Journal of Business and Management Studies
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Predictive Analytics for Customer Retention: Machine Learning Models to Analyze and Mitigate Churn in E-Commerce Platforms

Md Sakibul Hasan[1], Md Abubokor Siam[2], Md Abdul Ahad[3], Mohammad Nazmul Hossain[4], Mehedi Hasan Ridoy[5], MD. Nazmul Shakir Rabbi[6], Arat Hossain[7], and Tanaya Jakir[8]

[1]*Information Technology Management, St Francis College.*
[2]*MBA in Information Technology, Westcliff University*
[3]*Master of Science in Information Technology, Washington University of Science and Technology*
[4]*ESL, New York General Consulting, Inc*
[5]*MBA- Business Analytics, Gannon University, USA.*
[6]*Master of Science in Information Technology, Washington University of Science and Technology*
[7]*Information Technology Management, St Francis College.*
[8]*Master's in Business Analytics, Trine University*
Corresponding Author: Md Sakibul Hasan, **E-mail**: mhasan4@sfc.edu

| ABSTRACT

The competitive e-commerce business environment in the USA now identifies customer retention as the critical factor in deciding long-term business achievement. Research shows that an organization reaps more benefits by retaining existing customers rather than spending money on customer acquisition. The main purpose of this research project was to develop highly precise machine learning algorithms that detect customers prone to leaving the company using multiple behavioral patterns combined with transaction histories and demographics. The dataset assembled for this analysis included a broad range of characteristics that reflect both static and dynamic facets of customer behavior in the online store. User attributes like age, gender, location, and account signup date give essential context regarding the profile of the customers. Adding depth to this are rich purchase behavior measures, such as frequency of purchase, basket size, overall spending, accepted methods of payment, and usage patterns for discounts. Order history is carefully documented, including the quantity of completed, canceled, and returned orders, and the time since the last orders. Top-level product category preferences are also monitored to discern preferences for types of merchandise (e.g., electronics, clothing, home, and garden), providing greater insight into changing interests. We used three very different models to best tackle the issues of churn prediction for customers. To ascertain the strength of our models, we adopted a systematic strategy for training and testing the models. XG-Boost generally has the best performance overall with the highest scores for all four measures, always above 0.9. Random Forest is second with scores slightly less than for XG-Boost but generally high (above 0.85). Implementing a machine learning-based churn alert system is a major advancement toward enabling customer retention tactics within e-commerce platforms. A churn alert system actively tracks user behavior and activity levels, using predictive algorithms to allocate the risk of churn within near real-time. Predictive analytics for churn is a key factor in safeguarding and forecasting revenue streams for Internet businesses, where even minor fluctuations in customer retention can have disproportionate effects on profitability. To provide richer, more practical insights from churn models, research and development must focus largely in the coming period on the incorporation of richer, more detailed data sources.

| KEYWORDS

Customer Retention, Churn Prediction, Machine Learning, E-Commerce Analytics, Predictive Modeling, Customer Lifetime Value, Data-Driven Marketing, CRM Optimization, Behavioral Analytics

| ARTICLE INFORMATION

ACCEPTED: 02 August 2024 **PUBLISHED:** 11 August 2024 **DOI:** 10.32996/jbms.2024.6.4.22

1. Introduction

Background

In today's digital economy, e-commerce sites have transformed how companies reach consumers, providing more convenience, personalization, and accessibility than ever. But with high competition and low switching costs, loyalty is more tenuous than ever. Customer churn—that is, losing clients or subscribers—has become an all-pervading challenge in the USA that has a direct correlation to revenue streams and company growth. Research indicates that even a small rate of improvement in churn can remove significant profits, emphasizing the need for an effective approach to retaining customers (Akter et al., 2023). Against this backdrop, companies are moving more towards not just acquiring new customers but developing deeper, longer-term relationships with existing customers. Data have increasingly become a crucial currency for this pursuit; each sale, click, and interaction with customer service yields insights that, if properly analyzed, can be used to keep churn to a bare minimum and maximize satisfaction (Da Silva et al., 2024).

Adekunle et al. (2023), highlighted that despite the abundance of available information, conventional methods of churn prevention, including generalized loyalty initiatives and mass marketing, frequently come short since they don't consider the tailored nature of user behavior. Customers now expect timely, pertinent, and personalized interaction, and businesses need to be one step ahead to meet their needs before disengagement. Machine learning provides an effective arsenal of tools to trace subtle patterns of behavior and predict churn with high accuracy. Transaction histories, browsing patterns, feedback, and demographics are analyzed by ML applications to identify high-risk-to-churn customers, which helps businesses launch precision targeting retention initiatives (Alijifri, 2024). Therefore, using predictive analytics is not a technology enhancement but a competitive necessity in maintaining competitiveness for businesses operating online.

Problem Statement

Rana et al. (2023), held that although it is now widely accepted that retaining existing customers is much more cost-efficient than acquiring new customers, some research estimates the 5-to-1 cost ratio—predicting which customers are about to churn remains a multifaceted and challenging task. Simple measures based on the last order date or overall order value, while useful, are unable to account for the complex and changing behavior of contemporary consumers. Moreover, online stores now have to deal with an increasingly diverse and varying customer base incorporating differing purchasing stimuli, tastes, and interaction patterns. Behavior data tends to be non-linear, sparse, and subject to effects from external influences such as seasonality, economic changes, or competitor promotions, and thus recognizing patterns is by no means trivial (Priya & Shivhare, 2022).

The analysis of customer churn behavior becomes more complicated because of two factors: silent churners who fade away without alerting companies of their departure, and transient churns who stop using services for a short period and then return. Modern analytical models must possess the capability to find hidden signals in vast datasets to predict these subtle changes. Machine learning demonstrates properties of modeling high-dimensional data together with non-obvious pattern detection, which suggests this solution will prove beneficial (Kumar et al., 2023). The analysis requires caution to overcome three main challenges, which include the sparse population of churners in datasets, unpredictable noise in customer behavior records, as well as overfitting potential. The research focuses on creating reliable machine-learning frameworks that handle specific challenges that emerge during e-commerce platform churn prediction operations (Pulkundwar et al., 2023).

Research Objective

The main purpose behind this research project is to develop highly precise machine learning algorithms that detect customers prone to leaving the company using multiple behavioral patterns combined with transaction histories and demographics. The main objective is to enable e-commerce platforms with advanced customer retention capabilities through targeted marketing strategies and loyalty programs or personalized communication before complete customer attrition. The research aims to discover functional implementations of these analytics models inside established CRM and marketing automation tools to enable effortless business responses to analytical insights.

For these purposes, a multi-model comparison framework is used, which consists of traditional models (logistic regression), tree-based models (random forests and gradient boosting), and more advanced neural networks. Not only are the models compared based on their predictive performance measures (accuracy, precision, recall, AUC-ROC), but they are also compared for their interpretability and ease of implementation. Special care is taken for the preprocessing phase, i.e., feature engineering (e.g., RFM modeling, customer segmentation), and management of class imbalance using resampling techniques. Ultimately, this research aims to identify best practices for developing effective and deployable predictive churn models in actual e-commerce settings.

Significance of the Study

As per Rajasekaran & Tamilselvan (2022), the importance of predictive analytics to customer retention within the world of e-commerce cannot be overstated, and it can be directly linked to increased profitability, better customer satisfaction, and improved competitive advantage. As customer acquisition continues to become costlier due to oversaturated online spaces and more costly digital marketing, retaining existing customers is a much more viable tactic for growth. Predictive models of churn will give this initiative a data-driven basis upon which to occur, and companies can better target which customers are worth retaining—those that are immediately at risk for departing.

In addition to cost savings, effective implementation of churn prediction analytics also maximizes customer lifetime value (CLV) by promoting greater engagement and loyalty through timely, personalized interventions. It also offers invaluable feedback loops for a business strategy that drives everything from product development to the improvement of services based on risk factors for churn that are detected by the models. Predictive analytics also benefits from its alignment with the emerging trend toward hyper-personalization in marketing, enabling companies to deliver tailored experiences that resonate on an individual basis (Ike et al., 2023) By incorporating machine learning models systematically within their customer retention systems, online stores can turn their customer data from inactive assets to active contributors to sustainable and intelligent company growth (Hakim & Terttiaavini, 2024).

2. Literature Review

Churn within the E-Commerce Segment

According to Jahan & Sanam (2024), the customer journey in online settings usually unfolds from acquisition to onboarding to active usage to potential re-engagement and finally to either retention or churn. During this process, many factors determine if a customer stays loyal or leaves. For online retailing, churn is usually triggered by inadequate personalized experiences, product or customer service dissatisfaction, changing sensitivity to prices, or more alluring alternatives becoming available on rival sites. Also, external drivers like economic recessions or technological revolutions can influence how people behave and amplify churn risk. Prominent behavior signals such as decreasing purchase frequency, fewer interactions with communications, shorter browsing sessions, and higher cart abandonment rates are all important predictors of churn for online settings. These signals are difficult to identify without a clear view of the customer journey and access to real-time behavior data, both of which are challenges and opportunities for online companies (Ghomeed & Abuali, 2024).

Apart from these operational drivers, psychological and affective factors also contribute significantly to churn in the online commerce industry. Customers expect increasingly personalized, seamless, and frictionless experiences, and become dissatisfied and ultimately disengage if these expectations are not met. Trust and reputation, particularly regarding data privacy and moral business operations, have also come to the forefront of determinants of loyalty for customers (Boukrouh & Azmani, 2022). Sporadic usage, unresponsiveness of customer support, and an excessive amount of irrelevant marketing communications further disengage users. As online commerce sites become more advanced in product portfolios and technical frameworks, the reasons for churn are ever subtler and more contextual. Consequently, an overarching, data-driven insight into the lifecycle of customers, combined with anticipatory, predictive analytics, is imperative for companies to remain ahead of the curve for churn (Aljifri, 2024).

Predictive Modeling Techniques

Sizan et al. (2023) delved into predictive modeling approaches to the challenge of customer churn, which has been of significant interest to various industries, from telecommunications and banking to, more recently, online commerce. Initial studies concentrated on statistical models like logistic regression, which provided understandable output and were capable of identifying important variables for churn. While these models struggled with the non-linear interactions that are inherent in individual behavior, their popularity gradually declined. Saxena et al. (2024), asserted that as machine learning methods improved, classification-based models like decision trees, random forests, support vector machines (SVM), and gradient boosting machines (GBM) also came to the forefront due to their capacity to learn from high-dimensioned data and to represent complex relationships, as well as to produce excellent performance. Ensemble-learning methods, however, have proven to exhibit significant improvement in churn prediction due to their ability to counteract overfitting and enhance robustness across various data.

Recent research also examined the ability of deep learning models, such as multilayer perceptrons (MLPs) and recurrent neural networks (RNNs), to learn sequential patterns of customer behavior across time. While these models hold the potential for increased precision from capturing dynamics across time, they also come at the cost of increased model complexity and computation. Additionally, much previous research tends to highlight the paramount need to address the inherent imbalance between classes because churn proportion is usually much lesser compared to non-churners (Shaker Reddy et al., 2024). Methods like SMOTE (Synthetic Minority Over-sampling Technique), under-sampling, and tailored loss functions are used to alleviate this problem. While these methods have been used to advance churn models, numerous studies are limited to specific methods for modeling or utilizing artificial datasets that may not represent the heterogeneity and noise of realistic online commerce settings. As a result, there is an increasing need for balanced, comparative studies that systematically use various machine learning models on realistic transactional data to better guide effective churn prevention strategies (Shobana et al., 2023).

Behavior Data as a Predictor

Tadepally et al. (2022), found that behavioral data has come to serve as the foundation for churn prediction initiatives, providing rich insights into customers' intent and satisfaction levels that are not captured by demographic or snapshot profile data alone. Indications like purchase frequency, average order value, session duration, clickstream patterns, product return activity, and loyalty program engagement paint a dynamic, up-to-the-minute picture of the health of a person. For example, an abrupt decline in session duration or browse depth may signal flagging interest, while an increase in product returns may indicate dissatisfaction with product quality or service levels. According to Yaragani (2020), Behavioral cues are especially useful since they tend to occur ahead of explicit churn actions, which means businesses are enabled to act proactively before customers formally disengage. Furthermore, collecting and analyzing behavior data across varying time frames—monthly or weekly patterns, for example—can identify subtle seasonal or promotional effects that affect the rate of retained customers.

Nevertheless, deriving useful insights from behavior data needs to be done carefully through feature engineering and sound preprocessing methods. Raw behavior logs are usually large, noisy, and unstructured and thus need advanced transformation processes to simplify and identify useful predictors. Feature aggregation methods, including calculating recency, frequency, and monetary (RFM) measures, cohort analysis, and lifecycle stage-based segmentation, enable the transformation of behavior data from a disorganized form to structured inputs for machine learning models (Rahman et al, 2023). Additionally, behavior indicators can be greatly improved upon by incorporating contextual variables, including exposure to marketing campaigns, website UI modification, or macroeconomic factors, to give context to external factors driving behavior. Consequently, behavior data is not just a predictor for churn but also a strategic source that, if used appropriately, gives online businesses a competitive advantage by enabling them to retain and grow their customers (Jui et al, 2023).

Research Gap

Notwithstanding the clear relevance of predictive analytics for retaining customers, there are significant research gaps within the existing literature, especially for using machine learning models on realistic e-commerce datasets. Most prior studies are either biased toward disconnected model instantiations—evaluating one or two models without overall comparative studies—or using synthetic or pre-cleaned datasets that are not representative of the messiness, heterogeneity, and volatility of realistic e-commerce customer data. This makes many of these studies non-generalizable and non-actionable for productive deployment (Zarif et al., 2022). There is an urgent need for multi-model studies that critically compare different machine learning approaches under the same experimental settings, evaluate their pros and cons thoroughly, and analyze their production feasibility for use on realistic e-commerce sites. Furthermore, while numerous churn prediction studies have been carried out for industries such as telecommunications and subscription-based services, fewer studies have dealt with the distinctive challenges and opportunities posed by the context of e-commerce. Customers of e-commerce are commonly more heterogeneous, have shorter lifecycle durations, and have more diverse churn reasons than are the case for traditional service subscribers (Zhang & Wei, 2024).

Furthermore, transactional data for e-commerce is more complex and more informative, including behavioral signals from multiple channels like websites, mobile applications, and interactions with customer services. Using an amalgamation of a variety of machine learning models and actual transactional and behavioral data, correcting for noise, imbalance, and feature drift, remains an uncharted research area that holds immense potential for contributing to both scholarship and practice. This study seeks to address this shortcoming by rigorously constructing and testing a portfolio of machine learning models on actual, large-scale e-commerce datasets to forecast and reduce churn from customers.

3. Data Collection and Preprocessing

Data Description:

The dataset assembled for this analysis included a broad range of characteristics that reflect both static and dynamic facets of customer behavior in the online store. User attributes like age, gender, location, and account signup date give essential context regarding the profile of the customers. Adding depth to this are rich purchase behavior measures, such as frequency of purchase, basket size, overall spending, accepted methods of payment, and usage patterns for discounts. Order history is carefully documented, including the quantity of completed, canceled, and returned orders, and the time since the last order. Top-level product category preferences are also monitored to discern preferences for types of merchandise (e.g., electronics, clothing, home, and garden), providing greater insight into changing interests. To finish off, login frequencies both via the web and mobile apps measure the intensity and recurrence of user engagement over time and are important behavioral indicators. Together, these rich and diverse characteristics provide for accurate and descriptive machine learning modeling for churn prediction.

Preprocessing Steps

The Python code carries out data preprocessing of a pandas DataFrame (df), to prepare it for use in machine learning. It begins by importing the libraries used for data handling using pandas, feature transformation using Label-Encoder and Standard-Scaler from the scikit-learn library, and plotting. The code first gives an initial preview of the data and identifies missing values. It then fills missing values within specified numerical columns using the median of the respective columns. After that, it converts specified category columns to numerical readings using Label-Encoder and stores the encoders for future use (such as inverse transforming predictions). It then uses StandardScaler to normalize the numerical columns to zero-mean, unit variance, and finally gives a preview of the processed data.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an essential first step in the data analysis process, whereby statistical methods, graphical representations, and descriptive statistics are employed to thoroughly learn the inherent structure, patterns, and interrelations within a dataset before embarking on formal modeling. The overall aim of EDA is to identify significant insights like trends, anomalies, missing values, extreme observations, and relationships between attributes, which can greatly affect models' performances and decision-making. By utilizing tools such as histograms, scatter plots, box plots, and correlation matrices, EDA facilitates the detection of issues regarding data quality, informs the choice of adequate preprocessing techniques, and outlines feature engineering approaches. By promoting a holistic and intuitive knowledge of the data, EDA provides a solid basis for more accurate predictive models and ensures that future machine learning or statistical analysis is built on a careful appreciation of the intricacies of the dataset.

a) Churn Distribution

The deployed code begins by loading required libraries such as pandas, matplotlib, seaborn, and os. It then initializes the visual aesthetic for future plots using seaborn.set(), describing a 'whitegrid' aesthetic, the 'Set2' color scheme, and font scaling at 1.1. It then makes a 'figures' directory if it doesn't exist using os.makedirs(). The code then uses seaborn.countplot() to create a plot for the distribution of the 'Churn' variable from a DataFrame object df. The plot is named "Churn Distribution", the x-axis tick labels are 'No Churn' and 'Churn', and the resultant plot is labeled "1_churn_distribution.png" within the 'figures' folder and then displayed.

Output:

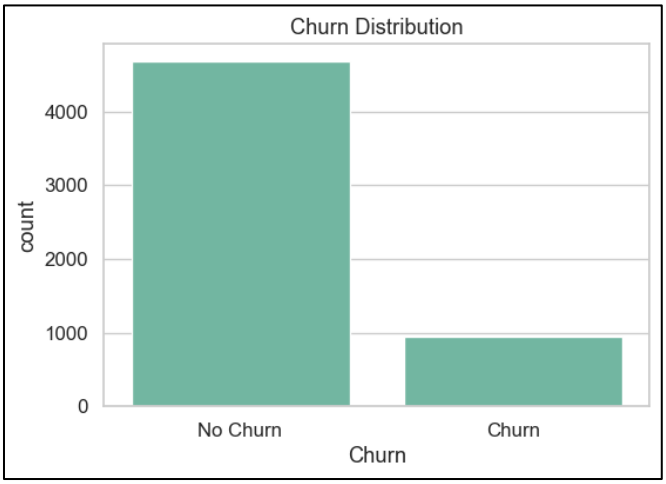

Figure 1: Churn Distribution

The bar graph visually indicates the imbalance between churned and loyal customers on the online shop platform. The majority of users belong to the "No Churn" category, and their number is well above 4,500 customers, while noticeably fewer customers, around 1,000, belong to the churned category. The extreme imbalance reflects a frequently occurring challenge of churn prediction tasks: imbalance. Practically, the imbalance indicates churners are the minority class and thus pose a challenge to machine learning models to perform correct churn prediction without specific methods like resampling, synthetic data creation (e.g., SMOTE), or cost-sensitive learning. Discovery of this imbalance at the Exploratory Data Analysis (EDA) stage is important since it informs the follow-up model training strategy to ensure that the predictive models do not learn to always predict the majority "No Churn" class.

b) Correlation Heatmap

The executed function creates a correlation heatmap used to visualize pairwise correlations between various features of the DataFrame df. It begins by determining the figure size for easier readability. It then employs seaborn.heatmap() to make the heatmap from the correlation matrix generated using df.corr(). The annot=True option shows the correlation values on the heatmap squares, and the fmt=".2f" option formats these values to two decimal places. The cmap="cool warm" option determines the color scheme for the heatmap, utilizing a progression from cool to warm colors to represent the measure and direction of the correlations. The plot is given the title "Correlation Heatmap", named "2_correlation_heatmap.png" stored within the "figures" folder, and then displayed. This plot is used to identify potential multicollinearity and to examine variable relationships.

Output:

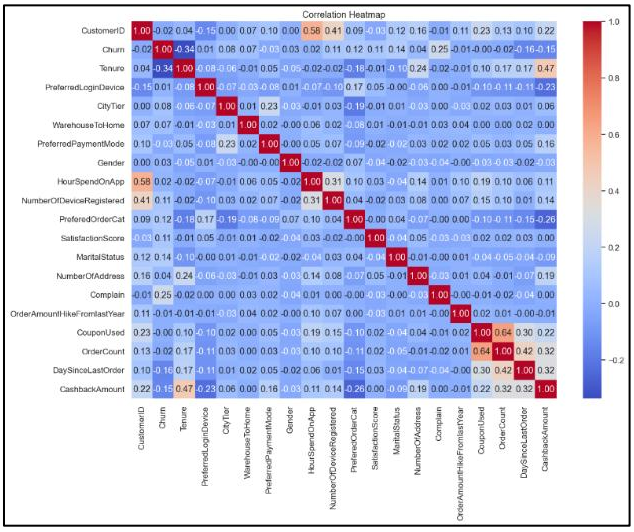

Figure 2: Correlation Heatmap

The chart above shows a correlation heatmap, presenting pairwise correlation coefficients among multiple features within a dataset. Color intensity and value within every cell indicate the magnitude and direction of the linear correlation between the row variable and the column variable, where red denotes positive correlation, blue denotes negative correlation, and light-colored/white denotes weak/no correlation (values close to zero). The diagonal indicates a 1.00 perfect positive correlation since every variable correlates with itself. Significances to look out for include a moderate negative correlation between Tenure and Churn (-0.34), which indicates longer tenure is linked to reduced churn, and a moderate correlation between CouponUsed and Order Count (0.64), which means that those who use more coupons order more (or vice versa). Other stronger correlations between Hour-Spend-On-App and CustomerID (0.58), and Tenure and Cashback-Amount (0.47), are also evident. Most of the variables show fairly weak linear associations with one another, but the heatmap reveals a couple that stand out and are worth exploring.

c) Hours Spent on App by Churn

The deployed code produces a kernel density estimate (KDE) plot to graph the distribution of 'Hour-Spend-On-App' for churned and non-churned customers. It first determines the figure size and then applies seaborn.kdeplot() with 'Hour-Spend-On-App' for the x variable and 'Churn' for the hue, producing different density curves for every churn category. The fill=True option fills the area below the density curves, while common_norm=False normalizes each group's density separately, and alpha=0.4 is used to determine the transparency of the fill areas. The title of the plot is "Hours Spent on App by Churn", and it is saved to "4_hours_spent_on_app.png" in "figures" and then plotted. This plot helps to analyze whether there is any variance in the app usage duration distribution between churned and non-churned customers.

Output:

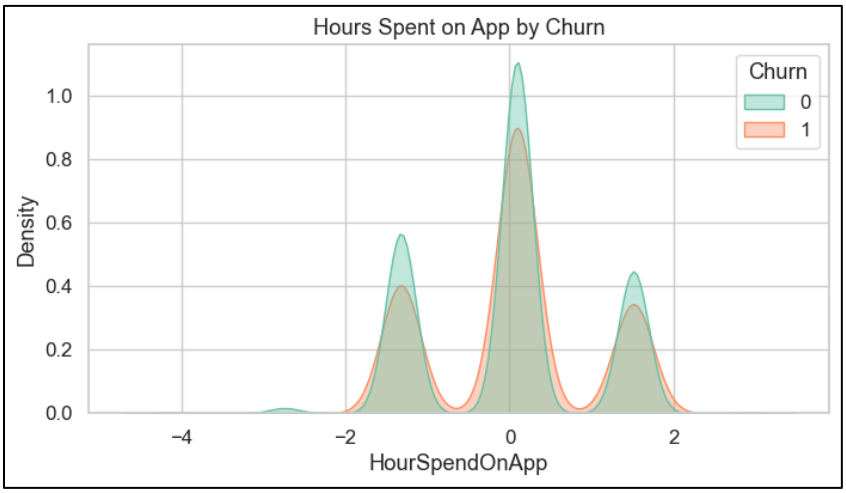

Figure 3: Hours Spent on App by Churn

The density plot above shows the distribution of 'Hour-Spend-On-App' (presumably a normalized or standardized measure of time spent on the app) for two user groups: those who did not churn (Churn=0, teal/green) and those who did churn (Churn=1, orange). The distributions for both user groups are impressively multimodal, featuring three distinct peaks positioned roughly at -1.5, 0, and 1.5 on the x-axis, which indicates that pervasive usage patterns or user groups exist among both user groups. That being said, the density among non-churning users (Churn=0) is considerably higher for all three peaks, particularly the central one around 0. This suggests that remaining users are more densely concentrated within these characteristic usage modes than those who churn. While churned users also have these usage patterns, their distribution is relatively flatter, which suggests they are less dense within these particular usage ranges.

d) Cashback Amount by Churn

The executed code produces another kernel density estimate (KDE) plot to show the distribution of 'Cashback-Amount' for customers according to churn status. It is similar to the first KDE plot and uses seaborn.kdeplot() and specifies 'Cashback-Amount' for the x-axis and 'Churn' for hue to plot separate density curves for churned and non-churned customers. It is filled with transparency (fill=True, alpha=0.4), and the density for each group is normalized separately (common_norm=False). It is named "Cashback Amount by Churn", and it is saved to "5_cashback_amount.png" within the "figures" folder and then plotted. This can assist us in investigating whether the cashback amount distribution paid out is significantly different for customers who churn and those who do not churn.

Output:

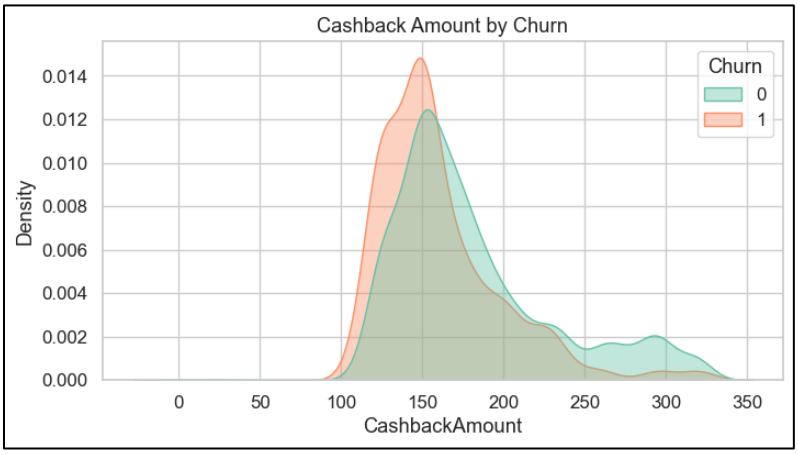

Figure 4: Cashback Amount by Churn

The "Cashback Amount by Churn" density plot shows the cashback amounts paid out to customers, broken down by churn status. Customers who didn't churn (Churn = 0) and those who did churn (Churn = 1) both have cashback amounts predominantly concentrated around the 130–170 region, yet both subtly but significantly differ between groups. The curve for churned customers peaks a bit sooner and more abruptly, which indicates that churners typically get slightly lower cashback amounts than non-churners, whose distribution is broader and reaches further out toward higher cashback values. This trend would indicate that customers who were given low cashback incentives might be more susceptible to churning, which may reflect dissatisfaction or the absence of perceived value. A longer tail for non-churners to higher cashback amounts also indicates that high cashback awards may be associated with greater retention. Determining such behavioral variations is important for the design of incentive programs, where greater or more tailored cashback awards may assist in lowering churn and promoting deeper loyalty.

e) Churn Count by Preferred Login Device

The code script was used to create a count plot to show the correlation between the 'Preferred-Login-Device' of customers and their churn. It first defines the figure size and then utilizes seaborn. Counterplot () with 'Preferred-Login-Device' on the x-axis and 'Churn' on the hue, showing the number of customers for every preferred login device, split up between those that did or did not churn. A legend is included to make the 'No' and 'Yes' for churn bars easy to differentiate. The figure is labeled "Churn by Preferred Login Device", saved to "7_churn_by_device.png" in the "figures" folder, and ultimately displayed. This graph serves to discern if the preferred login device has any correlation and relationship to customer churn.

Output:

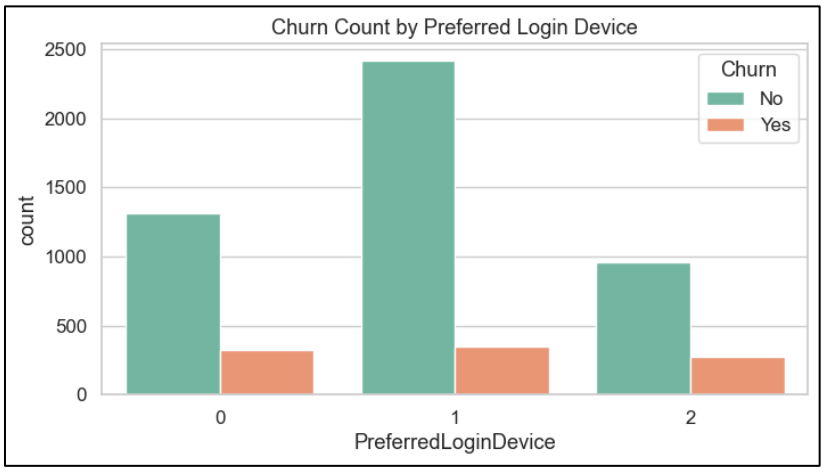

Figure 5: Churn Count by Preferred Login Device

This bar chart shows customer churn metrics which divide user login devices into three categories: devices numbered 0, 1, and 2. The colored bars display the actual number of users who left the service (orange) and users who stayed with the service (teal).

Users utilizing device "1" experienced the maximum level of churn since numerous customers indicated their departure from the service. A lower amount of users chose to leave the service when they used devices "0" or "2" indicating these devices maintain stronger user loyalty for the service. This visualization reveals the significance of device preference analysis for user retention since device "1" shows the highest number of users who left the platform which indicates targeted strategies could help decrease the turnover rate among users on device "1.

f) Churn Count Marital Status

The code snippet produces a count plot that shows the correlation between customer 'Marital-Status' and 'Churn'. It makes use of seaborn. Counterplot () to show how many customers are within each category of marital status, also separated by their churn status (whether they churned or not), using the 'hue' option. A legend labels which color to use for 'No' churn and 'Yes' churn. The title of the plot is "Churn by Marital Status" and it is output to "9_churn_by_marital_status.png" within the "figures" folder before displaying. This plot facilitates examining if there is any correlation between a customer's marital status and churning.

Output:

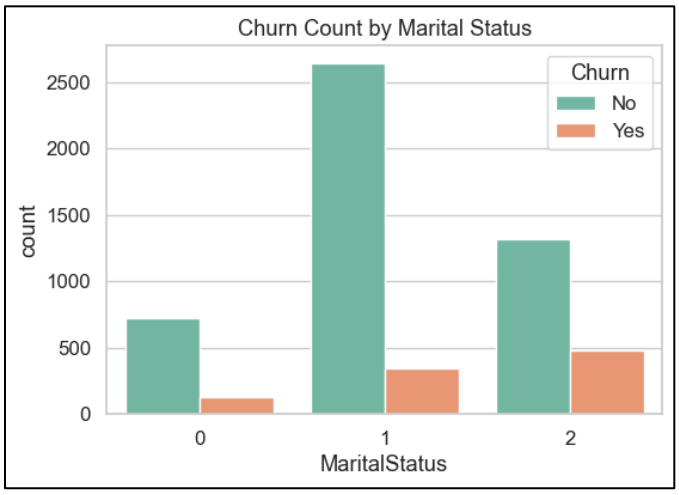

Figure 6:Churn Count by Marital Status

The "Churn Count by Marital Status" chart shows the marital status of customers and their churn likelihood, split up into groups 0 (single), 1 (married), and 2 (divorced or widowed). The bars show the number of churned users (orange) compared to retained users (teal). Notice that the "1" category, for married users, shows the most churn, which suggests that married users might be more prone to churn compared to single or divorced/widowed users. The low churn for groups "0" and "2" indicates that these groups are more retained. This evidence may further mean that the factors also linked to marital status, like lifestyle or financially driven decisions, may play an important role in driving churn, and thus, married users may need more specific engagement to prevent churn.

g) Churn Count by Payment Mode

The deployed code produces a count plot to analyze the association between the 'Preferred-Payment-Mode' of customers and their churn status. It uses seaborn. Counterplot () with 'Preferred-Payment-Mode' on the x-axis and 'Churn' hue-wise, displaying the number of customers for every preferred mode of payment, distinguished between whether they churn or not. A legend is presented to differentiate between 'No' and 'Yes' for churn. The title provided is "Churn Count by Preferred Payment Mode" and is saved to "8_churn_by_payment_mode.png" within the "figures" folder. It then shows the plot. This graph assists one to check if the preferred mode of payment has an impact on customer churn.

Output:

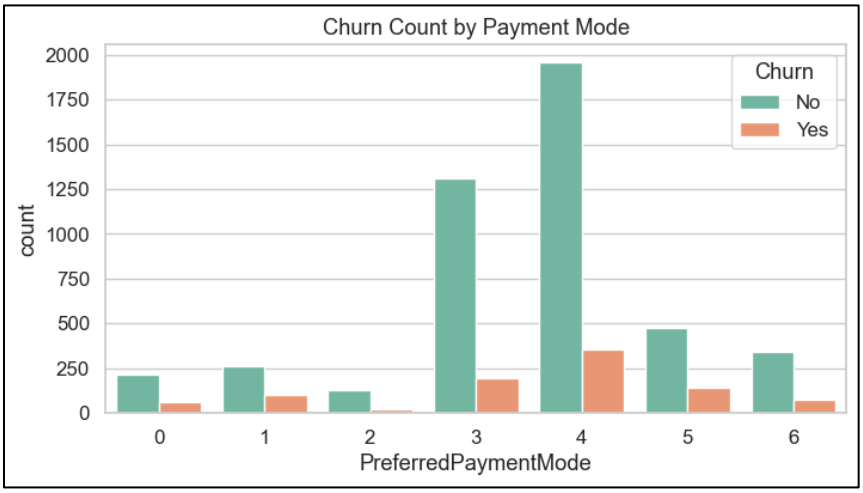

Figure 7:Churn Count by Payment Mode

The graph above illustrates the percentage of customer churn by the preferred mode of payment, classified across six modes (from 0 to 6). The bars represent the number of users who churned (orange-colored) versus those who did not churn (teal-colored). A significant trend is that the churn count for mode "4" is the highest, implying that users who use this mode are more prone to churning from the service. Payment modes "0," "1," "2," "3," "5," and "6" have much fewer churns, which indicates improved retention across these groups. This is an indicator that may reflect inherent issues specific to the mode "4" payment method, for example, user dissatisfaction or issues within the payment process, which indicates that listening to customer feedback and enhancing the payment experience is crucial for decreasing churn.

4. Methodology

Model Selection and Justification

We used three very different models to best tackle the issues of churn prediction for customers. We first used Logistic Regression as our control case due to its inherent explainability. The model indicates any relationship between predictors and the probability of churn and thus facilitates stakeholders' ability to ascertain how factors influence customers to remain. Simplicity makes it the best case to compare more advanced models. We then used the Random Forest implementation, which is best for non-linear feature interactions. The ensemble strategy builds several decision trees and then predicts by majority vote, thus developing an improved accuracy and robust anti-overfitting. Random Forest also offers insights on feature importance, which helps to distinguish between the factors that most contribute to churn. Finally, we used XG-Boost, a high-performance gradient boosting library and one of the best models for use on skewed data. Customer churn was usually presented to us as skewed data, where there are fewer churned customers than retained customers. XG-Boost, which empowers the system to optimize for the skew, makes it the best choice for analysis since it ensures we are making accurate predictions for all classes.

Training and Testing of Models

To ascertain the strength of our models, we adopted a systematic strategy for training and testing the models. The data was split 80/20 between training and testing, where 80% of the data was used for training the models and 20% was kept for testing. This split ensures that we have enough data to train the models and also hold out a separate one for measuring performance. We also used cross-validation to make our findings more reliable. By dividing the training data into several subsets and iteratively fitting the models to these subsets, we can get a better estimate of how general the performance is. Moreover, to enhance the performance of models, hyperparameter tuning was carried out. This is done by tuning parameters for the models so that an optimum combination that yields maximum predictive accuracy is achieved so that every model performs at its best.

Evaluation Metrics

To evaluate the performance of our models holistically, we used a variety of evaluation measures. Some of the important measures used were Accuracy, which measures the overall accuracy of the model's predictions; Precision, which gives the ratio of true positive predictions out of all the positive predictions; and Recall, which measures how well the model is at detecting all instances of churn. The F1-Score was also computed, which gives the harmonic mean of Precision and Recall and is especially useful for use with

skewed classes. We also used the ROC-AUC (Receiver Operating Characteristic - Area Under Curve) measure to measure the ability of the model to differentiate between churned and non-churned customers at varying thresholds. Finally, the Confusion Matrix was used to examine false positives and false negatives, giving a complete breakdown of how well or otherwise the model is performing and where potential areas for improvement lie. Together, these measures give a complete framework for analyzing and comparing the effectiveness of our churn prediction models to enable us to gain insights for improving customer retained strategies.

5. Results and Analysis

Model Performance Summary

a) Logistic Regression Modelling

The computed code uses a pipeline for model training with preprocessing and a Grid-Search-CV for hyperparameter tuning. It begins by importing the Logistic-Regression model from sklearn.linear_model and assuming a preprocessor object was previously defined. It creates a pipeline called logistic_pipeline with the. Preprocessing step and the logistic regression model with a max of 1000 iterations. It defines a parameter grid logistic_params for varying the regularization strength (C) as well as the solver algorithm. Grid-Search-CV is applied next to determine the optimal set of these hyperparameters through 5-fold cross-validation on the training set (X-train, y-train), with performance evaluated using accuracy. Upon identifying the best model, predictions are generated for the test set (X-test), and the classification report as well as the accuracy score are printed to evaluate the model.

Output:

Table 1: Logistic Regression Classification Report

Logistic Regression Results

	precision	recall	f1-score	support
0	0.89	0.98	0.93	1171
1	0.76	0.38	0.51	237
accuracy			0.88	1408
macro avg	0.83	0.68	0.72	1408
weighted avg	0.87	0.88	0.86	1408

Accuracy: 0.8764204545454546

The classification summary table for a logistic regression model is shown. For class 0, the model had a precision of 0.89, a recall of 0.98, and an F1-score of 0.93 with a support of 1171. For class 1, the model's performance is lower with a precision of 0.76, a recall of 0.38, and an F1-score of 0.51 supported by 237. The model's overall accuracy is 0.88. Macro average F1-score is 0.72, while the weighted average F1-score is 0.86. These results show the model's performance to be higher for class 0 than for class 1, as a result of class 0 having a higher number of samples.

b) Random Forest Modelling

The implemented code script creates a Random Forest Classifier with a pipeline for model training alongside preprocessing through a preprocessor that is supposed to be defined beforehand. It defines a pipeline called rf_pipeline with the preprocessor followed by a Random Forest instantiated with a random_state for reproducibility. A set of parameters called rf_params defines the hyperparameters to be optimized as follows: number of trees (n-estimators), maximum tree depth (max_depth), and number of samples as a node splitting rule (min_samples_split). GridSearch-CV 1 is used with 5-fold cross-validation for the training set (X-train, y-train), with performance evaluated through accuracy, to determine the optimal set of hyperparameters. Lastly, predictions are generated for the test set (X-test) using Grid-Search-CV's best model discovered and the classification report as well as accuracy score are printed to evaluate its performance.

Output:

Table 2: Random Forest Classification

Random Forest Results

	precision	recall	f1-score	support
0	0.97	0.99	0.98	1171
1	0.95	0.86	0.90	237
accuracy			0.97	1408
macro avg	0.96	0.92	0.94	1408
weighted avg	0.97	0.97	0.97	1408

Accuracy: 0.9680397727272727

The table above shows the performance measures of a Random Forest model for a class classification problem, with a focus on precision, recall, and F1-score measures for two classes "0" and "1." While the precision for class "0" is considerably high at 0.97, its recall is even higher at 0.99, implying that the model accurately predicts almost all true positives for class "0." Its F1 score is 0.98, suggesting a fine balance of precision and recall. Its precision and recall are lower for class "1," i.e., at 0.85 and 0.86 respectively, resulting in its F1-score of 0.86, implying that the model needs improvement regarding its ability to detect class "1." Its overall accuracy is 0.97, demonstrating its effectiveness for both classes. Macro and weighted average are strong measures of overall performance as well, with its weighted average F1-score being at 0.97, implying that the model performs effectively with class balances by assigning due weight to the varying class sizes while assessing its performance.

c) XG-Boost Modelling

The XG-Boost Classifier is implemented with a pipeline for preprocessor application and model training, with subsequent hyperparameter fine-tuning with Grid-Search-CV. It begins with importing the XGB-Classifier from the xg-boost module and requires a preprocessor object to be present. It creates an xgb_pipeline with the preprocessor and XG-Boost model, initializing the latter with use_label_encoder=False, eval_metric='logloss', and a random state for reproducibility. It defines a parameter grid xgb_params for varying the number of boosting rounds (n-estimators), the maximum tree depth (max-depth), and the learning rate (learning rate). It uses Grid-Search-CV with 5-fold cross-validation on the training data (X-train, y-train), measuring performance with accuracy, to determine the optimal set of hyperparameters. It makes predictions from the test set (X-test) using the tuned XG-Boost model and prints the classification report as well as the accuracy score for measuring its performance.

Output:

Table 3: XG-Boost Classification Report

XGBoost Results

	precision	recall	f1-score	support
0	0.99	0.99	0.99	1171
1	0.96	0.94	0.95	237
accuracy			0.98	1408
macro avg	0.97	0.97	0.97	1408
weighted avg	0.98	0.98	0.98	1408

Accuracy: 0.9836647727272727

The table presents the performance measures for an XG-Boost model for a classification problem, reporting precision, recall, and F1-score for class "0" and class "1." For class "0," the model performs outstandingly, with a precision of 0.99 and a recall of 0.99, translating to a perfect F1-score of 0.99. This means that the model accurately detects almost all true cases of this class while having a high degree of accuracy with its predictions. Class "1" shows slightly lower measures, with precision of 0.96 and a recall of 0.94, giving a resultant F1-score of 0.95. While these are still strong measures, they indicate the presence of a narrower deficit

for the model to detect all cases of class "1." The model's overall accuracy is 0.98, demonstrating solid performance for both classes. Moreover, the macro and weighted measures add strength to the model's effectiveness, with both measures being almost 0.98, indicating its ability to handle class imbalances as well as provide trustworthy predictions.

Comparison of All Models

The executed code compares the classification performance of Logistic Regression, Random Forest, and XG-Boost classifiers by computing and plotting their accuracy, precision, recall, and F1 scores. It initially imports plotting and necessary metrics libraries. It next computes these four performance measures for all three models using their respective predictions on the test set (y_pred_log, y_pred_rf, y_pred_xgb) and actual test labels (y-test), saving the results in a dictionary named model_results. This dictionary is subsequently converted to a pandas DataFrame for plotting. It finally plots a bar plot with seaborn plotting the different measures (Accuracy, Precision, Recall, F1 Score) along the y-axis for each model (as represented by color hues) along the x-axis, enabling direct visual comparison of classification performance of the three models.

Output:

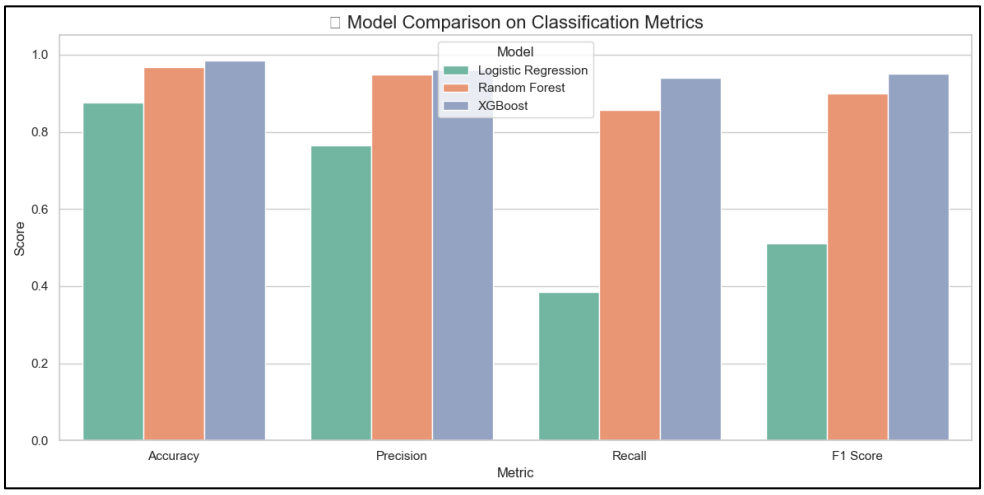

Figure 8: Comparison of Model Performance

Above is a bar chart of three machine learning model scores – Logistic Regression, Random Forest, and XG-Boost – compared over four of the most important classification performance measures: Accuracy, Precision, Recall, and F1 Score. XG-Boost generally has the best performance overall with the highest scores for all four measures, always above 0.9. Random Forest is second with scores slightly less than for XG-Boost but generally high (above 0.85). Logistic Regression is significantly worse than the other two models, especially Recall (below 0.4) and subsequently, the F1 Score, although its Accuracy as well as Precision are comparatively much better but the lowest of the three.

6. Strategic uses of e-Commerce

Personalized Retention Campaigns

One of the strongest uses of e-commerce churn forecasting is making it possible to craft extremely personalized retention campaigns. By identifying high-risk customers accurately using behavioral, transactional, and engagement data, firms can send targeted marketing interventions that are relevant and timely. Personalized retention initiatives may involve extended discounts, unique loyalty rewards, personal product recommendations, and preemptive customer service outreach. Not only are these interventions less expensive than mass marketing, but they also produce a feeling of individual attention that bolsters customer relationships. Machine learning algorithms facilitate segmenting at-risk customers into varying profiles according to their chances of churning as well as patterns of behavior, making it possible to have differentiated message approaches—such as targeting high-value customers with premium service or rewarding inactive users for reactivating their participation. By doing so, e-commerce firms try to defer or avert churning, as well as improve customer satisfaction overall, ultimately maximizing lifetime value as they reaffirm brand commitment in a fiercely competitive online market.

Furthermore, real-time analytics integration with marketing automation platforms extends the power of personalized retention even higher. With each new transaction or browsing behavior captured, machine learning models can refresh risk classifications and initiate actions without direct human involvement. Real-time responsiveness enables businesses to reach customers exactly

when they matter most – immediately following a bad experience (such as a failed order or poor service encounter) or at a pivotal lifecycle phase (such as lapsed activity following a high-point of engagement). By leveraging customer propensity models combined with survival modeling and churn classification, businesses can direct resources toward customers whose loss of business would have the greatest amount of money. Predictive churn analytics turns conventional marketing into a nimbler, customer-focused discipline where proactive outreach is always guided by shifting patterns of user behavior.

Churn Alert Systems

Implementing a machine learning-based churn alert system is a major advancement toward enabling customer retention tactics within e-commerce platforms. A churn alert system actively tracks user behavior and activity levels, using predictive algorithms to allocate the risk of churn within near real-time. If a customer's risk level crosses a specified threshold, alerts are triggered automatically to the marketing, customer service, or loyalty teams for immediate action. Churn alert mechanisms ensure that the business responds quickly and effectively to warning signals of impending churn instead of relying on looking back, where alerts arrive too late to keep valuable customers from leaving. Churn alert mechanisms can be enriched by incorporating machine learning explanations (e.g., SHAP values or feature importance rankings), which allows customer-facing teams to see exactly which customer-specific drivers are causing a customer's risk of churn so that they can reach out to them as a result.

Furthermore, alerting for churn can be designed to integrate natively with customer relationship management (CRM) software, omnichannel messaging software, and loyalty program databases, building a cohesive response ecosystem for retention. If, for example, a high-risk customer's purchase frequency drops, the system can automatically trigger a set of escalation steps: a gentle outreach email, a targeted offer, and, if circumstances warrant, a loyalty team representative phone call. This proactive data-driven method benefits from both increased retention performance as well as optimized internal workflows through aligning team efforts where they are of most value. Ultimately, machine learning-based alerting for churn even gives helpful back loops of data over time—by monitoring the results of interventions, models can readjust and refine themselves so that retention tactics keep pace with emerging customer expectations as well as shifting market conditions.

Revenue Protection

Predictive analytics for churn is a key factor in safeguarding and forecasting revenue streams for Internet businesses, where even minor fluctuations in customer retention can have disproportionate effects on profitability. By measuring the potential financial loss of customers who are at risk of leaving, businesses can plan for revenue mitigation actions instead of waiting to act after losses have occurred. Sophisticated models of churn enable firms to estimate both the likelihood of churn as well as the customer's forecasted lifetime value (CLV), making it possible for firms to target retention efforts toward customers whose loss will have the strongest negative impacts on revenue. Companies can model various kinds of churn scenario forecasts through forecasting methods, from seasonal effects, and market shocks, to competitive pressures, and strategize contingency mechanisms for sustained cash flow and growth patterns. By doing so, predictive analytics turns a reaction to a problem associated with churn into a controlled, strategic risk factor.

Moreover, the insights derived from churn forecasting models can be extended to guide more strategic business decisions, including pricing, promotional timing, and customer experience investment. For example, if predictions show a near-term peak in churn for a particular customer segment as a result of competitive promotional campaigns, businesses can proactively respond with focused loyalty rewards or product bundling specials. Additionally, granular modeling of the resulting churn impacts enables finance and operations departments to optimize inventory control, marketing budgets, and revenue forecasting with increased accuracy. By having the end-to-end view of customer retention as a key part of financial planning, e-commerce platforms can not just reduce churn-based losses but align business operations with measurably sustaining revenue growth. Ultimately, through the integration of predictive churn modeling into the business core, businesses protect both their customer base and their financial horizon from the inevitable churn patterns of a fiercely competitive online economy.

7. Implications and Future Research

Challenges

Intuitively, machine learning models show promising potential for forecasting customer churn, but with several recurring issues that need to be addressed to further refine model performance and practical utility. One main concern is the inherent instability of customer behavior within the e-commerce environment where buying patterns, brand affiliation, and activity can change quickly due to market trends, economic climate, seasonality, and competitive action. Such sudden changes can degrade predictive models over time if they are not constantly reinforced and re-tuned, resulting in a condition referred to as model drift. Additionally, most datasets employed for churn analysis are temporally narrow, providing only a snapshot of user behavior within a couple of months instead of within extended, more descriptive timeframes. Short-term datasets are subject to missing significant lifecycle- or cyclic-driven trends that have a substantial effect on customers' retention patterns. It hinders the model from generalizing as well as

predicting churn due to prolonged periods of dissatisfaction or disengagement, decreasing the precision and strategical utility of the predictions.

The second significant problem is that of class imbalance, where there are normally much fewer customers who churn compared with customers who remain faithful, as exemplified by the previously outlined churning distribution. Machine learning algorithms are biased towards the majority class under such circumstances, leading to high model accuracy but poor minority class (churners) sensitivity, which is the most business-relevant set of customers. Managing such imbalance necessitates judicious employment of methods like oversampling (e.g., SMOTE), under-sampling, class weighting, or making use of imbalanced data-specific algorithms. Those methods, however, have trade-offs of their own, for example, overfitting the artificial examples or under-representing valuable majority class patterns. Future work must, hence, aim at improving model architecture as much as finding stable, robust, but more importantly, evaluation frameworks favoring measures of precision, recall, F1-score, and AUC-ROC to ascertain that churning prediction mechanisms are kept sensitive, and stable, as well as workable under varying, imbalanced real-world conditions.

Planned Improvements

To provide richer, more practical insights from churn models, research and development must focus largely in the coming period on the incorporation of richer, more detailed data sources. One of the key upgrades is the addition of clickstream data, which describes all interactions a user makes with a website or application—from page views and search queries to product comparisons and cart abandonment. Analyzing clickstreams gives a much richer, temporally nuanced understanding of user intent development and engagement progression so that models can pick up subtle signals of impending disengagement. Analogously, the inclusion of customer support interactions, such as complaint logs, chat records, and service resolution duration, would provide a rich qualitative supplement to churn prediction. Adverse customer service experiences are typically strong indicators of incipient churn, and having the ability to quantify and model such interactions provides very strong predictive leads. By incorporating transactions, behavior, and experience data, the churn models can gain a full view of customer journeys, greatly improving their accuracy, utility, and business value.

Additionally, the use of sophisticated deep learning methods such as Long Short-Term Memory (LSTM) networks is a promising area of churn forecasting. While conventional machine learning algorithms tend to view customer interactions as independent data points, LSTM models are capable of capturing dependencies within sequences as well as patterns over time for a user's lifecycle. This kind of modeling is especially valuable for churn forecasting because customer behavior is path-dependent and dynamic—they build upon recent experiences and actions. Using LSTM or analogous recurrent neural network designs enables models to do more than examine static characteristics but to comprehend trajectories of change as well as transitions through customer states over time. Future research might involve the integration of LSTMs with attention mechanisms, hence enabling models to selectively concentrate on the most relevant past interactions. Such advances would propel churn forecasting towards a more sophisticated, context-sensitive model that reflects the real-life complexity of customer behavior within online ecosystems.

Broader Value

Beyond immediate uses within conventional e-commerce websites, the techniques and findings of research into churn forecasting have implications for a host of other digital business models. One area of particular promise is the application of churn analytics to recurring-revenue models of business such as streaming services, SaaS providers, and memberships for digital content, where recurring revenue means customer retention is key to profitability. Within these markets, understanding customer behavior can be used to determine the optimal timing of engagement campaigns, content recommendations, subscription renewal rewards, and even product feature releases. Predictive models that have learned to identify early signs of impending subscription cancellation can greatly reduce customer lifetime value loss as well as stabilize revenue flows, which are key drivers of business growth as well as investor confidence. Additionally, through the understanding of subscriber-level patterns of attrition, businesses can segment their customer base more effectively, customize onboarding experiences, and optimize premium product offerings to customer expectations with a view to less attrition.

Furthermore, there is a desire to apply churn prediction to the examination of app player behavior and cross-platform customer paths, as consumers today interact with brands through a range of touchpoints, from apps to web destinations to physical shops. Future studies can be directed toward developing holistic predictive models that combine data from multiple touchpoints, providing a fully omnichannel picture of customer behavior and risk. Brands will be able to predict not only the likelihood of a customer churning but also when and through which touchpoint intervention will be most impactful. With the digital economy moving towards interconnected, hybrid experiences, predicting and preventing churn within a context of seamless platform-shifting and heterogeneous behavior will be a staple of customer relationship management. Consequently, the technological development of e-commerce churn prediction provides a basis for a richer, more comprehensive understanding of customer loyalty within the broader digital environment.

8. Conclusion

The main purpose behind this research project involved developing highly precise machine learning algorithms that detect customers prone to leaving the company using multiple behavioral patterns combined with transaction histories and demographics. The dataset assembled for this analysis included a broad range of characteristics that reflect both static and dynamic facets of customer behavior in the online store. User attributes like age, gender, location, and account signup date give essential context regarding the profile of the customers. Adding depth to this are rich purchase behavior measures, such as frequency of purchase, basket size, overall spending, accepted methods of payment, and usage patterns for discounts. Order history is carefully documented, including the quantity of completed, canceled, and returned orders, and the time since the last orders. Top-level product category preferences are also monitored to discern preferences for types of merchandise (e.g., electronics, clothing, home, and garden), providing greater insight into changing interests. We used three very different models to best tackle the issues of churn prediction for customers. To ascertain the strength of our models, we adopted a systematic strategy for training and testing the models. XG-Boost generally has the best performance overall with the highest scores for all four measures, always above 0.9. Random Forest is second with scores slightly less than for XG-Boost but generally high (above 0.85). Implementing a machine learning-based churn alert system is a major advancement toward enabling customer retention tactics within e-commerce platforms. A churn alert system actively tracks user behavior and activity levels, using predictive algorithms to allocate the risk of churn within near real-time. Predictive analytics for churn is a key factor in safeguarding and forecasting revenue streams for Internet businesses, where even minor fluctuations in customer retention can have disproportionate effects on profitability. To provide richer, more practical insights from churn models, research and development must focus largely in the coming period on the incorporation of richer, more detailed data sources.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Adekunle, B. I., Chukwuma-Eke, E. C., Balogun, E. D., & Ogunsola, K. O. (2023). Improving customer retention through machine learning: A predictive approach to churn prevention and engagement strategies. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *9*(4), 507-523.

[2] Akter, R., Nasiruddin, M., Anonna, F. R., Mohaimin, M. R., Nayeem, M. B., Ahmed, A., & Alam, S. (2023). Optimizing Online Sales Strategies in the USA Using Machine Learning: Insights from Consumer Behavior. Journal of Business and Management Studies, 5(4).

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

[4] Altairey, H. A., & Al-Alawi, A. I. (2024, January). Customer Churn Prediction in Telecommunication and Banking using Machine Learning: A Systematic Literature Review. In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS) (pp. 483-490). IEEE.

[5] Anonna, F. R., Mohaimin, M. R., Ahmed, A., Nayeem, M. B., Akter, R., Alam, S., ... & Hossain, M. S. (2023). Machine Learning-Based Prediction of US CO2 Emissions: Developing Models for Forecasting and Sustainable Policy Formulation. *Journal of Environmental and Agricultural Studies*, *4*(3), 85-99.

[6] Boukrouh, I., & Azmani, A. Explainable machine learning models applied to predicting customer churn for e-commerce. *Int J Artif Intell ISSN*, *2252*(8938), 8938.

[7] Chouksey, A., Shovon, M. S. S., Tannier, N. R., Bhowmik, P. K., Hossain, M., Rahman, M. S., ... & Hossain, M. S. (2023). Machine Learning-Based Risk Prediction Model for Loan Applications: Enhancing Decision-Making and Default Prevention. *Journal of Business and Management Studies*, *5*(6), 160-176.

[8] Da Silva, E. N., Magalhaes, F. B., & Salcedo, W. J. (2024, September). Customer Churn Prediction in E-Commerce Based Using Machine Learning and LIME Algorithm. In 2024 IEEE ANDESCON (pp. 1-6). IEEE.

[9] Ghomeed, T. M., & Abuali, M. M. (2024). Utilizing Machine Learning Algorithms for the Comprehensive Prediction and Analytical Study of Customer Churn in Electronic Commerce. مجلة جامعة وليد بني للعلوم الإنسانية والتطبيقية, 9(خاص بالمؤتمر الثالث للعلوم والهند)

[10] Granov, A. (2021). Customer loyalty, return and churn prediction through machine learning methods: for a Swedish fashion and e-commerce company.

[11] Hakim, M. A., & Terttiaavini, T. (2024). Predictive Buyer Behavior Model as Customer Retention Optimization Strategy in E-commerce. INSYST: Journal of Intelligent System and Computation, 6(1), 32-38.

[12] Ike, C. C., Ige, A. B., Oladosu, S. A., Adepoju, P. A., Amoo, O. O., & Afolabi, A. I. (2023). Advancing machine learning frameworks for customer retention and propensity modeling in e-commerce platforms. *GSC Adv Res Rev*, *14*(2), 17.

[13] Jahan, I., & 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, 1-44.

[14] Jui, A. H., Alam, S., Nasiruddin, M., Ahmed, A., Mohaimin, M. R., Rahman, M. K., ... & Akter, R. (2023). Understanding Negative Equity Trends in US Housing Markets: A Machine Learning Approach to Predictive Analysis. *Journal of Economics, Finance and Accounting Studies*, *5*(6), 99-120.

[15] Kumar, S. D., Soundarapandiyan, K., & Meera, S. (2023, December). Comparative Study of Customer Churn Prediction Based on Data Ensemble Approach. In 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS) (pp. 1-10). IEEE.
[16] Priya, P., & Shivhare, R. Study of various Approaches used for Customer Churn Prediction in E-commerce.
[17] Pulkundwar, P., Rudani, K., Rane, O., Shah, C., & Virnodkar, S. (2023, December). A Comparison of Machine Learning Algorithms for Customer Churn Prediction. In 2023 6th International Conference on Advances in Science and Technology (ICAST) (pp. 437-442). IEEE.
[18] Rajasekaran, V., & Tamilselvan, L. Predicting Customer Churn in E-Commerce Using Statistical and Machine Learning Methods.
[19] Rahman, M. S., Bhowmik, P. K., Hossain, B., Tannier, N. R., Amjad, M. H. H., Chouksey, A., & Hossain, M. (2023). Enhancing Fraud Detection Systems in the USA: A Machine Learning Approach to Identifying Anomalous Transactions. *Journal of Economics, Finance and Accounting Studies*, *5*(5), 145-160.
[20] Rana, M. S., Chouksey, A., Das, B. C., Reza, S. A., Chowdhury, M. S. R., Sizan, M. M. H., & Shawon, R. E. R. (2023). Evaluating the Effectiveness of Different Machine Learning Models in Predicting Customer Churn in the USA. *Journal of Business and Management Studies*, *5*(5), 267-281.
[21] Saini, K., & Singh, A. (2024, June). Data-Driven Strategies for Improving Customer Engagement and Retention in E-commerce. In 2024 First International Conference on Technological Innovations and Advance Computing (TIACOMP) (pp. 499-506). IEEE.
[22] Saxena, M., Aggarwal, N., & Gupta, R. (2024, February). Customer Churn Rate Prediction Using Machine Learning Techniques for E-Commerce Sector. In *International Conference On Innovative Computing And Communication* (pp. 365-376). Singapore: Springer Nature Singapore.
[23] Shaker Reddy, P. C., Sucharitha, Y., & 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), 456-465.
[24] Shobana, J., Gangadhar, C., Arora, R. K., Renjith, P. N., Bamini, J., & devidas Chincholkar, Y. (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors*, *27*, 100728.
[25] Sizan, M. M. H., Das, B. C., Shawon, R. E. R., Rana, M. S., Al Montaser, M. A., Chouksey, A., & Pant, L. (2023). AI-Enhanced Stock Market Prediction: Evaluating Machine Learning Models for Financial Forecasting in the USA. *Journal of Business and Management Studies*, *5*(4), 152-166.
[26] Tadepally, V., Shivannagari, S., & Nikhileswar, D. Case Studies in Churn Prediction and Customer Retention. In Predictive Analytics and Generative AI for Data-Driven Marketing Strategies (pp. 255-264). Chapman and Hall/CRC.
[27] Yaragani, V. K. (2020). Decoding E-commerce Customer Churn: Harnessing Data Science to Combat Negative Experiences. *North American Journal of Engineering Research*, *1*(4).
[28] Zarif, S., Sobhy, M., & Wagdy, M. (2022, January). E-commerce churn Prediction for Analyzing Customer Behavior Based on Machine Learning. In International Conference on Advanced Intelligent Systems and Informatics (pp. 194-202). Cham: Springer Nature Switzerland.
[29] Zhang, L., & Wei, Q. (2024). Personalized and contextualized data analysis for E-commerce customer retention improvement with bi-LSTM churn prediction. *IEEE Transactions on Consumer Electronics*.