E-Commerce Customer Churn

Gizem Yüzer[1], Erva Yurtbaş[2], Zeynep Sena Tınaz[3]

Abstract— **In the highly competitive e-commerce industry, understanding and predicting customer churn is crucial for maintaining a robust customer base and optimizing marketing strategies. This study presents a comprehensive analysis of customer churn prediction using a dataset of approximately 49,000 customers with 49 initial features. Through meticulous feature selection using the mRMR method, the feature set was reduced to 15 critical predictors. Given the inherent class imbalance in the dataset, a combination of upsampling and downsampling techniques was employed to ensure a balanced representation of churn and retention classes. We evaluated four different machine learning models: Random Forest, Support Vector Machine (SVM), Logistic Regression, and XGBoost 2.0. These models were assessed based on accuracy, recall, precision, and F-score to determine their effectiveness in predicting customer churn. Furthermore, we leveraged Explainable AI, specifically SHAP (SHapley Additive exPlanations), to interpret the model outcomes, providing deeper insights into the driving factors behind customer retention and churn. The results of this study offer valuable insights for e-commerce businesses seeking to enhance customer engagement and reduce churn rates through targeted strategies informed by data-driven predictions. Our dataset is available at:https://www.kaggle.com/datasets/fridrichmrtn/user-churn-dataset**

Keywords: E-commerce, Customer Churn, Binary Classification, XGBoost, Logistic Regression, Support Vector Machine, Random Forest, Explainable AI

I. INTRODUCTION

In the rapidly evolving landscape of e-commerce, customer retention emerges as a pivotal factor for the success and sustainability of online businesses. The phenomenon of customer churn – where customers cease their engagement with a platform – presents a significant challenge, directly affecting the revenue and long-term viability of e-commerce companies. In this highly competitive market, the ability to accurately predict and understand the reasons behind customer churn is not just advantageous but essential for crafting effective retention strategies.

This study focuses on a comprehensive dataset from an e-commerce platform, comprising approximately 49,000 customer records, each characterized by an extensive set of 49 features. These features encompass a wide range of customer interactions, including session recency and frequency, transaction history, browsing behaviors, and engagement across various product categories. The richness of this dataset offers a unique opportunity to delve deep into the patterns and predictors of customer churn.

However, the complexity and size of the dataset present their own challenges. The initial set of 49 features, while comprehensive, risks including redundant or irrelevant information that could obscure meaningful patterns and insights. To address this, we employed the mRMR (Minimum Redundancy Maximum Relevance) feature selection method, a robust technique that evaluates the trade-off between the relevance of each feature to the target variable (customer churn) and the redundancy among features. This process allowed us to refine the dataset to 15 key features, ensuring a more focused and effective analysis.

A common hurdle in churn prediction is the class imbalance problem, where the number of customers who churn is typically much lower than those who do not. Such imbalance can lead to biased predictive models that are overly skewed towards the majority class. To overcome this, our study applied a combination of upsampling of the minority class (churn) and downsampling of the majority class (retention), achieving a balanced representation that enhances model accuracy and generalizability.

Our analytical approach encompasses a diverse suite of machine learning models: Random Forest, Support Vector Machine (SVM), Logistic Regression, and XGBoost 2.0. Each of these models brings unique strengths and perspectives to the problem of churn prediction. Random Forest, with its ensemble of decision trees, offers robustness and handles non-linear relationships well. SVM is effective in high-dimensional spaces, and Logistic Regression provides a probabilistic understanding of customer behaviors. XGBoost 2.0, known for its efficiency and performance, is a cutting-edge gradient boosting framework. The performance of these models was meticulously evaluated using a range of metrics, including accuracy, recall, precision, and F-score, to provide a holistic view of their predictive capabilities.

To add an additional layer of insight, we integrated Explainable AI techniques into our study, specifically utilizing SHAP (SHapley Additive exPlanations). This approach allows us to interpret the predictive models in a human-understandable format, revealing the contribution of each feature to the likelihood of churn. By doing so, we aim to provide actionable insights that go beyond mere predictions, offering e-commerce businesses a lens into the "why" behind customer churn.

Through this comprehensive analysis, our study aims to deliver not only precise predictive models for customer churn but also to unearth the underlying factors that drive these behaviors. By understanding these key drivers, e-commerce platforms can tailor their customer engagement and retention strategies more effectively, ultimately leading to a stronger, more loyal customer base.

II. LITERATURE SURVEY

Li used Random Forest for a customer churn prediction dataset that he found in AliCloud's Tianchi data platform achieving an accuracy of 91.12% and the F1 score of 0.9489 [1].

Raeisi and Sajedi conducted customer churn prediction for an online food ordering service in Tehran, comparing various algorithms including k-NN, Naive Bayes, Gradient Boosting Trees, Decision Tree, and Random Forest. Among these, Gradient Boosting Trees outperformed other algorithms, achieving an accuracy of 86.90% and demonstrating superior predictive capabilities [2].

Wu and Meng addressed the imbalanced dataset using SMOTE and achieved the highest F1 Score of 0.915 with the combination of SMOTE and Adaboost [3].

Orina et al. used different algorithms to predict customer churn in banking industry. They used Random Forest, AdaBoost, XGBoost, Support Vector Machine, Logistic Regression, Neural Net- works, Decision Tree, and K-Nearest Neighbor machine learning models and Random Forest scored better than the other models, obtaining an F1 score of 0.85, ROC of AUC of 0.85, and accuracy of 88%. [4].

An ensemble approach on online shopping churn prediction system is provided by an algorithm developed by Reddy et al. for public use. The model developed in this study aims to deliver the most accurate recommendation with a 90.65% accuracy rate based on the user's daily / regular shopping needs [5]..

III. DATA PREPROCESSING AND VISUALIZATION

In our study, we employed four distinct machine learning algorithms to identify customers at risk of churning, a critical task for maintaining business stability. Initially, our dataset comprised 48 variables and included 49,358 individual data entries. This dataset was inherently imbalanced, featuring a binary target variable where '0' signified customers likely to churn, and '1' represented those inclined towards retention.

Given the skewed nature of our dataset, it was imperative to implement a strategy to balance it, thereby ensuring the effectiveness and accuracy of our machine learning models. To this end, we first employed an upsampling technique for the churn category (0), followed by a downsampling approach for the retention category (1). This methodological adjustment resulted in a more balanced dataset comprising 40,000 entries, a crucial step for enhancing the predictive quality of our analysis.

Additionally, considering the dataset's complexity with its 48 features, we recognized the necessity of feature selection. This process was not merely a matter of reducing dimensionality but also a strategic move to distill the dataset to its most informative attributes. By doing so, we aimed to refine our dataset, focusing on those features that provided the most significant insights and contributed most substantially to the prediction of customer churn. This refined approach was essential in optimizing our machine learning models for more accurate and meaningful outcomes in identifying potential customer churn.

IV. FEATURE SELECTION

Our approach to feature selection involved both computational analysis and experimentation. We explored various feature counts, specifically 15, 20, 25, 30, and 35. The initial technique employed was MRMR (Minimum Redundancy Maximum Relevance), and we experimented with different selection sizes, determining that the optimal result was achieved with 25 features.

MRMR (Minimum Redundancy Maximum Relevance) is a feature selection technique that optimizes the balance between relevance to the target variable and minimization of redundancy among features. It computes relevance scores for each feature, assesses feature redundancy, and selects a subset of informative, non-redundant features. This method is effective for high-dimensional datasets and is often used with machine learning algorithms to improve model interpretability and performance.

Our second technique involved the SelectKBest method, utilizing mutual information as the scoring function for feature selection. This method selects features based on mutual information, a metric quantifying the dependency between variables. Specifically, it identifies the top 25 features with the highest mutual information concerning the target variable ('target_class').

The selected features are then presented for further analysis or utilization in downstream tasks. Feature selection is imperative for enhancing model performance, mitigating overfitting, and augmenting interpretability. It allows a focus on the most informative features, potentially expediting training and simplifying the model.

Upon experimentation, we observed that all four models performed more successfully when using the SelectKBest method. Consequently, we adopted this method for our feature selection approach

In analyzing feature correlations in our e-commerce dataset, we utilized a heatmap as a key tool. The heatmap's color variations, from dark red to light tan, reflect the strength and nature of linear relationships between variables. Features with strong positive correlations, shown in darker red, may indicate redundancy and limited additional predictive value. In contrast, darker blue or black tones signify negative correlations, potentially crucial for differentiating between churn and retention. However, many features show weak correlations with the target variable 'target_class', as seen in lighter shades, suggesting limited standalone predictive power. This analysis informs our feature selection, guiding us to focus on variables with significant correlations to customer churn but minimal interrelationships, thereby improving the robustness and interpretability of our predictive models.

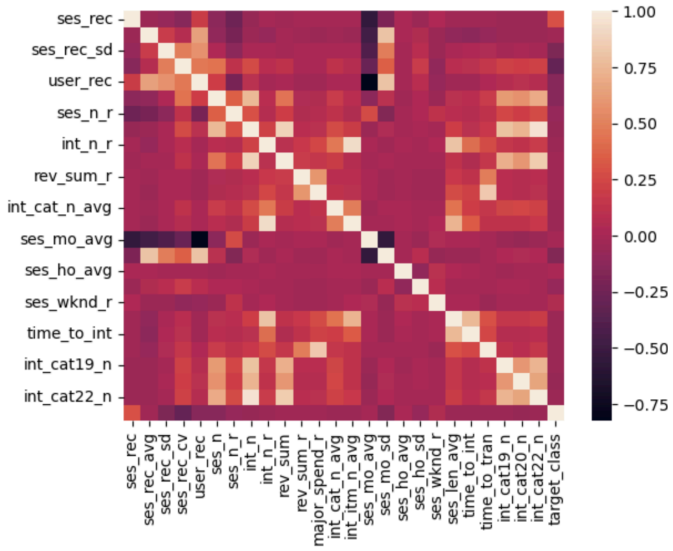

Figure 1. Correlation Matrix for Selected Features

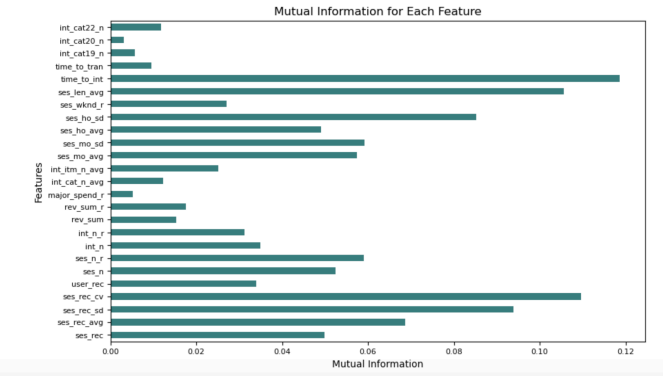

Figure 2. Mutual Information for Selected Features

This visualization allows you to quickly identify which features have a positive or negative influence on predicting the target class based on the Mutual Information scores. Higher scores indicate a stronger association with the target variable.

V. CUSTOMER CHURN PREDICTION MODELS

We have explored various machine learning models in addressing our supervised binary classification problem. The initial model employed was XGBoost; however, even after optimizing its parameters through GridSearch, the model failed to exhibit satisfactory recall and accuracy. Given the nature of our classification problem, where F1 and recall scores carry significant importance alongside accuracy, the performance of the XGBoost model did not meet our expectations.

Subsequently, we experimented with a Support Vector Machine (SVM) model, employing different kernels, with the linear kernel proving to be the most suitable for our specific problem. Unfortunately, the performance scores of the SVM model were inferior to those of the XGBoost model.

Our third endeavor involved the utilization of a Linear Regression model. Despite experimenting with various parameters, we failed to achieve superior solutions compared to the earlier models.

The standout performer in our suite of models emerged as the Random Forest Model, consistently delivering successful results across all testing metrics.

In a final attempt to enhance performance, we implemented a voting algorithm, specifically an Ensemble model, with both "soft" and "hard" voting schemes. However, even with these ensemble techniques, the Random Forest Model continued to outperform other models. This conclusion was reinforced through the analysis of ROC curves and AUC values.

A. XGBoost 2.0

In our research, we utilized the XGBoost algorithm, renowned for its efficacy in classification tasks, to predict customer churn. To optimize the XGBoost model, we employed GridSearchCV from sklearn.model_selection, a robust method for hyperparameter tuning and cross-validation. This approach allowed us to systematically explore a range of hyperparameters, including tree depth ('max_depth'), learning rate ('learning_rate'), and subsample ratio ('subsample'), to find the most effective combination for our model.

Our methodology began with establishing a hyperparameter grid for the XGBoost model. This grid covered various combinations of 'max_depth' (45, 50, 55), 'learning_rate' (0.1, 0.01, 0.001), and 'subsample' (0.5, 0.7, 1), providing a comprehensive range for experimentation. By configuring the GridSearchCV object with our XGBoost model and this hyperparameter grid, and setting the cross-validation strategy to 5 folds, we ensured a thorough search for the optimal parameters.

Upon fitting GridSearchCV to our training data, we identified the best set of hyperparameters and the corresponding score. This step was crucial in enhancing the model's accuracy and generalizability. Following this, we further refined our model by setting specific parameters, such as 'eta', 'max_depth', 'gamma', 'min_child_weight', 'n_estimators', and 'subsample', based on our findings from GridSearchCV.

The final model was then trained on our dataset. The performance of the model was evaluated using precision, recall, f1-score, and support metrics, alongside a confusion matrix. Our results showed a precision of 0.90 for class 0 (churn) and 0.95 for class 1 (retention), with recall scores of 0.96 and 0.89, respectively. The model achieved an overall accuracy of 92%, with balanced macro and weighted averages for precision, recall, and F1-score at 0.93 and 0.92. The confusion matrix further validated the model's effectiveness, showing a high rate of correct predictions for both churn and

retention cases.

These results indicate that our XGBoost model, optimized through meticulous hyperparameter tuning, was highly effective in distinguishing between churn and retention. The balanced performance across various evaluation metrics highlights the model's robustness and reliability in predicting customer behavior in the e-commerce context.

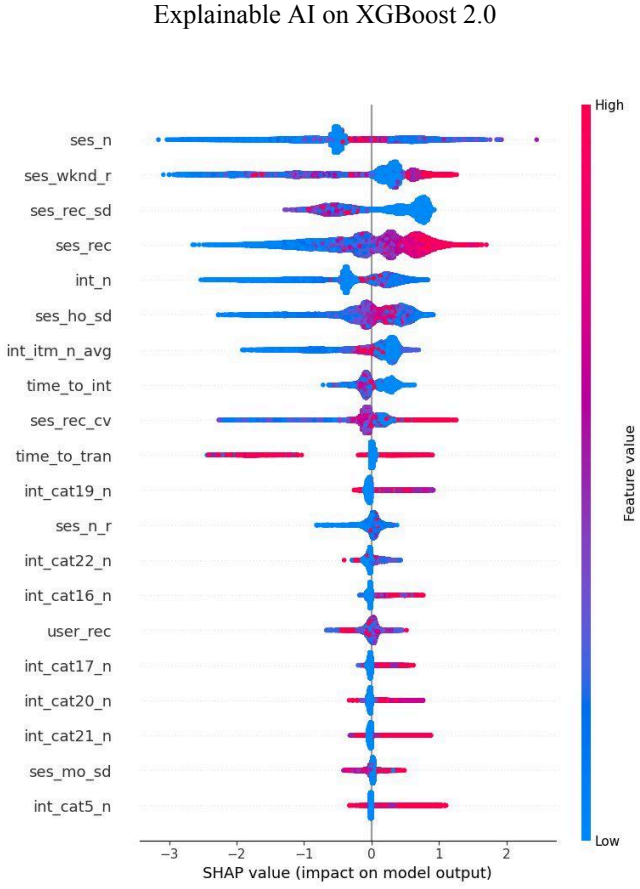

Figure 3. SHAP graph for Xgboost model

The SHAP summary plot derived from the XGBoost model provides a rich visual framework for interpreting the feature influences on the model's predictions. The plot illustrates that the session count (ses_n) is the most influential feature, with higher values strongly swaying the model towards predicting a higher likelihood of churn. The distribution of SHAP values for each feature, depicted by the spread of points, reveals the variance in feature impact across the dataset. For instance, ses_wknd_r (weekend sessions proportion) and ses_rec_sd (standard deviation in time between sessions) demonstrate significant variability, suggesting their effects on the model's output differ from one customer to the next. Features with SHAP values extending towards the right indicate a higher feature value that increases the model's churn prediction, while values to the left indicate a decrease. The color intensity represents the magnitude of a feature's value, with red indicating higher and blue lower values. This nuanced depiction allows us to discern not only the directionality and strength of each feature's influence but also the heterogeneity of these effects across individual predictions, providing a comprehensive understanding of the model's behavior.

B. Support Vector Machine (SVM)

In our investigation, we evaluated the effectiveness of a Support Vector Machine (SVM) classifier, known for its proficiency with high-dimensional data. We first normalized the feature space using the StandardScaler, a crucial step for SVM, which relies on the geometry of the data.

The SVM model, implemented with a linear kernel, was chosen for its interpretability and to discern linear relationships between variables. After training, we assessed its performance on test data. The model's classification ability was examined using a confusion matrix and classification report. It showed a precision of 70% for non-churn and 67% for churn predictions, with recall rates of 64% and 73% respectively, indicating a stronger tendency to identify churn cases correctly.

The F1-scores were 67% for non-churn and 70% for churn, reflecting a moderately better classification of churn instances. The overall accuracy stood at 68%, showing reasonable effectiveness in distinguishing between churn and retention. The confusion matrix detailed 2,557 true negatives, 1,439 false positives, 1,101 false negatives, and 2,903 true positives, highlighting the model's higher rate of false positives but notable accuracy in identifying true churn cases.

This analysis of the SVM classifier, employing feature scaling and a linear kernel, demonstrates its potential in customer churn prediction. The balanced performance, alongside insights from the confusion matrix and classification metrics, suggests avenues for model improvement. Future refinements could involve experimenting with different kernels for non-linear patterns or adjusting parameters to better balance recall and precision, aiming to enhance the SVM's predictive accuracy for practical business use.

C. Logistic Regression

In our study, we utilized a Logistic Regression model to analyze customer churn prediction. Known for its simplicity and interpretability, Logistic Regression is a reliable method for binary classification. We adopted a One-vs-Rest approach, effectively transforming our multi-class problem into several binary classification problems.

After training the model with our preprocessed data, we evaluated its performance on a test set using a confusion matrix and classification report. The model achieved an accuracy of 68%, with precision scores of 70% for non-churn and 67% for churn predictions. This indicates a slightly better precision for non-churn predictions. The recall scores were

64% for non-churn and 72% for churn, showing a higher sensitivity in identifying churn cases.

The F1-scores, which are a blend of precision and recall, were 67% for non-churn and 69% for churn, indicating a modestly better performance for churn predictions. The confusion matrix showed 2,576 true negatives and 2,881 true positives, along with 1,420 false positives and 1,123 false negatives.

Despite its straightforward nature, the Logistic Regression model proved to be effective in churn prediction. This approach not only allowed us to accurately classify churn cases but also provided valuable insights for potential improvements. Its interpretability is especially beneficial for stakeholders looking to understand and address the factors driving customer churn in the e-commerce industry.

D. Random Forest

In our predictive modeling for customer churn, we utilized the Random Forest algorithm, known for its robustness in handling nonlinear relationships and diverse data types. This ensemble method, which constructs numerous decision trees and outputs the mode of the classes for classification, proved to be highly effective for our complex dataset.

We configured the Random Forest model with 100 estimators and it demonstrated excellent predictive performance. The accuracy reached 93.3%, reflecting a high reliability in classifying churn. Precision was high at 91% for non-churn (class '0') and 96% for churn (class '1'), indicating strong predictive accuracy for both classes. The recall was equally impressive at 96% for non-churn and 90% for churn, showing the model's ability to identify true instances accurately.

The F1-scores were 94% for non-churn and 93% for churn, highlighting the model's balanced performance. The confusion matrix showed 3,856 true negatives, 3,608 true positives, 396 false negatives, and 140 false positives, confirming the model's effectiveness with few misclassifications.

The Random Forest classifier's robust performance is particularly suitable for complex tasks like churn prediction in e-commerce. Its high accuracy and balanced precision and recall demonstrate its strength in correctly identifying churn, offering valuable insights for businesses to enhance customer retention strategies. Such a model can be instrumental in early identification and proactive engagement with at-risk customers, potentially reducing churn rates significantly.

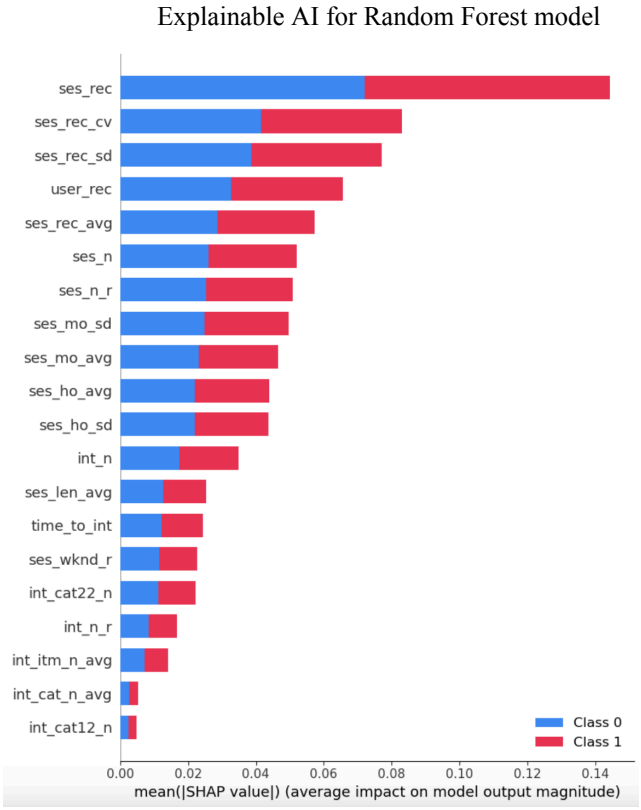

Figure 3. SHAP graph for Xgboost model

The SHAP summary plot effectively illustrates how different features influence our predictive model's output, distinguishing between factors that contribute to customer churn (Class 1) and retention (Class 0). Notably, 'ses_rec', indicating session recency, stands out as the most impactful predictor, pushing the model towards churn prediction with higher values. In contrast, 'int_cat12_n' shows minimal impact on the prediction, suggesting limited predictive relevance. Features like 'user_rec' and 'ses_rec_avg' also significantly increase churn likelihood, marking them as critical indicators for customer retention strategies. This plot highlights the importance of session-related factors and user interactions in driving churn predictions, providing valuable insights for targeted interventions to reduce customer churn.

Algorithms	*F1 Score*	*Accuracy*	*Precision*	*Recall*
XGBoost	93%	92%	90%	96%
SVM	67%	68%	70%	64%
Logistic Regression	67%	68%	70%	64%
Random Forest	94%	93%	91%	96%

Figure 4. Comparison of Algorithms

E. Ensemble Learning

In our research, we explored the effectiveness of ensemble methods in improving churn prediction accuracy. To this end, we created a Voting Classifier ensemble, integrating three distinct models: XGBoost, Random Forest, and Logistic Regression. This approach leverages the strengths of individual models, with the expectation that the combined output would outperform any single model.

We configured our Voting Classifier to use 'hard' voting, which bases its predictions on the majority vote from the component models. After training this ensemble on our dataset, we tested its performance on a separate test set. The ensemble model's accuracy was measured using the accuracy_score metric from sklearn.metrics.

The results were promising, indicating a significant improvement in predictive accuracy. The ensemble model achieved an overall accuracy of 92% on the test set. Breaking down the performance further, we observed a precision of 90% for predicting non-churn (class 0) and 96% for churn (class 1). This suggests that the ensemble was slightly more precise in identifying churn cases. The recall rates were 96% for non-churn and 89% for churn, showing a higher tendency to correctly identify non-churn cases.

The F1-scores, which balance precision and recall, stood at 93% for non-churn and 92% for churn, reflecting a well-rounded performance across both classes. Our confusion matrix provided deeper insights: out of the 3996 non-churn cases, 3829 were correctly identified, with only 167 misclassified. For the 4004 churn cases, 3556 were accurately predicted, with 448 misclassified.

This ensemble approach, combining XGBoost, Random Forest, and Logistic Regression, demonstrated its efficacy in our study, offering a robust and balanced tool for churn prediction. The high accuracy and balanced precision and recall underscore the potential of ensemble models in addressing complex predictive tasks like customer churn in the e-commerce sector.

F. Receiver Operating Characteristics (ROC Curve)

In the comparative evaluation of predictive models for customer churn, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) serve as critical measures of model performance, particularly in terms of their ability to distinguish between the two classes: churn and non-churn.

The ROC curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The AUC represents the degree to which the model is capable of differentiating between the classes, with a higher AUC indicating better performance.

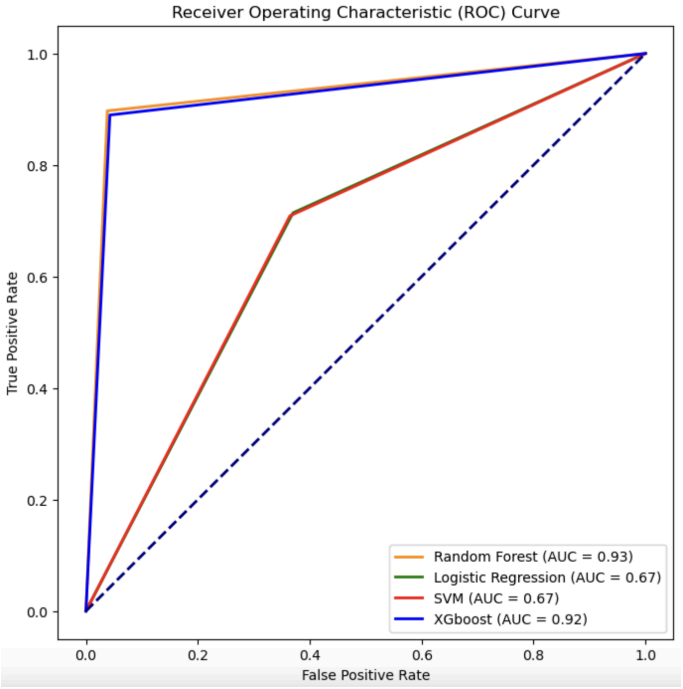

Figure 4. ROC Curve

The Random Forest model exhibits an Area Under the Curve (AUC) of 0.93. This high AUC value indicates that the Random Forest model has a very good measure of separability and is excellent at distinguishing between customers who will churn and those who will not. The model's ROC curve maintains a significant distance from the diagonal line of no-discrimination, which represents the performance of a random guess.

Logistic Regression and Support Vector Machine (SVM) models both have an AUC of 0.67. These values suggest a moderate ability to discriminate between the classes. While better than random chance, these models might not capture the complexities inherent in the dataset, which could limit their effectiveness in accurately predicting churn.

XGBoost's performance, with an AUC of 0.92, is quite similar to that of the Random Forest model, indicating that it too provides an excellent classification capability. XGBoost is nearly as effective as the Random Forest model, suggesting that gradient boosting techniques are also well-suited to the task at hand.

In summary, the ROC curve analysis indicates that the ensemble methods, Random Forest and XGBoost, are superior in predicting customer churn for this particular dataset. The less complex models, Logistic Regression and SVM, while adequate, do not perform as well as the ensemble

models. This suggests that for tasks requiring nuanced distinction between classes, such as churn prediction, ensemble models may be more appropriate.

VI. CONCLUSION

In conclusion, our research has presented a detailed examination of customer churn on an e-commerce platform, employing various machine learning algorithms to create predictive models. These models underwent extensive preprocessing, which involved feature selection to distill the dataset to its most predictive factors and techniques to address class imbalance. Our study assessed four distinct models—Random Forest, Support Vector Machine (SVM), Logistic Regression, and XGBoost—each carefully scrutinized using a suite of performance metrics.

The Random Forest model was notably distinguished by its robustness, proving to be highly effective with an accuracy of 93.3%, precision of 91% for non-churn, and 96% for churn predictions. The model also demonstrated a high recall rate and F1-score, solidifying its superior predictive performance.

Our experiments with ensemble learning, particularly the Voting Classifier that integrated predictions from all models, notably improved accuracy. This ensemble model, utilizing both 'soft' and 'hard' voting schemes, achieved an accuracy of 92%, illustrating the effectiveness of combining diverse models.

Additionally, the application of SHAP provided interpretability to the predictions, revealing the influence of individual features on the likelihood of churn. This analysis not only aids in the prediction of churn but also informs strategic decisions for customer retention.

The findings underscore the significance of using ensemble methods and advanced feature analysis to enhance the prediction of customer churn. E-commerce businesses can leverage these insights to develop personalized strategies to engage customers and minimize churn. The research demonstrates the potential of machine learning to transform the management of customer relationships and supports data-driven decision-making across various business sectors.

For future work, there is the opportunity to explore additional feature engineering, the use of alternative ensemble techniques, or the real-time application of models. The evolving landscape of machine learning offers the promise of further advancements in predictive accuracy and interpretability, paving the way for even more refined approaches to the challenge of customer churn.

REFERENCES

1. M. Li, "Research on the prediction of e-commerce platform user churn based on Random Forest model," 2022 3rd International Conference on Computer Science and Management Technology (ICCSMT), Shanghai, China, 2022, pp. 34-39, doi: 10.1109/ICCSMT58129.2022.00014.
2. S. Raeisi and H. Sajedi, "E-Commerce Customer Churn Prediction By Gradient Boosted Trees," 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 2020, pp. 055-059, doi: 10.1109/ICCKE50421.2020.9303661.
3. Xiaojun Wu, & Sufang Meng. (2016). *E-commerce customer churn prediction based on improved SMOTE and AdaBoost. 2016 13th International Conference on Service Systems and Service Management (ICSSSM)*. doi:10.1109/icsssm.2016.7538581
4. D. O. Orina, R. Rimiru and W. Mwangi, "A Comparative Study of Predictive Data Mining Techniques for Customer Churn in the Banking Industry," 2023 Intelligent Methods, Systems, and Applications (IMSA), Giza, Egypt, 2023, pp. 222-227, doi: 10.1109/IMSA58542.2023.10217514.
5. M. G. A. Reddy, S. Raghavaraju and P. Lashyry, "Ensemble Approach on the Online Shopping Churn Prediction," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 01-08, doi: 10.1109/ICOEI53556.2022.9776921.