LSTM did better than all four models under examination as it had the highest F1-score (0.5105) and AUC-ROC (0.5108), and this may indicate that it was able to identify churn patterns better to some extent. Both LightGBM and XGBoost were performance disappointments without the help of GridSearchCV, and a slight improvement in hyperparameters only brought a slight improvement, whereas recall proved to be a significant obstacle. Since CNNs are not a suitable solution for tabular data, the worst results were provided by CNN. Despite the fact that LSTM yielded better results than the other algorithms, the scores were still poor to indicating that either a new algorithm, improved feature engineering, or a new approach is required to improve the accuracy of churn prediction.

5 Chapter 5: Conclusion and Recommendation

5.1 5.1 Conclusion

This paper aimed at applying four models of machine learning, namely, XGBoost, LightGBM, CNN, and LSTM, to predict e-commerce client attrition. Models were then evaluated based on measures such as recall, accuracy, precision, F1-score, and AUC-ROC after feature engineering and preprocessing. Even though the hyperparameter tuning helped to maximize the results of tree-based models, LSTM managed to win the race with a small margin. The deficient results of CNN also proved that it is not a good fit with tabular data. The data quality and modeling methodologies can be improved, as the research demonstrated that even though machine learning would allow obtaining information about churn prediction, model performance was still low.

5.2 5.2 Recommendation

These findings permit coming up with a number of recommendations. First, visual insights uncovered a decrease in churn in those categories that contain longer-term contracts, so e-commerce ventures ought to investigate them. Second, more attention should be paid to high-paying clients who might as well turn to sectors in else they do not feel well-appreciated. Technically, before the application of deep learning, to optimise feature value, it can be sometimes beneficial to apply simpler models as baselines, e.g., Logistic Regression or Decision Trees. In order to enhance accuracy in forecasts and interpretability of the model, more focus should be laid on data enrichment, such as behavioral history or consumer reaction.

6 Chapter 6: Future Work

To better align the upcoming research with the capabilities of LSTM and CNN models, it will be possible to include time-series or sequential consumer behavior data. Ensemble methods, either through stacking or mixing, can also be used to enhance performance on the whole. Two of the options of the dimensionality reduction and feature selection techniques that may be explored are Recursive Feature Elimination (RFE) and principal component analysis (PCA). Finally, understanding model decisions would be assisted by the application of XAI tools like SHAP or LIME, which in turn would allow better business planning and trust in automated systems. These changes may increase the predictive ability of the model and its application to a high degree.

7 References

Al Rahib, M.A., Saha, N., Mia, R. and Sattar, A., 2024. Customer data prediction and analysis in e-commerce using machine learning. *Bulletin of Electrical Engineering and Informatics*, *13*(4), pp.2624-2633.

Azevedo, A., 2019. Data mining and knowledge discovery in databases. In *Advanced methodologies and technologies in network architecture, mobile computing, and data analytics* (pp. 502-514). IGI Global Scientific Publishing.

Berger, P. and Kompan, M., 2019. User modeling for churn prediction in E-commerce. *IEEE Intelligent Systems*, *34*(2), pp.44-52.

De, S. and Prabu, P., 2022. Predicting customer churn: A systematic literature review. *Journal of Discrete Mathematical Sciences and Cryptography*, *25*(7), pp.1965-1985.

Hasan, M.S., Siam, M.A., Ahad, M.A., Hossain, M.N., Ridoy, M.H., Rabbi, M.N.S., Hossain, A. and Jakir, T., 2024. Predictive Analytics for Customer Retention: Machine Learning Models to Analyze and Mitigate Churn in E-Commerce Platforms. *Journal of Business and Management Studies*, *6*(4), pp.304-320.

Imani, M., 2025. Comparing Traditional Machine Learning and Advanced Gradient Boosting Techniques in Customer Churn Prediction: A Telecom Industry Case Study.

Imani, M., Joudaki, M., Beikmohamadi, A. and Arabnia, H.R., 2025. Customer Churn Prediction: A Review of Recent Advances, Trends, and Challenges in Conventional Machine Learning and Deep Learning.

Jahan, I. and Sanam, T.F., 2024. A comprehensive framework for customer retention in E-commerce using machine learning based on churn prediction, customer segmentation, and recommendation. *Electronic Commerce Research*, pp.1-44.

Li, X. and Li, Z., 2019. A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm. *Ingénierie des Systèmes d'Information*, *24*(5).

Ljupco Markuseski, L., Zdravkoski Igor, I.Z., Andonovski Miroslav, M.A. and Jovanoska Aleksandra, A.J., 2019, November. Knowledge discovery databases (kdd) process in data mining. In *INTERNATIONAL CONFERENCE PROCEEDING* (pp. 529-539). Faculty of Economic Prilep.

Mammadov, R., 2024. *Customer Churn Prediction.* [Online] Available at: https://www.kaggle.com/datasets/rashadrmammadov/customer-churn-dataset

Neeraj, M., Daniel, E., Durga, S. and Seetha, S., 2025, April. Explainable Multi-Stage Churn Prediction using Graph Neural Network in Telecom Sector. In *2025 8th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1100-1105). IEEE.

Pondel, M., Wuczyński, M., Gryncewicz, W., Łysik, Ł., Hernes, M., Rot, A. and Kozina, A., 2021, July. Deep learning for customer churn prediction in e-commerce decision support. In *Business Information Systems* (pp. 3-12).

Raeisi, S. and Sajedi, H., 2020, October. E-commerce customer churn prediction by gradient boosted trees. In *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)* (pp. 055-059). IEEE.

Shaker Reddy, P.C., Sucharitha, Y. and Vivekanand, A., 2024. Customer Churn Prevention For E-commerce Platforms using Machine Learning-based Business Intelligence. *Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering)*, *17*(5), pp.456-465.

Shikhli, M.S.E. and Hammad, A.M., 2018, June. Data Acquisition Model for Analyzing Schedule Delays Using KDD: Knowledge Discovery and Datamining. In *Proceedings of the 4th International Conference on Frontiers of Educational Technologies* (pp. 67-74).

Shobana, J., Gangadhar, C., Arora, R.K., Renjith, P.N., Bamini, J. and devidas Chincholkar, Y., 2023. E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors*, *27*, p.100728.

Ulas, B.T., Imer, S., Kalayci, T.A. and Asan, U., 2021, October. Modeling Customer Churn Behavior in E-commerce Using Bayesian Networks. In *Global Joint Conference on Industrial Engineering and Its Application Areas* (pp. 283-300). Cham: Springer International Publishing.

Vanneschi, L., Horn, D.M., Castelli, M. and Popovič, A., 2018. An artificial intelligence system for predicting customer default in e-commerce. *Expert Systems with Applications*, *104*, pp.1-21.

Win, T.T. and Bo, K.S., 2020, November. Predicting customer class using customer lifetime value with random forest algorithm. In *2020 International conference on advanced information technologies (ICAIT)* (pp. 236-241). IEEE.

Xia, G. and He, Q., 2018, March. The research of online shopping customer churn prediction based on integrated learning. In *2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018)* (pp. 259-267). Atlantis Press.

Yayun, Z.H.U.A.N.G., 2018. Research on E-commerce customer churn prediction based on improved value model and XG-boost algorithm. *Management Science and Engineering*, *12*(3), pp.51-56.

Zaghloul, M., Barakat, S. and Rezk, A., 2025. Enhancing customer retention in Online Retail through churn prediction: A hybrid RFM, K-means, and deep neural network approach. *Expert Systems with Applications*, *290*, p.12\