Customer Churn Prediction using Machine Learning: A Case Study of E-commerce Data

Jingyuan Li
Beijing Jiaotong University
No.3 Shangyuan Village,
Haidian District

ABSTRACT

In the highly competitive e-commerce industry, customer churn represents a major challenge to profitability and sustainability. This study aims to develop a robust predictive model for customer churn using a publicly available e-commerce dataset. The research leverages various machine learning algorithms, including Logistic Regression, Random Forest, XGBoost, and LightGBM, to compare performance. We address class imbalance with SMOTE and utilize SHAP and LIME for model interpretability. Our results demonstrate the effectiveness of the Random Forest model, achieving a ROC AUC of 0.9850. This study provides valuable insights into the factors driving customer churn, offering actionable recommendations for businesses to reduce churn rates and enhance customer retention strategies.

General Terms

Machine Learning, E-commerce, Churn Prediction.

Keywords

Customer Churn, Random Forest, XGBoost, E-commerce Data.

1. INTRODUCTION

Customer churn is a critical issue across various industries, and predicting it has been a subject of extensive research. Recent advancements in machine learning have enabled the development of more accurate models for churn prediction, particularly in industries like e-commerce where competition is fierce. Research has demonstrated that by leveraging data-driven models, businesses can implement effective retention strategies that reduce churn rates and increase profitability (Peddarapu et al., 2022). These strategies become even more effective when coupled with advanced machine learning techniques that capture complex customer behavior and provide actionable insights.

2. LITERATURE REVIEW

In the context of machine learning applications for customer churn prediction, various models have been studied. For instance, deep learning approaches have shown promise in improving predictive performance in complex e-commerce environments, particularly where customer behavior is highly variable (Pondel et al., 2021). Studies emphasize the significance of addressing class imbalance, which can skew results and impact the model's ability to accurately identify churners. Techniques like SMOTE have been widely adopted to mitigate this issue, leading to better precision and recall in a variety of applications (Zimal et al., 2023).

In addition, research in the B2B e-commerce sector has highlighted the use of support vector machines and parameter optimization to enhance churn prediction. These models are particularly effective in environments with noisy and imbalanced data, demonstrating strong generalization capabilities (Gordini & Veglio, 2017)..

3. DATA AND METHODOLOGY

3.1 Data Description

The dataset used in this study was sourced from Kaggle and contains 5,597 records after data cleaning. It includes various features such as customer tenure, preferred login device, city tier, and satisfaction scores. The target variable is binary, indicating whether a customer has churned (1) or not churned (0). Table 1 provides a summary of the key features used in this study.

Table 1. Feature Description

Feature	Description
CustomerID	Unique identifier for each customer
Tenure	Number of years the customer has been with the company
PreferredLoginDevice	Customer’s preferred login device (mobile, desktop, etc.)
CityTier	Tier of the customer’s city (market regions)
WarehouseToHome	Distance from the warehouse to the customer's home
SatisfactionScore	Customer’s satisfaction score (1-5)
Churn	Whether the customer has churned (1) or not (0)

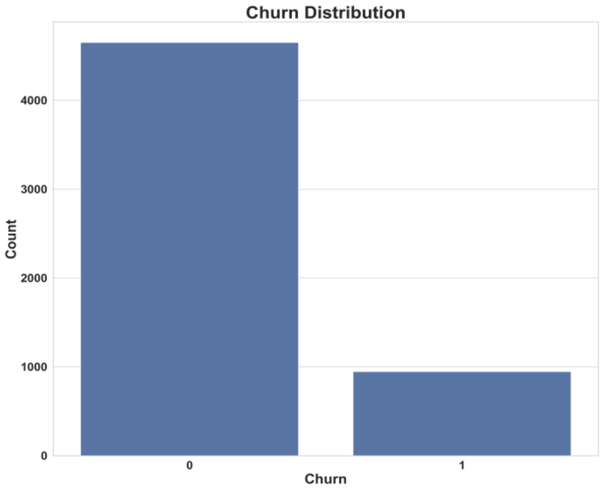

Figure 1: Churn Distribution Bar Chart

The chart illustrates that most customers do not churn, necessitating techniques like SMOTE to balance the dataset.

3.2 Data Preprocessing

Data preprocessing is a critical step to ensure the quality and consistency of the dataset. Missing values in numerical variables were imputed using the median, while categorical variables were imputed with the mode. Outliers in features like OrderAmountHikeFromlastYear were removed using the fourth quartile method, which eliminated 33 extreme records.

3.3 Feature Engineering

To improve model performance, several feature engineering techniques were applied, including one-hot encoding for categorical features and the creation of new interaction terms. All numerical features were standardized using Standard Scaling to mitigate unit discrepancies.

3.4 Data Visualization

To better understand the distribution of features and their relationship to churn, several visualizations were generated. Figure 2 shows the distribution of satisfaction scores for churned and non-churned customers.

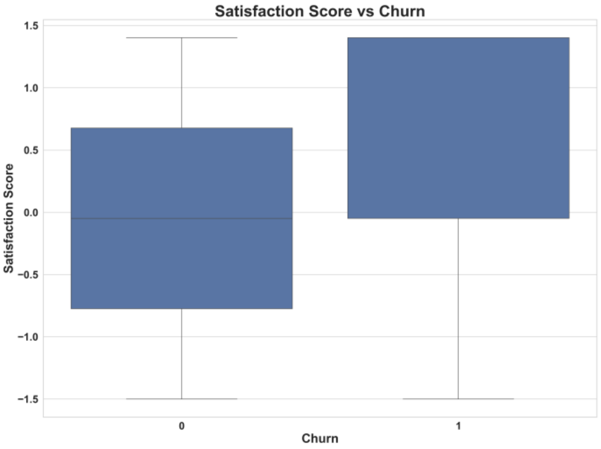

Figure 2: Satisfaction Score vs Churn Boxplot

This boxplot illustrates that churned customers tend to have lower satisfaction scores, but with greater variance compared to non-churned customers.

Similarly, Figure 3 highlights the relationship between customer complaints and churn.

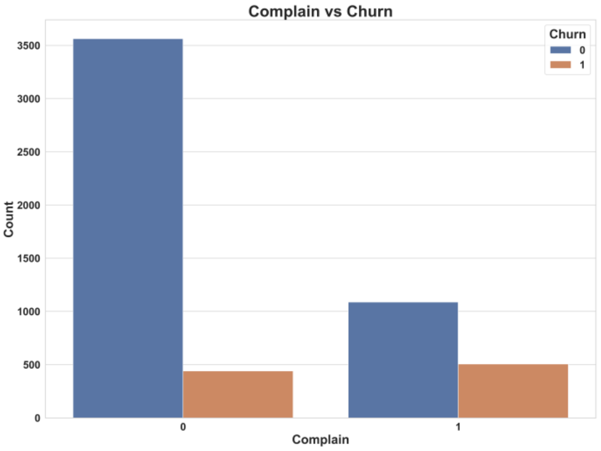

Figure 3: Complain vs Churn Countplot

This countplot demonstrates that customers who file complaints are more likely to churn.

3.5 Class Imbalance Handling

The dataset exhibited a significant class imbalance, with approximately 83% of customers not churning and 17% churning. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training dataset. This technique artificially oversamples the minority class (churned customers) to achieve a balanced class distribution, which mitigates the risk of bias in predictive models (Zimal et al., 2023).

4. MODEL DEVELOPMENT

4.1 Model Selection:

For the binary classification problem of customer churn, various machine learning algorithms were explored. Logistic Regression served as the baseline model due to its simplicity and interpretability. Additionally, more complex models, including Decision Trees, Random Forests, XGBoost, and LightGBM, were selected to capture non-linear patterns in the data.

The evaluation metrics used to compare model performance included Accuracy, Precision, Recall, F1-Score, and ROC AUC. Given the imbalanced nature of the dataset, particular emphasis was placed on Precision, Recall, and ROC AUC to ensure robust performance in identifying churned customers.

4.2 Model Performance

Table 2 compares the performance of the models.

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.78	0.95	0.42	0.55	0.8403
Decision Tree	0.92	0.96	0.74	0.78	0.8826
Random Forest	0.95	0.98	0.83	0.86	0.9851
XGBoost	0.96	0.98	0.88	0.89	0.9872
LightGBM	0.93	0.96	0.78	0.80	0.9708

Table 2: Model Performance Comparison

The Random Forest and XGBoost models demonstrated the best performance, achieving ROC AUC scores of 0.9851 and 0.9872, respectively. Due to its overall balanced performance, Random Forest was selected for further hyperparameter tuning using Grid Search.

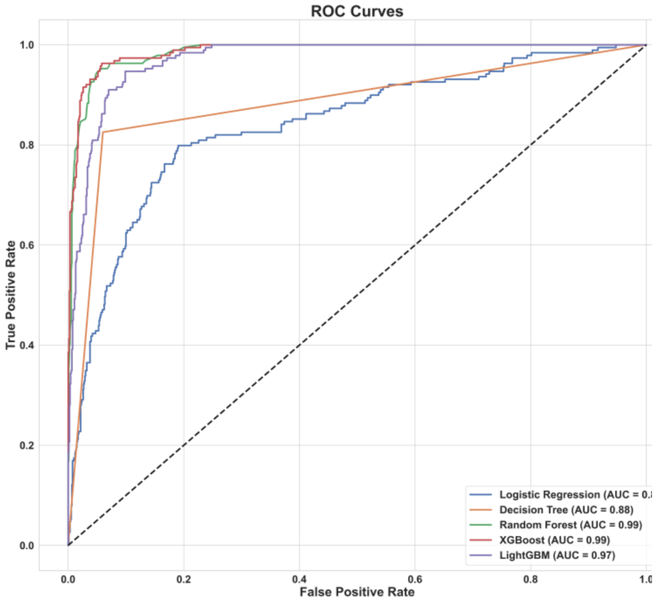

Figure 4: ROC Curves for Various Models

The ROC curves illustrate that the Random Forest and XGBoost models outperform others, with ROC AUC scores nearing 1.

4.3 Confusion Matrix

To further assess the performance of the Random Forest model, a confusion matrix was constructed (Figure 5). This matrix shows the true positives, false positives, true negatives, and false negatives.

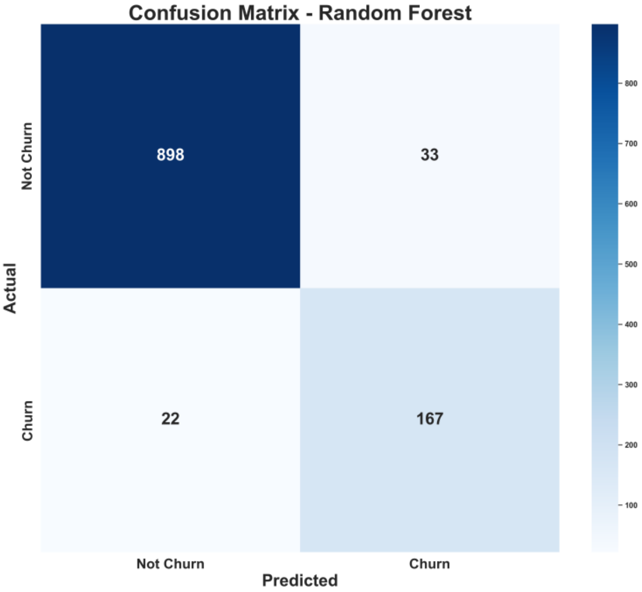

Figure 5: Confusion Matrix - Random Forest

The confusion matrix highlights the model's ability to predict both churned and non-churned customers accurately, with minimal misclassifications.

4.4 Model Interpretation

SHAP and LIME Analysis:

To enhance model interpretability, SHAP and LIME techniques were employed. SHAP values were used to identify the most important features contributing to churn predictions, while LIME provided localized explanations for individual customer predictions.

The SHAP Summary Plot (Figure 6) revealed that OrderCount_per_Tenure, poly_Tenure^2, and SatisfactionScore were among the top features influencing churn. This aligns with business intuition, as customers with lower satisfaction scores and shorter tenure are more likely to churn (Guo-en Xia & Qingzhe He, 2018).

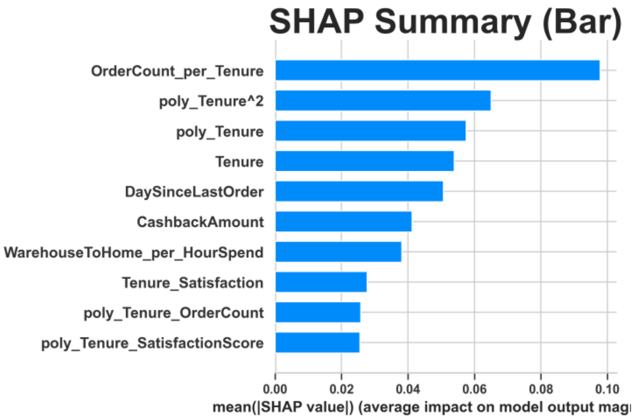

Figure 6: SHAP Summary Bar Plot

The SHAP plot shows that features such as OrderCount_per_Tenure and SatisfactionScore have the greatest impact on churn predictions.

Additionally, Figure 7 displays a SHAP Summary Scatter Plot, which visualizes the relationship between feature values and their contribution to the churn prediction.

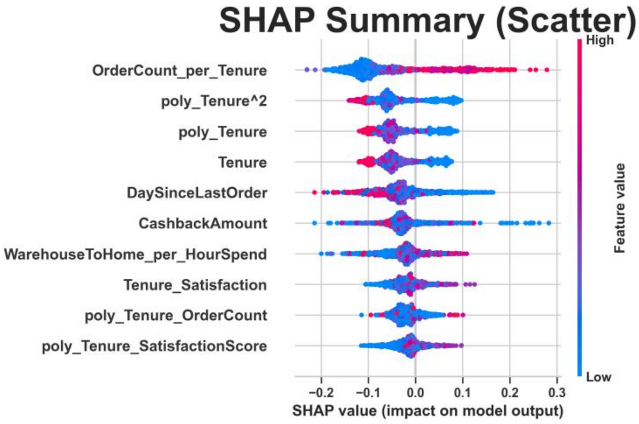

Figure 7: SHAP Summary Scatter Plot

High values of OrderCount_per_Tenure (red points) are associated with lower churn probabilities, while high values of DaySinceLastOrder are linked to a higher probability of churn.

Figure 8 presents the LIME explanation for the Random Forest model's predictions. This detailed breakdown of feature contributions highlights why specific customers were classified as likely to churn or not churn.

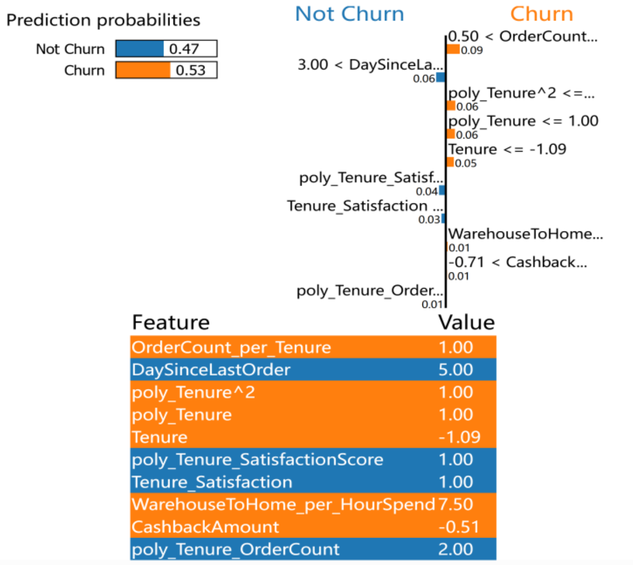

Figure 8: LIME Explanation of Churn Predictions

The LIME plot illustrates the localized feature contributions for a specific prediction, offering greater transparency into why the model classified a customer as likely to churn.

5. DISCUSSION

The findings from this study highlight the importance of using machine learning models to predict customer churn effectively. By understanding the key factors driving churn, businesses can implement targeted strategies to retain high-risk customers. Features such as Tenure, SatisfactionScore, and OrderCount were identified as strong predictors of churn, aligning with findings from previous studies (Pondel et al., 2021). Businesses can leverage these insights to design targeted retention strategies. For instance, low satisfaction scores may indicate the need for immediate customer engagement interventions, while short-tenured customers could be offered loyalty incentives to enhance retention. This study's insights suggest that proactive outreach programs focusing on customer satisfaction and engagement could significantly reduce churn rates and improve long-term profitability. A key observation from our results is the superior performance of XGBoost, similar to the findings of Gordini & Veglio (2017), who used SVMs and Random Forests for churn prediction. Machine learning models, particularly XGBoost and Random Forest, prove to be highly effective for churn prediction, as highlighted in previous studies (Zhang et al., 2023).

In comparison to Gordini & Veglio, our findings further highlight that models with enhanced interpretability, such as Random Forest coupled with SHAP, provide businesses with not only predictive capabilities but also actionable insights into customer behavior. This allows companies to fine-tune their marketing and customer service strategies based on the identified churn-driving factors. Future research could explore the integration of customer feedback data to enhance churn predictions even further.

6. CONCLUSION

In conclusion, this study makes a valuable contribution to the growing body of research on customer churn prediction by demonstrating the effectiveness of machine learning models, particularly Random Forests and XGBoost, in achieving high predictive accuracy. The application of SMOTE to address class imbalance, along with the use of SHAP and LIME for model interpretability, enhances the practical applicability of these models in the e-commerce industry. However, certain limitations exist, such as the use of a single dataset from Kaggle, which may not fully reflect the broader complexities of customer behavior across different e-commerce platforms.

Looking forward, there are several key areas where future research could further strengthen the predictive power and applicability of churn prediction models. For example, integrating additional data sources, such as detailed customer interaction history, real-time feedback, and personalized purchase patterns, could offer a more comprehensive understanding of the factors driving churn. This would not only improve the accuracy of churn predictions but also provide more actionable insights for businesses aiming to reduce churn.

Another promising direction for future work involves the use of more advanced deep learning techniques, such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks. These methods are particularly well-suited for analyzing sequential data, allowing for a better understanding of how customer behavior evolves over time, which is often a crucial factor in churn prediction.These directions offer promising avenues for further improvement in the field of churn prediction and customer retention.

7. REFERENCES

[1] Xia, G., & He, Q. (2018). The Research of Online Shopping Customer Churn Prediction Based on Integrated Learning. , 259-267. https://doi.org/10.2991/MECAE-18.2018.133.

[2] Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. Industrial Marketing Management, 62, 100-107. https://doi.org/10.1016/J.INDMARMAN.2016.08.003.

[3] Zimal, S., Shah, C., Borhude, S., Birajdar, A., & Patil, P. (2023). Customer Churn Prediction Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2023.49142.

[4] Peddarapu, R., Ameena, S., Yashaswini, S., Shreshta, N., & PurnaSahithi, M. (2022). Customer Churn Prediction using Machine Learning. 2022 6th International Conference on Electronics, Communication and Aerospace Technology, 1035-1040. https://doi.org/10.1109/ICECA55336.2022.10009093.

[5] Pondel, M., Wuczynski, M., Gryncewicz, W., Lysik, L., Hernes, M., Rot, A., & Kozina, A. (2021). Deep Learning for Customer Churn Prediction in E-Commerce Decision Support. , 3-12. https://doi.org/10.52825/bis.v1i.42.

[6] Zhang, X.; Guo, F.; Chen, T.; Pan, L.; Beliakov, G.; Wu, J. A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research. J. Theor. Appl. Electron. Commer. Res. 2023, 18, 2188-2216. https://doi.org/10.3390/jtaer18040110

[7] Kaggle. (2021). E-commerce customer churn analysis and prediction dataset. Retrieved on September 20, 2024, from https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction/data