Customer Churn Prediction on E-commerce Online Retail

Executive Summary:

This project analyzed transactional data between 2010 and 2011 from a UK-based e-commerce online retail store. Feature engineering, data cleaning, and exploratory data analysis were performed on the dataset. The data was converted from transactional-level granularity to customer level granularity. Customer churn was defined as not making a purchase within 90 days of 09/12/2011.

The best-performing model achieved strong predictive ability (recall of ~85%, ~80% ROC AUC), with total money spent and total items bought emerging as the most important predictors. Low-spending, low-volume customers had churn rates above 50%, while high-value customers churned less than 15% of the time.

The findings suggest that increased effort should be spent on retaining customers who buy a small number of items and spend a minimal amount of money. Targeted advertisements and coupons for these customers could be useful.

Introduction:

Customer retention is a critical challenge in e-commerce, as acquiring new customers is often more costly than keeping existing ones. Understanding which customers are at risk of leaving and why they churn can help businesses design more effective marketing and loyalty strategies.

This project focuses on churn prediction for a UK-based online retail store. The dataset covers all transactions between December 1, 2010, and December 9, 2011. To frame the analysis, customer churn is defined as a customer who did not make a purchase within 90 days of December 9, 2011.

The main objectives of this project are:

1. To predict the likelihood of customer churn with high accuracy, prioritizing correct identification of retained customers.
2. To identify the key drivers of churn.
3. To provide actionable, data-driven recommendations for reducing churn.

Data Description:

The online retail dataset was obtained from the UCI machine learning repository. It contains transactional data from a UK-based online retail store, with data between December 1, 2010, and December 9, 2011. It contains 541,909 transaction-level records across 8 columns, including InvoiceNo, CustomerID, Quantity, UnitPrice, InvoiceDate, StockCode, Description, and Country. After cleaning the data, it contained information about 4,364 customers. 33% of these customers had churned.

After feature engineering, the following features were used to build the model: "TotalItemsBought", "TotalMoneySpent", "NumberOfCancelledOrders", "NumberOfNonCancelledOrders", "NumberOfInvoices", "MeanPriceOfInvoice", "StdDvOfInvoice", "LargestInvoice", "MeanOfTotalItemsPurchased", "StdDvOfTotalItemsPurchased", "LargestNumOfTotalItemsPurchased", "Country_United Kingdom."

Methodology:

Feature engineering was performed in order to convert the dataset from transactional-level granularity to customer level granularity. A column for churn was created. A customer has churned if they have not made a purchase after September 10, 2011. All rows with missing customer IDs were removed, which eliminated all missing values. Duplicate rows were also removed. The fully cleaned dataset contained 4,364 rows and 13 columns.

Exploratory data analysis was performed to examine the distribution of features and correlation between features. Since all features were heavily skewed, models that required a normal distribution for each feature, such as logistic regression, were avoided.

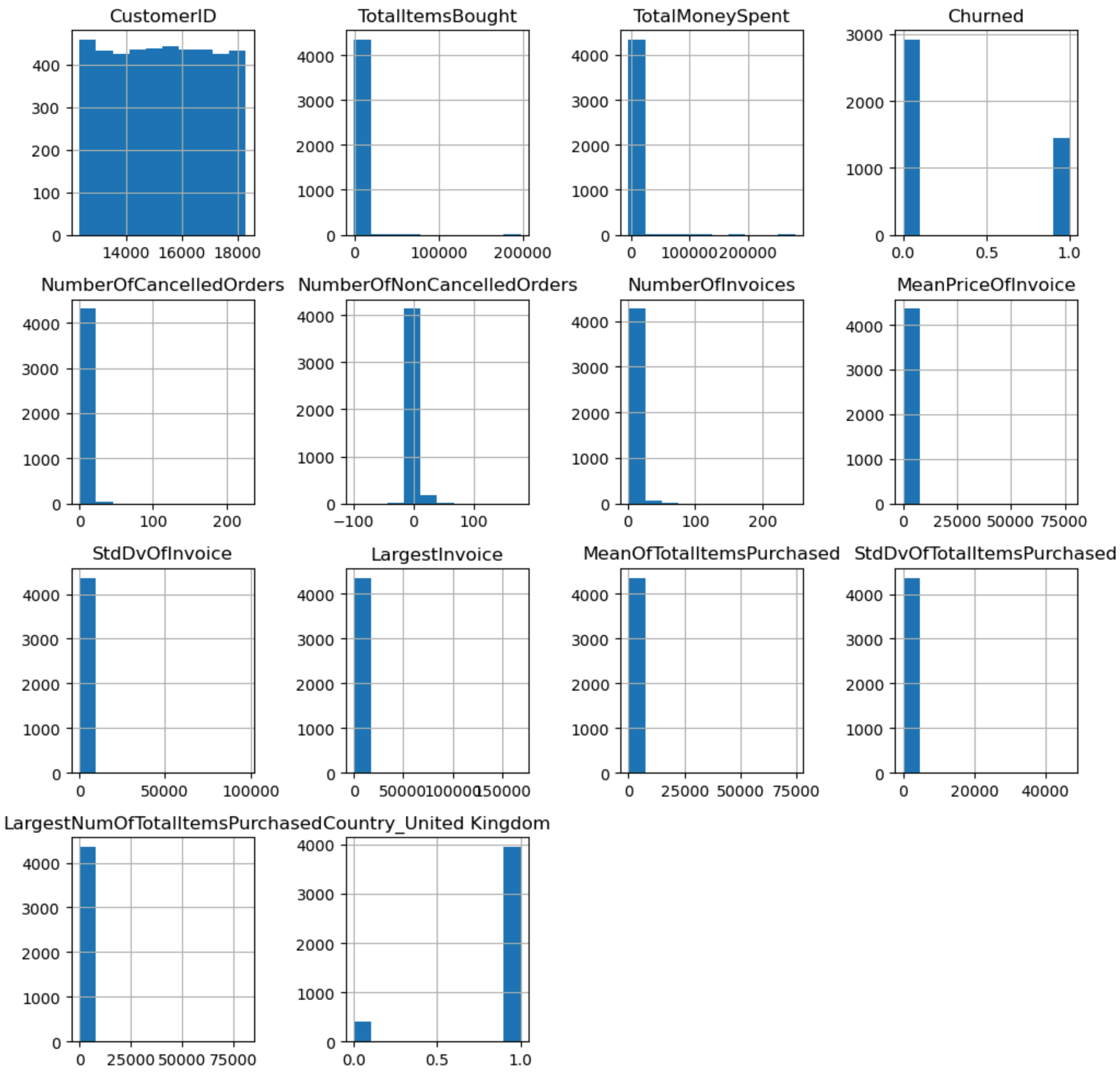
To prepare for building the model, the data was split into a training set (60%), a validation set (20%), and a test set (20%). Feature values were scaled, such that all values were between 0 and 1. The target classes were balanced using SMOTE to synthetically upsample the minority class and tomek links to clean noisy boundary samples.

Several classification models were tested to predict customer churn, including random forest, xgboost, k nearest neighbors, and stacking with voting. Cross validation was performed to tune hyperparameters. The classification threshold for each model was adjusted from the default value of 0.5 to the value that maximized Youden's J statistic, as determined from the ROC curve. The following error metrics were calculated for each model: accuracy, precision, recall, F1 score, ROC AUC, PR AUC, and the confusion matrix. Based on the main objectives of this analysis, recall and ROC AUC were prioritized for evaluating the success of a model's performance to minimize costs and retain as many customers as possible. The metrics for the training set and test set were compared to check for overfitting.

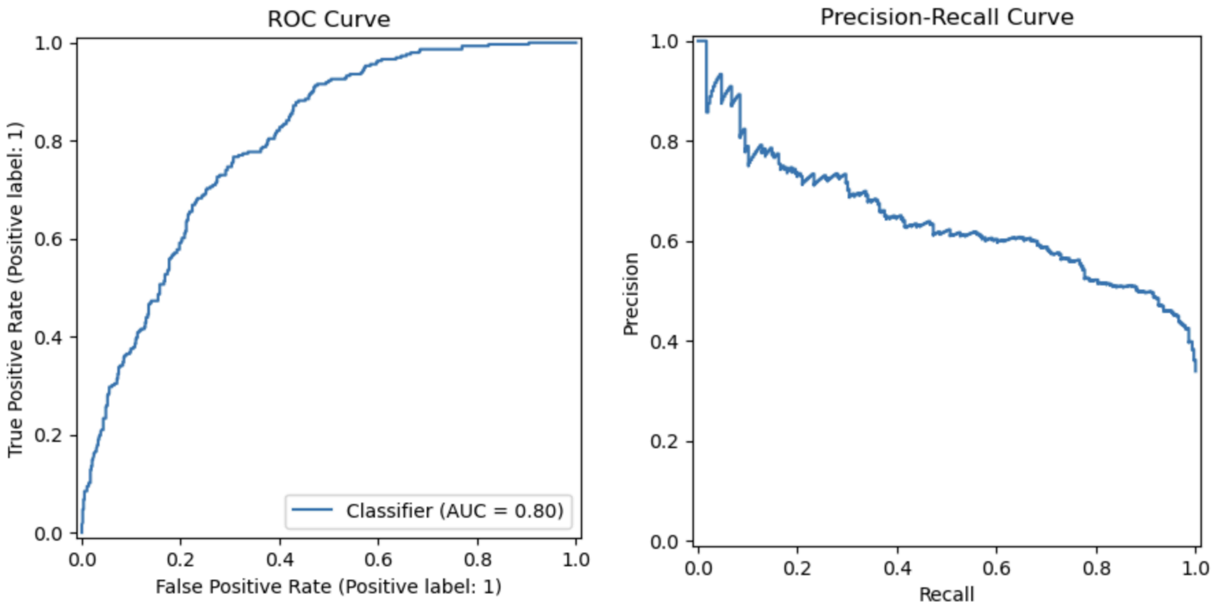
The model that was determined to be best was interpreted with permutation feature importance and partial dependency plots. The most important features were further analyzed to see how they impact customer churn.

Results:

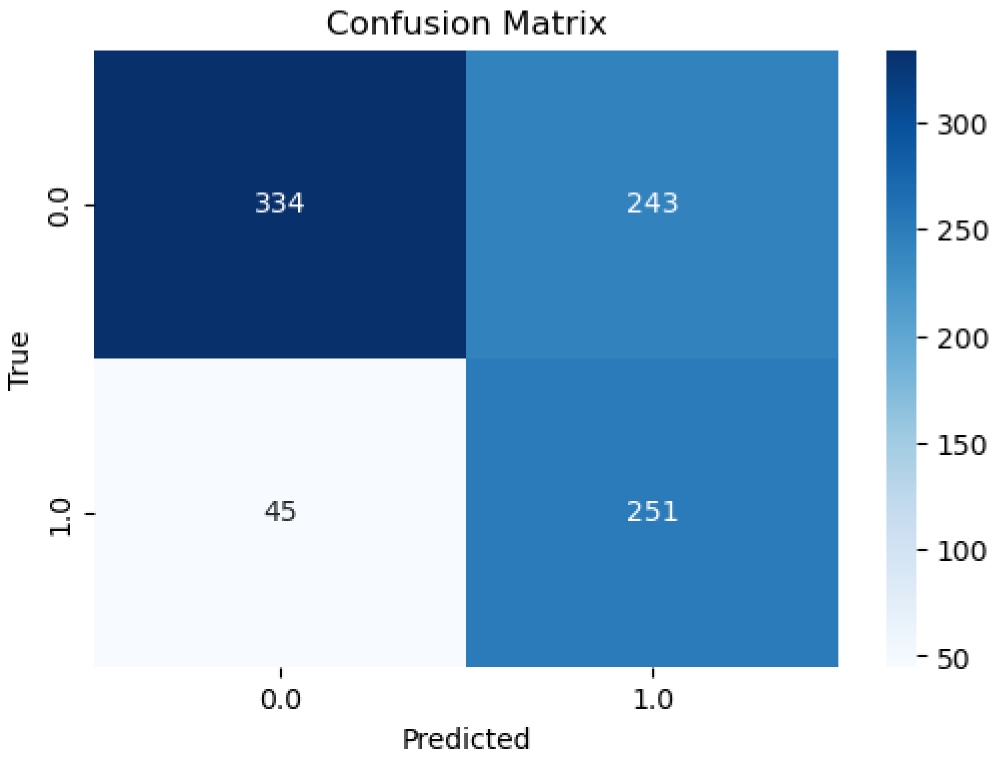
Exploratory data analysis revealed that the distributions of all the features were heavily skewed.



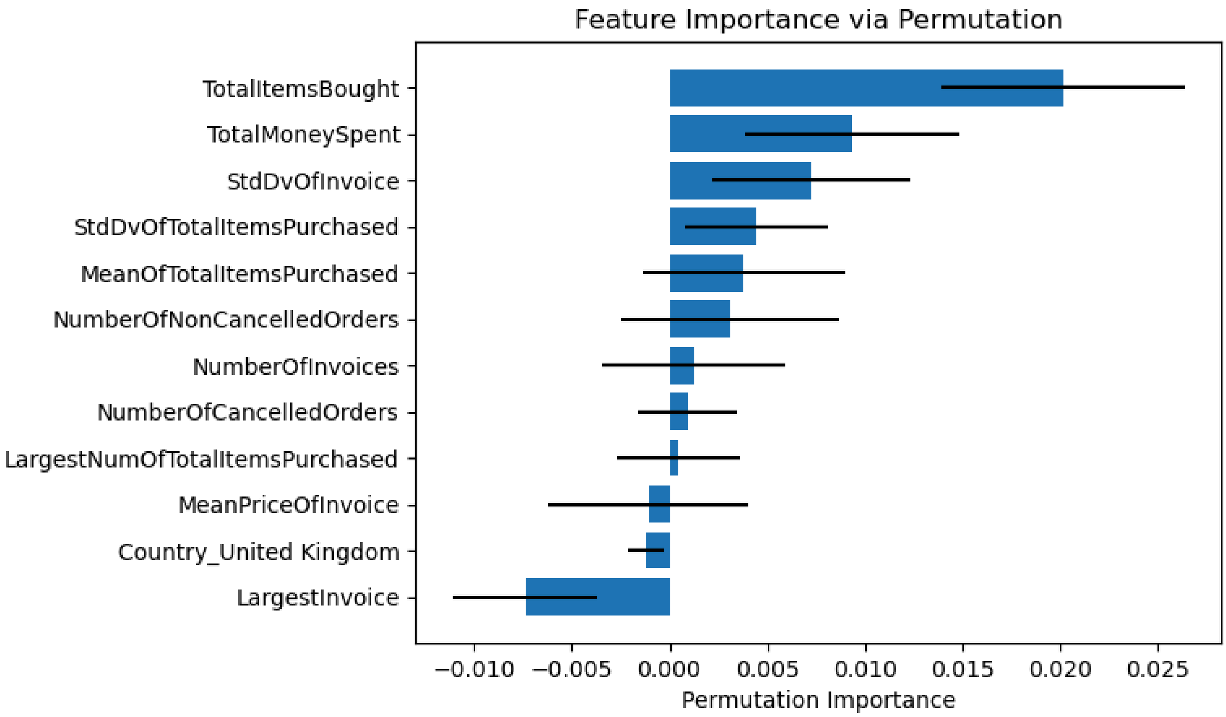
Of the four classification models tested, random forest performed the best. The model achieved recall of 84.8% and ROC AUC of 79.7%, making it the most effective for identifying at-risk customers while minimizing false negatives. It also had an accuracy of 67.0%, precision of 50.8%, F1 score of 63.5%, and PR AUC of 64.2%. The ROC curve and PR curve are shown below.



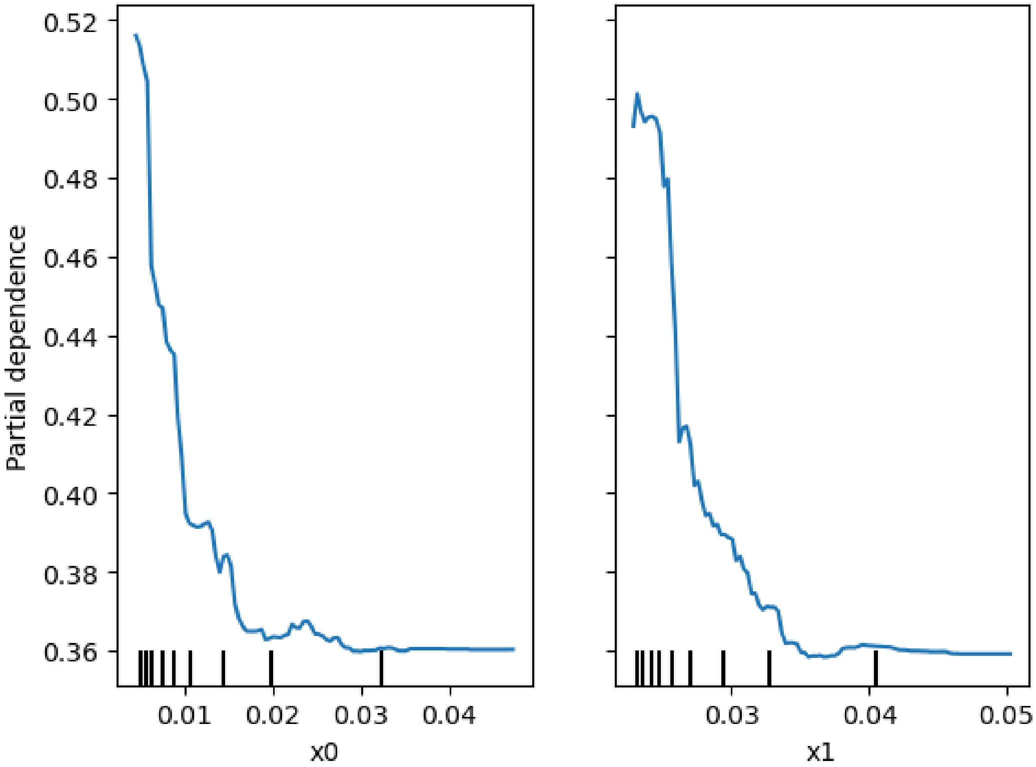
The confusion matrix of the model can be seen below.



The model was also tested on the training data to see if overfitting occurred. The results for the training data were 91.9% recall and 87.0% ROC AUC. This is about 7% higher for both metrics, which indicates mild overfitting, but not enough to be concerned. Permutation feature importance showed that total items bought and total money spent were the two most influential predictors of churn.



The partial dependency plots for those features are shown below (x0 is total items bought, and x1 is total money spent).



Summary statistics of churn rates by these features were analyzed. Customers buying fewer than 200 items or spending less than $300 had churn rates around 56%, whereas customers buying more than 800 items or spending more than $900 had churn rates of only 10–14%.

Discussion:

The analysis demonstrates that the number of items purchased and amount of spending are the strongest predictors of churn. Customers who buy more items and spend more money are significantly less likely to churn, while low-value customers show much higher churn rates. This finding aligns with the expectation that high-spending customers are more loyal, while low-spending shoppers are more likely to lapse.

From a business perspective, this suggests that retention efforts should prioritize low-spending customers. Targeted campaigns such as personalized discounts, coupons, or loyalty rewards could help encourage repeat purchases and reduce churn within this group. The costs of these efforts are minimal compared to the cost of a customer churning.

Some slight overfitting was present in the model. Random forest models tend to overfit on noisy data. More pruning measures can be taken to better generalize the model to unseen data. Collecting more data with stronger predictors may improve the generalizability of the model.

Conclusion and Next Steps:

The random forest model was determined to be the best with a recall of about 85% and a ROC AUC of 80%. The most important features are total items bought and total money spent. The next steps for analyzing this data are to add more features relating to customer exposure to advertisements, such as ad watch time and where the ad was watched. This could give more insight into how advertising affects customer churn.