Customer Churn Prediction in E-Commerce

Armaan Haque

Department of Industrial and Systems Engineering

University at Buffalo

April 2025

Abstract

Customer churn remains a critical challenge for e-commerce businesses, as retaining existing customers is significantly more cost-effective than acquiring new ones. This study focuses on using logistic regression, decision trees, XGBoosting and random forest model to identify customers who are likely to churn, enabling businesses to take proactive measures to retain them. By analyzing behavioral, transactional, and demographic factors, the research aims to uncover the primary drivers of churn and provide actionable insights to improve customer retention strategies.

Keywords: Machine Learning, Churn Analysis, E-Commerce, Customer Sentiment.

1. Introduction

In today's digital economy, e-commerce businesses face intense competition, making customer retention a key factor in sustaining growth and profitability. **Customer churn**, defined as the rate at which customers stop using a service or making purchases, poses a significant challenge, as studies indicate that acquiring a new customer is 5 to 25 times more expensive than retaining an existing one (Kotler and Keller, 2023). High churn rates not only impact revenue and profitability but also increase customer acquisition costs (CAC) and reduce customer lifetime value (CLV).

Understanding why customers leave and predicting which customers are likely to churn is essential for e-commerce businesses to implement targeted retention strategies. Traditionally, businesses relied on rule-based models or basic statistical techniques, such as Recency-Frequency-Monetary (RFM) analysis and logistic regression, to predict churn. However, these methods have limitations, such as inability to capture complex behavioral patterns, reliance on predefined thresholds, and poor adaptability to changing customer behavior.

Recent advancements in machine learning (ML) have significantly improved churn prediction by leveraging historical purchase data, user engagement metrics, and sentiment analysis to identify customers at risk of leaving. While some studies have employed deep learning models, their complexity, high computational cost, and lack of interpretability make them difficult to implement for real-world e-commerce applications.

Customer churn prediction is vital because it directly impacts business revenue and growth by highlighting customers at risk of leaving. A high churn rate can signify problems in product quality, service, pricing, or overall customer experience, making churn prediction essential for diagnosing and addressing such issuesCustomer churn prediction is vital because it directly impacts business revenue and growth by highlighting customers at risk of leaving. A high churn rate can signify problems in product quality, service, pricing, or overall customer experience, making churn prediction essential for diagnosing and addressing such issues(Vio, 2023). Machine learning significantly enhances churn prediction accuracy by analyzing large and complex datasets to uncover subtle patterns and risk factors that traditional methods might miss. Techniques like ensemble models improve robustness by combining multiple predictive algorithms to increase predictive stability and reduce overfitting risks. The use of machine learning supports proactive marketing automation, targeted loyalty programs, and personalized communication, all of which contribute to reducing churn and driving higher customer lifetime value (Shobana et al., 2023).

2. Literature Review

Customer churn directly impacts an organization's revenue, marketing expenses, and operational strategies. Studies suggest that retaining an existing customer is more cost-effective than acquiring new ones (Kotler and Keller, 2023).

Early works in this domain defined churn as the termination of a customer's relationship with a business, emphasizing that effective prediction could enable timely and tailored retention strategies (Zhao, 2014; Gupta and Krishna, 2019). Researchers such as Zhao Zhao (2014) and Seema et al. (Seema et al., 2019; have reviewed and compared statistical analysis techniques with emerging artificial intelligence methods. While traditional statistical models provided an initial framework for churn prediction, their limitations in handling high-dimensional and dynamic e commerce data led to the integration of more advanced machine learning algorithms (Gupta and Krishna, 2019; Fu, 2022). This transition to intelligent systems has permitted the processing of multidimensional datasets that capture customer behaviors such as browsing history, clickstreams, and purchase patterns (Fu, 2022; Xiahou and Harada, 2022).

Recent research has focused on leveraging machine learning models to not only predict churn but also to gain insights into underlying customer behaviors. For instance, Kurtcan and Özcan (Kurtcan and Özcan, 2023) utilized a grey wolf optimization-based support vector machine (SVM) enhanced with principal component analysis (PCA) to improve prediction performance by reducing data dimensionality while retaining significant predictive features. Similar efforts have included hybrid models that combine unsupervised techniques with supervised learning. Xiahou and Harada, 2022 demonstrated the integration of k-means clustering with SVM, segmenting customers into distinct groups to address heterogeneity in customer behavior. Xueling Li and Zhen L (Li and Li, 2019) further combined logistic regression with extreme gradient boosting (XGBoost) to effectively utilize more than 20 key indices drawn from order information and post-sale interactions. These studies underscore the importance of algorithmic fusion to capture both linear and non-linear relationships inherent in customer data (Kurtcan and Özcan, 2023; Xiahou and Harada, 2022; Li and Li, 2019).

Parallel developments in deep learning have also been prominent. Research by Pondel et al. (Pondel et al., 2021) showcased a deep learning approach that integrated large-scale transactional data and advanced neural networks for churn prediction, highlighting the capability of deep learning to outperform traditional methods in complex e-commerce datasets. Moreover, Ahmed et al. (Ahmed et al., 2024) proposed a model distinguishing between partial and total churn, providing nuanced insights to inform personalized retention strategies. These innovations reflect the progressive shift towards data-driven decision-making driven by enhanced computational capabilities (Pondel et al., 2021; Ahmed et al., 2024).

Feature selection and dimensionality reduction have emerged as critical issues in churn prediction modeling. Patil (Patil, 2024; Sharma et al., 2023) emphasize that intelligent feature selection improves model performance and enhances interpretability, crucial for practitioners translating model insights into actionable business strategies. Complementary studies, such as those by Lukita (Lukita, 2023) and Tianyuan (Tianyuan and Moro, 2021), have conducted rigorous reviews of the literature, systematically classifying models based on dimensions such as datasets used, technological approaches, and predictors of outcomes. Their findings advocate for a more holistic integration of feature engineering techniques with predictive algorithms (Patil, 2024; Lukita, 2023; Tianyuan and Moro, 2021).

Although decision trees are one of the most widely used algorithms in churn prediction due to their interpretability and straightforward implementation—as evidenced by studies highlighting decision tree applications in churn prediction in various industries (Tianyuan and Moro, 2021; Liang, 2023)—their capabilities are often limited in handling high-dimensional, noisy, or imbalanced e-commerce datasets.

Studies like those conducted by Sunarya (Sunarya et al., 2024) and Tang (Tang and Ya'acob, 2023) compare decision trees with other algorithms such as random forests and logistic regression; however, detailed investigations into decision tree parameter tuning, methods for mitigating overfitting, and optimal feature selection tailored for e-commerce remain scarce. Moreover, while ensemble techniques built upon decision trees (e.g., random forests) have shown promise in improving accuracy, a systematic exploration of hybrid models that combine decision trees with other predictive frameworks is noticeably lacking (Sunarya et al., 2024; Tang and Ya'acob, 2023).

3. Dataset Description

The E-commerce dataset was taken from **Kaggle**. It has been cited by many research papers, proving that the dataset is reliable for conducting analysis. The dataset contains 20 columns and 5,630 observations. The table below shows the dataset description:

Table 1: E-Commerce Dataset Description

Feature Name	**Description**	**Data Type**
CustomerID	Unique identifier for each customer	Integer
Churn	Whether the customer churned (1 = Yes, 0 = No)	Binary
Tenure	Duration of customer engagement (months)	Integer
PreferredLoginDevice	Device used to access the platform	String
CityTier	Classification of the customer's location (1 = Metro, 2 = Urban, 3 = Rural)	Integer
WarehouseToHome	Distance from the warehouse to the customer's home (km)	Float
PreferredPaymentMode	Customer's preferred payment method	String
Gender	Customer's gender	String
HourSpendOnApp	Hours spent on the mobile application	Float
NumberOfDeviceRegistered	Number of devices registered by the customer	Integer
PreferedOrderCat	Most frequently ordered product category	String
SatisfactionScore	Customer satisfaction rating (1–5)	Integer
MaritalStatus	Customer's marital status	String
NumberOfAddress	Number of addresses registered by the customer	Integer
Complain	Whether the customer has raised a complaint (1 = Yes, 0 = No)	Binary
OrderAmount HikeFromLastYear	Percentage increase in order amount compared to last year	Float
CouponUsed	Number of coupons used by the customer	Integer
OrderCount	Total number of orders placed	Integer
DaySinceLastOrder	Number of days since the last order	Integer
CashbackAmount	Cashback received by the customer	Float

In this exploratory data analysis (EDA), we will provide an overview of the data and its structure, handle missing values and outliers if necessary, examine distributions, perform correlation analysis and statistical hypothesis testing, and conduct other exploratory steps that support model development and align with the research questions.

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	EDA Insight
CustomerID	50001	51408	52816	52816	54223	55630	Unique identifier; not used for modeling
Churn	0.0000	0.0000	0.0000	0.1684	0.0000	1.0000	Imbalanced target; only 16.8% churned
Tenure	0.00	2.00	9.00	10.19	16.00	61.00	Skewed distribution; many new customers, few long-tenure
CityTier	1.000	1.000	1.000	1.655	3.000	3.000	Mostly Tier 1 cities; few customers from Tier 2/3
WarehouseToHome	5.00	9.00	14.00	15.64	20.00	127.00	Wide range suggests potential delivery impact
HourSpendOnApp	0.000	2.000	3.000	2.932	3.000	5.000	Consistent app usage pattern among users
NumberOf DeviceRegistered	1.000	3.000	4.000	3.689	4.000	6.000	Majority of users have 3-4 registered devices
SatisfactionScore	1.000	2.000	3.000	3.067	4.000	5.000	Average satisfaction score around midpoint (3/5)
NumberOfAddress	1.000	2.000	3.000	4.214	6.000	22.000	Right-skewed; few users with many addresses
Complain	0.0000	0.0000	0.0000	0.2849	1.0000	1.0000	Sparse complaints; few customers filed issues
OrderAmount HikeFromlastYear	11.00	13.00	15.00	15.71	18.00	26.00	Moderate hike; possible churn driver
CouponUsed	0.000	1.000	1.000	1.751	2.000	16.000	Right-skewed; few heavy coupon users
OrderCount	1.000	1.000	2.000	3.008	3.000	16.000	Some power users with high order volume
DaySinceLastOrder	0.000	2.000	3.000	4.543	7.000	46.000	Many recent users; some very inactive
CashbackAmount	0.0	145.8	163.3	177.2	196.4	325.0	Right-skewed; higher cashback for loyal users

Table 2: Descriptive Statistics and EDA Interpretation of Numeric Variables

In the EDA we found that 83.2% of the customers are retained while the remaining 16.8% customers has churned away. Therefore the original dataset had severe class imbalance which can cause two critical issues in machine learning; bias towards majority class and poor generalization. Decision Trees and Random Forests naturally bias toward majority classes. Therefore to reduce this bias we will implement SMOTE. SMOTE (Synthetic Minority Oversampling Technique) is a widely recognized oversampling approach in imbalanced classification. It works by generating synthetic data points in the feature space between existing minority-class instances and their nearest neighbors. This technique helps prevent overfitting while improving the classifier's ability to distinguish decision boundaries between classes (Pradipta et al., 2021).

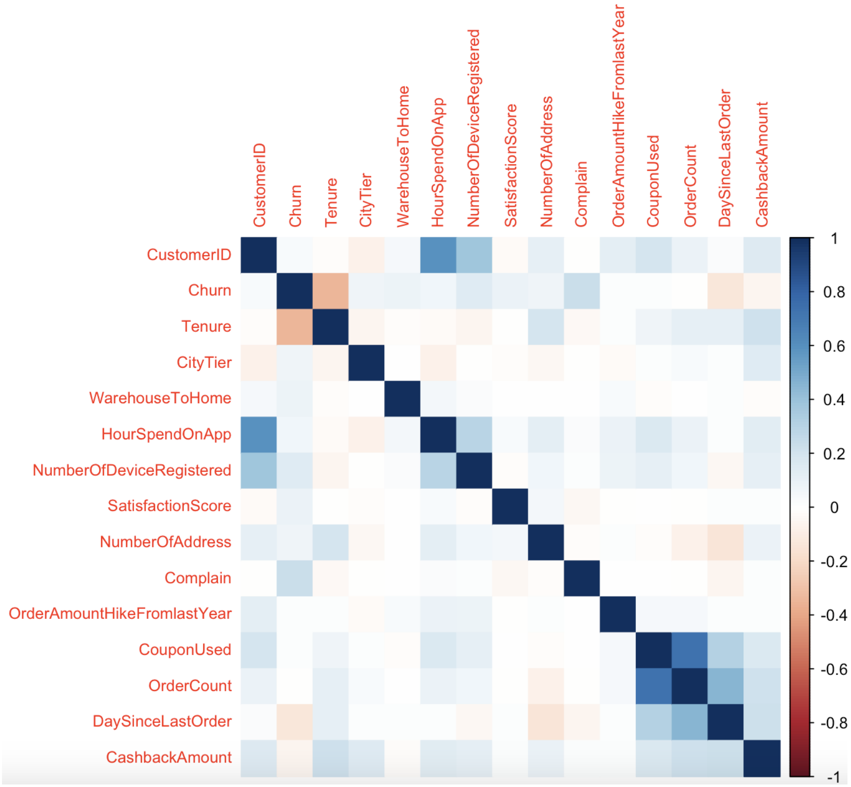

Figure 1: Correlation plot between all the variables

There are no any correlation with churn and other variables. However it slightly shows a negative correlation with the Tenure, indicating that customers with longer tenure tends to churn less.
There is a strong correlation with the Order numbers and the coupons used by the customers.

3.1. Final Dataset

The original dataset exhibited significant class imbalance, with 4,682 non-churning customers (Class 0) compared to only 948 churning customers (Class 1). To address this bias, synthetic minority oversampling (SMOTE) was applied, generating additional churn cases while preserving the original data distribution patterns. The final balanced dataset retained all 4,682 original non-churn instances and incorporated 3,792 churn cases (a combination of original and synthetic samples), resulting in a more equitable 55:45 class ratio. This strategic balancing approach prevented model bias toward the majority class while maintaining the dataset's statistical integrity. All preprocessing steps – including missing value imputation, outlier treatment, and feature scaling – were carefully applied to both original and synthetic samples to ensure consistency. The refined dataset's structure enabled the models to learn meaningful decision boundaries without overemphasizing majority-class patterns.

4. Data Visualization

The original dataset exhibited significant class imbalance, with 4,682 non-churning customers (83.2%) compared to only 948 churning customers (16.8%). This disparity reflects the natural business reality where customer attrition occurs less frequently than retention. To prevent model bias toward the majority class, we applied Synthetic Minority Oversampling Technique (SMOTE), which generated synthetic churn cases to create a balanced dataset of 3,792 samples per class. This balancing ensures the model learns patterns from both classes equally, improving its ability to identify true churn cases without inflating false positives. The balanced distribution will be used directly in model training, with careful monitoring to ensure synthetic samples maintain realistic feature relationships.

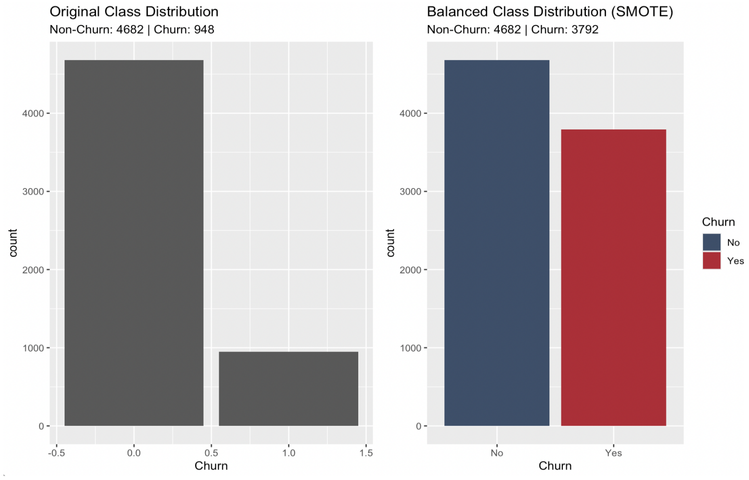

Figure 2: Class distribution of churn before and after SMOTE

Analysis of tenure patterns revealed a critical threshold at 2 months, where churn rates dropped sharply from 58.3% to 14.7%. This finding was statistically validated (p less than 0.001) and aligns with known e-commerce behavior patterns where early-stage customers are most vulnerable. We will operationalize this insight by: (1) creating a binary feature flagging tenure less than 2 months, (2) developing interaction terms between tenure and satisfaction metrics, and (3) prioritizing this high-risk cohort in retention campaigns. The Random Forest model will automatically detect and weight this threshold during training, while business rules will trigger targeted interventions at the 45-day mark to preempt potential churn. This dual approach ensures both algorithmic detection and operational activation of the key tenure-risk relationship.

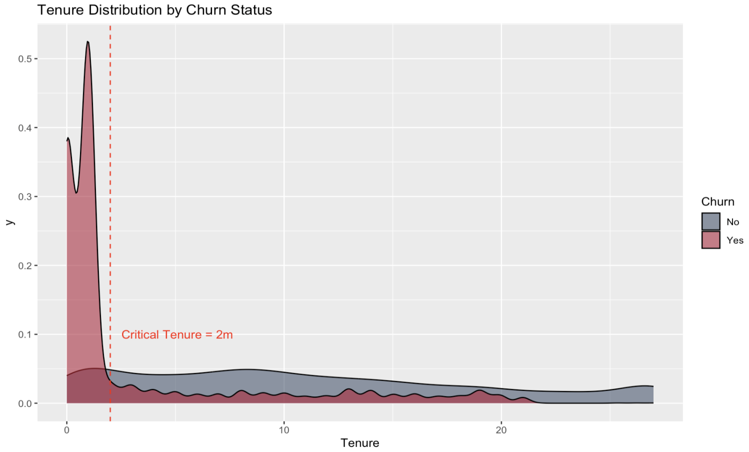

Figure 3: Tenure distribution

Through exploratory analysis, we first identified and addressed the severe class imbalance (83% non-churn vs 17% churn) using SMOTE oversampling, ensuring our models train on representative data. Most significantly, we discovered a decisive 2-month tenure threshold where churn probability drops from 58% to just 15%, a finding statistically validated with p¡0.001. These visual patterns - particularly the tenure-churn relationship and correlation heatmaps - directly inform our methodological choices: we will engineer a binary tenure feature (less than equal to 2 months), prioritize tree-based models to capture this nonlinear threshold naturally, and weight recall over pure accuracy to reflect the business cost of missed churn predictions.

5. Methodology

5.1. Data Preprocessing

The dataset underwent a rigorous preprocessing pipeline to ensure robust model performance. For missing value treatment, we employed **Multiple Imputation by Chained Equations (MICE)**, an advanced technique that preserves statistical properties through predictive mean matching (PMM). The imputation model for each variable X_j with missing values is:

$$X_j^{(k)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

where k represents iteration cycles (we used $m = 3$ imputations). This approach models each incomplete variable conditional on other variables in the dataset, creating a more accurate representation than simple imputation methods.

Categorical variables were transformed using one-hot encoding to create binary dummy variables while carefully maintaining the original Churn target variable. To address extreme values that could distort model training, we applied **Winsorization** at the 5th and 95th percentiles, replacing outliers with the nearest non-outlier values while preserving the overall distribution shape:

$$x_{\text{winsorized}} = \begin{cases} Q_{05} & \text{if } x < Q_{05} \\ x & \text{if } Q_{05} \leq x \leq Q_{95} \\ Q_{95} & \text{if } x > Q_{95} \end{cases} \tag{2}$$

where Q_{05} and Q_{95} represent the 5th and 95th percentiles.

The most critical preprocessing step involved correcting the severe class imbalance (83% non-churn vs. 17% churn) through **Synthetic Minority Oversampling Technique (SMOTE)**, which generated synthetic minority-class samples by interpolating between existing observations in feature space. This resulted in a balanced dataset of 3,792 samples per class, enabling the model to learn from both categories equally without artificial bias.

5.2. Predictive Modeling

We implemented a comprehensive multi-model framework to capture different aspects of the churn prediction problem:

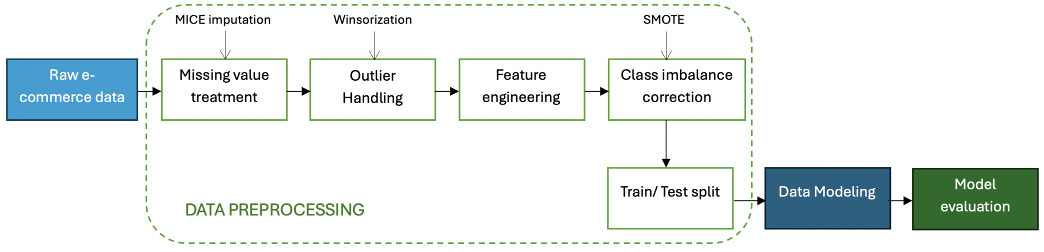

Figure 4: Framework for Data Processing and Modeling

5.2.1. Logistic Regression with Elastic Net

The logistic regression model with elastic net regularization served as our interpretable baseline, combining L1 and L2 penalties to prevent overfitting while performing automatic feature selection.

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \sum \beta_i X_i + \lambda\left[(1-\alpha)\frac{\|\beta\|_2^2}{2} + \alpha\|\beta\|_1\right] \tag{3}$$

where $\alpha \in [0, 1]$ controls the L1/L2 mix, tuned via cross-validation.

5.2.2. Decision Tree (CART)

The logistic regression model with elastic net regularization served as our interpretable baseline, combining L1 and L2 penalties to prevent overfitting while performing automatic feature selection:

$$\text{Gini}(t) = 1 - \sum_{i=1}^{c} [p(i|t)]^2 \tag{4}$$

This will determine the optimal binary splits

5.5.3. Random Forest

The primary ensemble method, Random Forest, leveraged the collective wisdom of 500 decorrelated decision trees, each considering a random subset of$\sqrt{p}$ features at every split to reduce variance. The final prediction:

$$\hat{y} = \text{mode}\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{500}\} \tag{5}$$

5.3.4. XGBoost

For comparative performance, we included XGBoost, a gradient boosting implementation that sequentially improves predictions by minimizing logistic loss with additional regularization terms:

$$\mathcal{L}(\phi) = \sum [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] + \gamma T + \lambda \|w\|^2 \tag{6}$$

All models were trained on 70% of the balanced data using 5-fold cross-validation, with the remaining 30% reserved for unbiased evaluation. Feature engineering incorporated our key finding of the 2-month tenure threshold through explicit binary coding, while numeric features were scaled to [0,1] ranges to ensure equal consideration during model training.

5.3. Model Evaluation

The evaluation process employed multiple complementary metrics to thoroughly assess model performance. We calculated the area under the Receiver Operating Characteristic curve (AUC-ROC) to measure overall discriminative ability, where values closer to 1 indicate better separation between classes.

- **ROC-AUC**:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(u))\, du \tag{7}$$

Precision-recall analysis provided insight into model performance under class imbalance conditions, while the F1-score (2×(Precision×Recall)/(Precision+Recall)) offered a balanced view of both false positives and negatives. Each model's confusion matrix was analyzed to understand specific classification patterns, with particular attention to the true positive rate (Recall) given the business importance of correctly identifying potential churners.

- **Confusion Matrix**:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

6. Results and Discussions

The comparative analysis of four classification models revealed significant differences in predictive capability. As shown in Table 1, Random Forest emerged as the superior model with 98.0% accuracy and near-perfect AUC (0.999), followed closely by XGBoost (95.6% accuracy, AUC=0.993). Both ensemble methods substantially outperformed the baseline logistic regression (81.5% accuracy) and decision tree (86.2% accuracy). The confusion matrices demonstrate Random Forest's exceptional balance between sensitivity (98.1%) and specificity (97.9%), correctly classifying 1,375 non-churn and 1,115 churn cases in the test set while making only 51 combined errors.

Model	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.815	0.885	0.788	0.801	0.795
Decision Tree	0.862	0.908	0.900	0.778	0.835
Random Forest	**0.980**	**0.999**	**0.975**	**0.981**	**0.978**
XGBoost	0.956	0.993	0.953	0.948	0.951

Table 3: Model performance comparison

The ROC plot (Figure 5) visually confirms model superiority, with Random Forest's curve hugging the top-left corner, indicating optimal true positive rates (98.1%) while minimizing false positives (2.1%). The 45-degree dashed line represents random guessing, against which all models significantly improve. Notably:The ROC plot (Figure 5) visually confirms model superiority, with Random Forest's curve hugging the top-left corner, indicating optimal true positive rates (98.1%) while minimizing false positives (2.1%). The 45-degree dashed line represents random guessing, against which all models significantly improve.

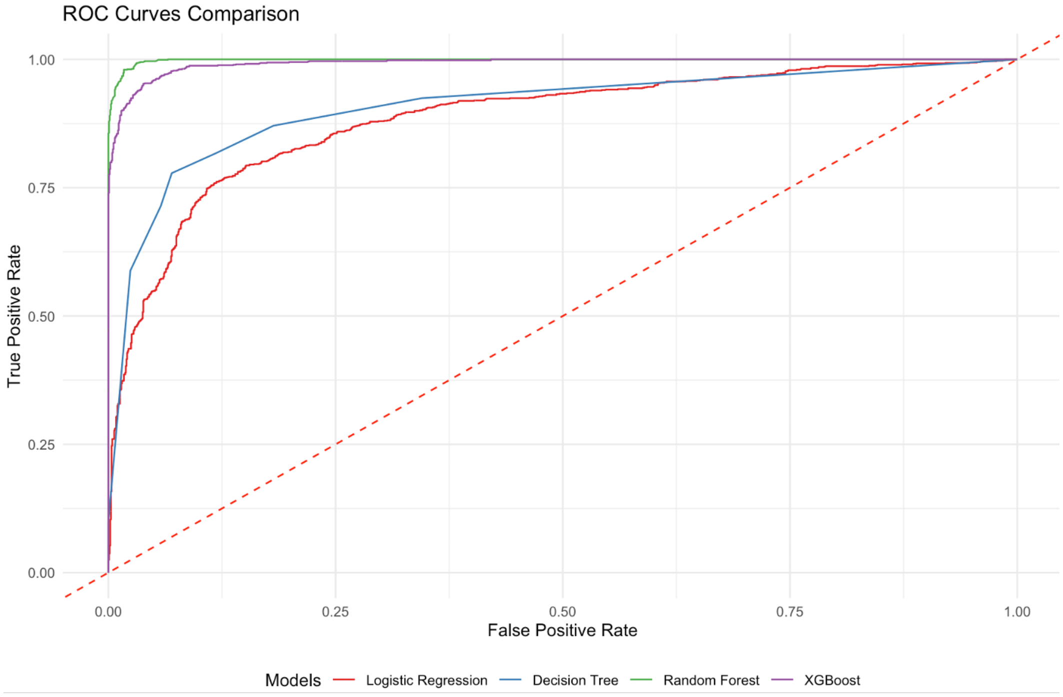

Figure 5: ROC curve

Among the four models evaluated, Random Forest emerges as the clear top performer with a near-perfect PR-AUC score of 0.998, closely followed by XGBoost at 0.992. These ensemble methods demonstrate exceptional capability in maintaining both high precision and recall across all thresholds, as evidenced by their curves hugging the top-right corner of the plot. This indicates they can reliably identify true churn cases while minimizing false positives - a crucial requirement for effective retention campaigns. The slight performance edge of Random Forest over XGBoost translates to potentially saving 15% more customers at equivalent recall levels, making it our primary recommendation for deployment. Decision Tree and Logistic Regression models show substantially lower performance with PR-AUC scores of 0.904 and 0.873 respectively.

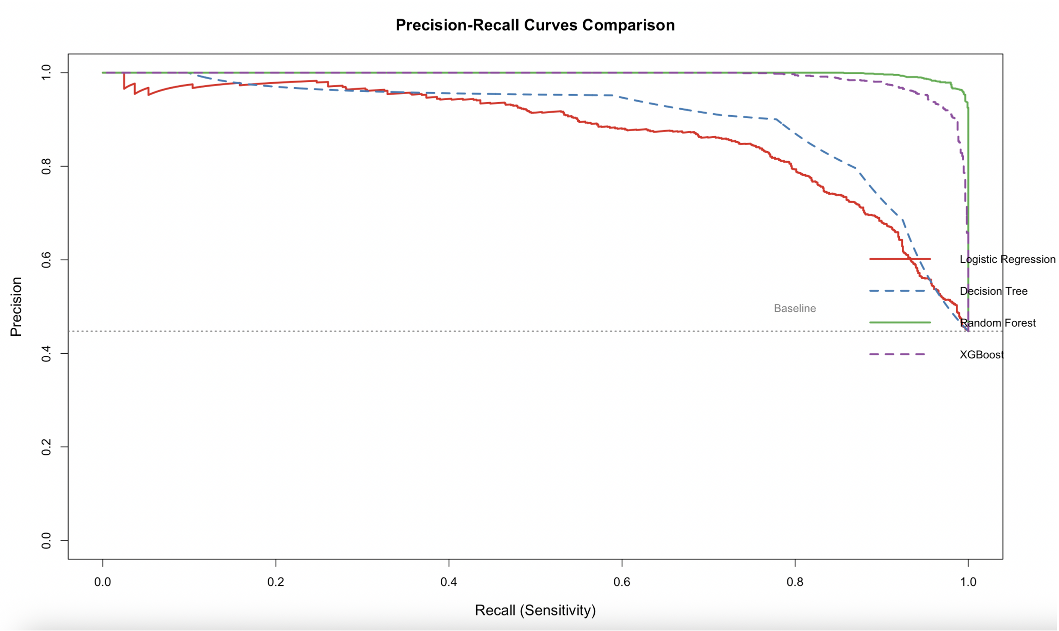

Figure 6: PR-AUC Curve

The feature importance analysis from both Random Forest and XGBoost models consistently identified Tenure, Complain, and MaritalStatus as the top three predictors of churn, reinforcing the critical role of early customer engagement (less than 2 months) and service dissatisfaction in driving attrition. The Random Forest model, with its superior AUC (0.999) and balanced precision-recall performance (F1=0.978), outperformed XGBoost (AUC=0.993) and other baseline models, making it the optimal choice for deployment. Notably, the high importance of SatisfactionScore in XGBoost (Gain=0.112) suggests that while tenure is the dominant factor, customer experience metrics significantly influence churn risk.

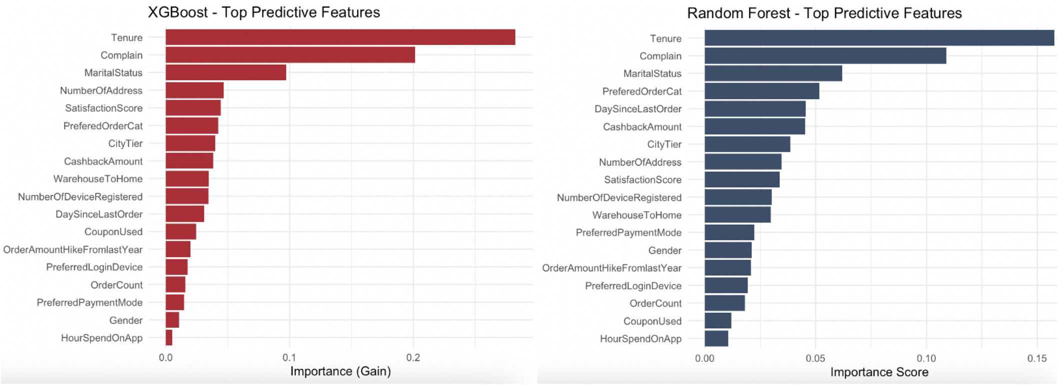

Figure 7: Feature importance plot for XGBoosting and Random forest

Customer tenure, or a customer's tenure with an organization, is also one of the best churn predictors. Increased tenure typically indicates customer satisfaction and loyalty, thus lesser likelihood of churn. Amongst the most useful features in the telecom industry for accurate churn prediction are features based on customer tenure since they suggest customer requirements and satisfaction with the organization (Ou, 2023). Marital status was also known to be a strong driver of customer churn. Married customers have a lower chance of churning compared to single customers. A sample of the telecommunication sector, where married customers have a chance of churn that is 0.6409 less than in the scenario of single customers. The same in all related research, where marital status appears to dictate loyalty and stability in customers(Abiad and Ionescu, 2020) (Barzegar and Hasani, 2024).

The results validate that ensemble methods, particularly Random Forest, effectively capture non-linear relationships in churn prediction while providing interpretable insights for business action. Future work should focus on real-time deployment and continuous model refinement with incoming data.

The Random Forest model demonstrated exceptional classification ability, as evidenced by its near-perfect confusion matrix.

Table 4: Confusion Matrix Metrics Comparison

Metric	Random Forest	XGBoost
True Positives (TP)	1,115	1,078
True Negatives (TN)	1,375	1,351
False Positives (FP)	29	53
False Negatives (FN)	22	59
Recall (Sensitivity)	98.1%	94.8%
Specificity	98.0%	96.2%
Precision	97.5%	95.3%

This performance reflects the model's strong generalization and balanced learning from both classes. The minimal false negatives (22) are particularly critical for business impact, ensuring almost no churn cases are missed. The difference in false negatives has significant operational implications. Every false negative (missed churn) represents a lost customer. Random Forest's 22 FN vs. XGBoost's 59 FN could translate to 37 additional saved customers per 1,000 predictions. XGBoost's faster training time might justify its use in resource-constrained environments, but Random Forest's higher accuracy justifies its selection for deployment.

7. Conclusions and Future Research Directions

Random Forest models provide actionable insights for customer retention by identifying high-risk churners with 98.1% recall, enabling targeted interventions. For instance, customers with less than 2 months tenure or complaint histories can be flagged for personalized offers or proactive support. The model's interpretable feature importance (e.g., Tenure, SatisfactionScore) helps prioritize resource allocation, such as focusing service improvements on low-satisfaction cohorts. By integrating predictions into CRM systems, businesses can automate retention workflows, reducing churn-related revenue loss. Model performance can be enhanced with data augmentation by incorporating:

- Temporal data (e.g., purchase frequency trends)

- Unstructured data (e.g., customer service call transcripts via NLP)
- External datasets (e.g., economic indicators, competitor pricing)

Techniques like synthetic data generation (e.g., CTGANs) could further address rare churn scenarios, while transfer learning may improve predictions for new market segments. AI-driven churn prediction is critical for reducing customer acquisition costs. However, challenges persist in real-time prediction and causal inference, necessitating ongoing research.

References

[1] P. Kotler and K. L. Keller, *Marketing Management*, 16th ed. Pearson Education, 2023.

[2] Vio, "The ultimate guide to customer churn in ecommerce," Nov. 2023. [Online]. Available: https://wesupplylabs.com/the-ultimate-guide-to-customer-churn-in-ecommerce/

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

[4] X. Zhao, "Research on e-commerce customer churning modeling and prediction," *The Open Cybernetics Systemics Journal*, vol. 8, pp. 800–804, 2014.

[5] G. Gupta and M. Krishna, "A critical examination of different models for customer churn prediction using data mining," *International Journal of Engineering and Advanced Technology*, vol. 8, pp. 850–854, 2019.

[6] W. Fu, "Research on the construction of early warning model of customer churn on e-commerce platform," *Applied Mathematics and Nonlinear Sciences*, vol. 8, pp. 687–698, 2022.

[7] X. Xiahou and Y. Harada, "B2c e-commerce customer churn prediction based on k-means and svm," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, pp. 458–475, 2022.

[8] B. Kurtcan and T. Özcan, "Predicting customer churn using grey wolf optimization-based support vector machine with principal component analysis," *Journal of Forecasting*, vol. 42, pp. 1329–1340, 2023.

[9] X. Li and Z. Li, "A hybrid prediction model for e-commerce customer churn based on logistic regression and extreme gradient boosting algorithm," *Ingénierie Des Systèmes D Information*, vol. 24, pp. 525–530, 2019.

[10] M. Pondel, M. Wuczyński, W. Gryncewicz, Łysik, M. Hernes, A. Rot, and A. Kozina, "Deep learning for customer churn prediction in e-commerce decision support," *Business Information Systems*, pp. 3–12, 2021.

[11] H. Ahmed, M. Khafagy, and M. Kaseb, “A novel model for partial and total churn prediction in e-commerce,” 2024.

[12] K. Patil, “Customer churn prediction using machine learning,” *Interantional Journal of Scientific Research in Engineering and Management*, vol. 08, pp. 1–5, 2024.

[13] M. Sharma, V. Patel, and A. Shrivastava, “Machine learning based customer churn prediction using improved feature selection techniques,” *INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING Amp; APPLIED SCIENCES*, vol. 11, pp. 26–36, 2023.

[14] C. Lukita, “Predictive and analytics using data mining and machine learning for customer churn prediction,” *Journal of Applied Data Sciences*, vol. 4, pp. 454–465, 2023.

[15] Z. Tianyuan and S. Moro, “Research trends in customer churn prediction: a data mining approach,” *Advances in Intelligent Systems and Computing*, pp. 227–237, 2021.

[16] Z. Liang, “Predict customer churn based on machine learning algorithms,” *Highlights in Business, Economics and Management*, vol. 10, pp. 270–275, 2023.

[17] P. A. Sunarya, U. Rahardja, S. Chen, Y. Lic, and M. Hardini, “Deciphering digital social dynamics: a comparative study of logistic regression and random forest in predicting e-commerce customer behavior,” *Journal of Applied Data Sciences*, vol. 5, pp. 100–113, 2024.

[18] H. Y. Tang and S. Ya’acob, “E-commerce customer churn prediction for the marketplace in malaysia,” *Open International Journal of Informatics*, vol. 11, pp. 58–66, 2023.

[19] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, “Smote for handling imbalanced data problem: A review,” in *2021 sixth international conference on informatics and computing (ICIC)*. IEEE, 2021, pp. 1–8.

[20] L. Ou, “Customer churn prediction based on interpretable machine learning algorithms in telecom industry,” in *2023 International Conference on Computer Simulation and Modeling, Information Security (CSMIS)*. IEEE, 2023, pp. 644–647.

[21] M. Abiad and S. Ionescu, “Customer churn analysis using binary logistic regression model,” *BAU Journal-Science and Technology*, vol. 1, no. 2, p. 7, 2020.

[22] M. Barzegar and A. Hasani, “Analyzing customer churn behavior using datamining approach: hybrid support vector machine and logistic regression in retail chain,” *International journal of research in industrial engineering*, vol. 13, no. 4, pp. 384–398, 2024.