Capstone Project Report: E-Commerce Customer Churn Prediction

Table of Contents

1. Introduction
The aim of this project was to develop a predictive model for customer churn within a retail business. Customer churn is a critical metric for any business, as retaining existing customers is often more cost-effective than acquiring new ones. We focused on understanding and predicting customer churn patterns to help the business proactively take steps to retain valuable customers.

2. Problem Statement
The problem was defined as follows:

Given historical data on customer interactions and purchase behavior, can we build a model to predict whether a customer will churn?

The goal was to create a binary classification model that distinguishes between customers who are likely to churn and those who are likely to remain active.

3. Data Preprocessing
Data preprocessing was a crucial step, including handling missing values, encoding categorical variables, and creating features to capture customer behavior. We also performed random undersampling to address class imbalance, ensuring the model's accuracy was not biased towards the majority class.

4. Exploratory Data Analysis
An exploratory data analysis was conducted to gain insights into customer demographics, purchase patterns, and correlations. This helped us understand which features were most relevant for predicting churn.

5. Feature Engineering
We engineered several features, such as "DaysSinceFirstPurchase," "OrderStdDeviation," and "CumulativeTime," to capture customer-specific characteristics and purchase history. These features were used to enhance the model's predictive power.

6. Modeling and Training Validation
We used an XGBoost classifier as our baseline model, which achieved an impressive validation accuracy of approximately 97.97%. The confusion matrix and classification report provided additional insights into the model's performance.

7. Hyperparameter Tuning with Random Undersampling
Recognizing the importance of mitigating class imbalance, we employed random undersampling of churn customers and conducted hyperparameter tuning to enhance the model's performance further. The ROC AUC score after tuning was 0.871, indicating significant improvement.
With this corrected information, the report accurately reflects the ROC AUC score obtained with the RUS dataset.

8. Model Evaluation and Performance
The model's evaluation included visualizations of the ROC curve, confusion matrix, precision-recall curve, and F1 score vs. threshold. These visualizations provided insights into model performance, suggesting a need for trade-offs between precision and recall.

9. Recommendations

Recommendation 1: Investigate Class Imbalance Techniques

Given the class imbalance, explore other techniques such as oversampling, synthetic data generation, or adjusting class weights to enhance the model's performance further.

Recommendation 2: Fine-Tune Threshold

Fine-tuning the threshold may help optimize the model for specific business goals. Depending on the cost of false positives and false negatives, the threshold can be adjusted to maximize precision, recall, or F1 score.

Recommendation 3: Continuous Monitoring

Churn prediction is an ongoing process. Implement continuous monitoring and retraining of the model to adapt to changing customer behavior and trends.

10. Conclusion
In conclusion, this project addressed the problem of customer churn prediction in the retail business. The project involved data preprocessing, feature engineering, modeling, and

hyperparameter tuning. The model demonstrated promising results, with room for improvement in addressing class imbalance and threshold optimization.

The findings suggest that with further refinement and continuous monitoring, the business can proactively identify and retain customers at risk of churning, ultimately leading to improved customer retention and business growth.