Report: Capstone Project

Customer Churn Prediction

Executive Summary

**Problem Statement**

Client company as a DTH (Direct to Home) service provider is facing a lot of competition in the current market and it has become a challenge to retain the existing customers in the current situation. Hence, the company wants to develop a model through which they can do churn prediction of the accounts and provide segmented offers to the potential churners. In this company, account churn is a major problem because one account can have multiple customers. Hence by losing one account, the company might be losing more than one customer.

**Solution Approach**

Based on usage and other demographic factors, we will build a prediction model via which we can identify potential churners, and provide recommendations to the company for running targeted marketing campaigns to minimize churn.

**Target Metric**

Our best performing model will be the model that minimizes false negatives, and therefore maximizes the Recall metric. False negatives occur when the model predicts the customer will NOT churn, and then they do. This will cost the company money if the customer attrites. A false positive would be if the model predicts the customer WILL churn, but they do not. In this case, the company would likely spend a small amount of resources to target that customer in the marketing campaign.

**Final Results**

Our best-performing model is the Random Forest Classifier, with a Recall score of .84 on training data, and .844 on validation data. The most important features are “Tenure of the account”, “if the account issued a complaint within the last 12 months”, “days since the account contacted customer care”, and “if the account is Regular Plus”. We have concluded that the marketing campaign should target the following customers:

1. New accounts that were set up within the past 12 months.
2. Accounts that issued at least one complaint within the past 12 months.
3. Accounts that have contacted Customer Care within the past 3 months.
4. ‘Regular Plus’ accounts.

Data Exploration

**Data Overview**

| Variable | Description |
| --- | --- |
| AccountID | Account unique identifier |
| Churn | Account churn flag (Target), 1 - Churn, 0 - No Churn |
| Tenure | Tenure of account in months |
| City\_Tier | Tier of primary customer’s city |
| CC\_Contacted\_L12m | How many times all the customers of the account has contacted customer care within last 12 months |
| Payment | Preferred payment mode of the customers on the account |
| Gender | Gender of the primary customer of the account |
| Service\_Score | Satisfaction score given by customers of the account on service provided by the company |
| Account\_user\_count | Number of customers tagged with this account |
| account\_segment | Account segmentation on the basis of spend |
| CC\_Agent\_Score | Satisfaction score given by customers of the account on customer care service provided by company |
| Marital\_Status | Marital status of the primary customer of the account |
| rev\_per\_month | Monthly average revenue generated by account in last 12 months |
| Complain\_l12m | Any complaints has been raised by account in last 12 months |
| rev\_growth\_yoy | Revenue growth percentage of the account (last 12 months vs last 24 to 13 months) |
| coupon\_used\_l12m | How many times customers have used coupons to do the payments in last 12 months |
| Day\_Since\_CC\_connect | Number of days since no costumers in the account has contacted the customer care |
| cashback\_l12m | Monthly average cashback generated by the account in last 12 months |
| Login\_device | Preferred login device of the customers in the account |

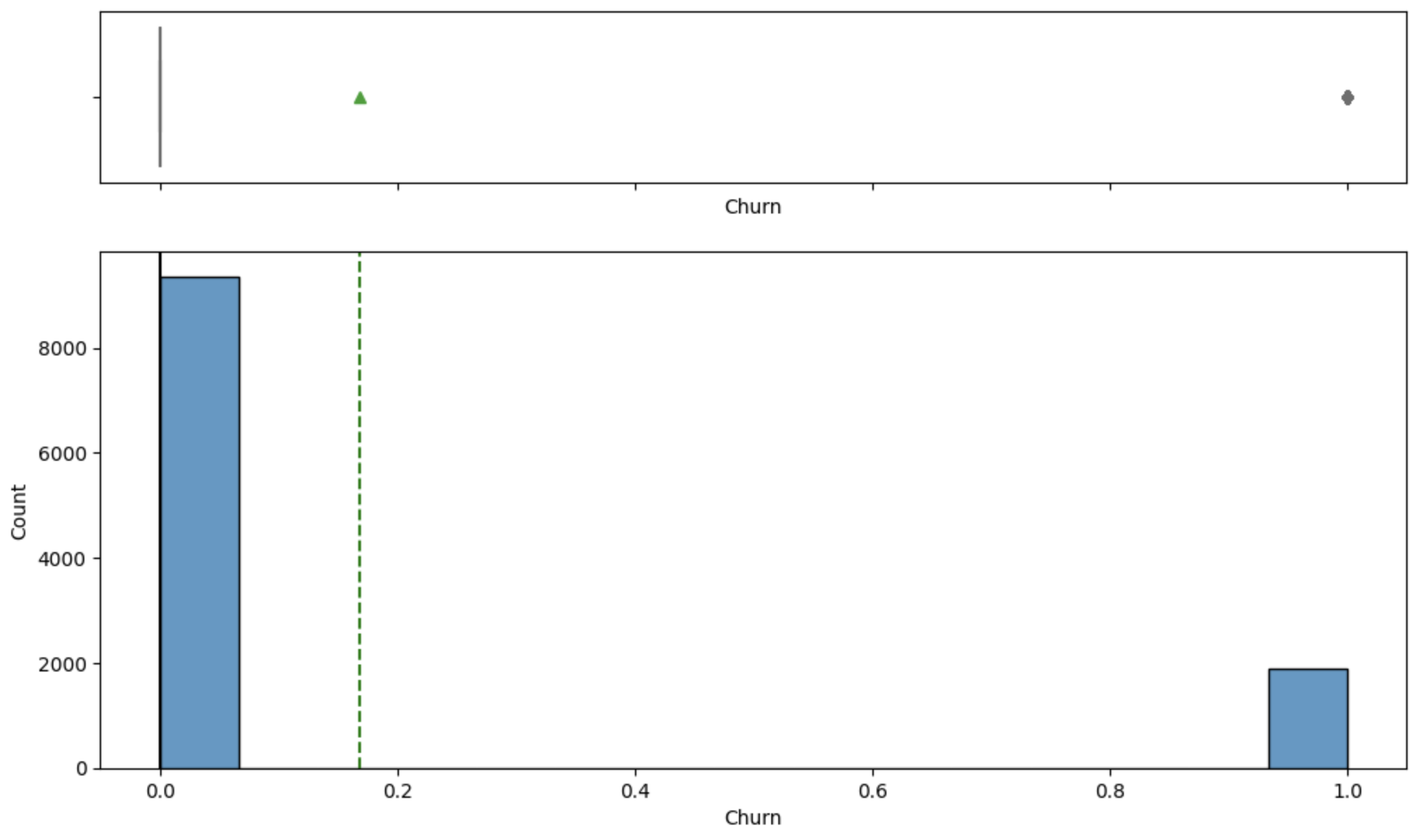
**Data Quality**

There is quite a bit of missing data, as well as data with irregular values. We will deal with the missing/irregular data before creating our models.

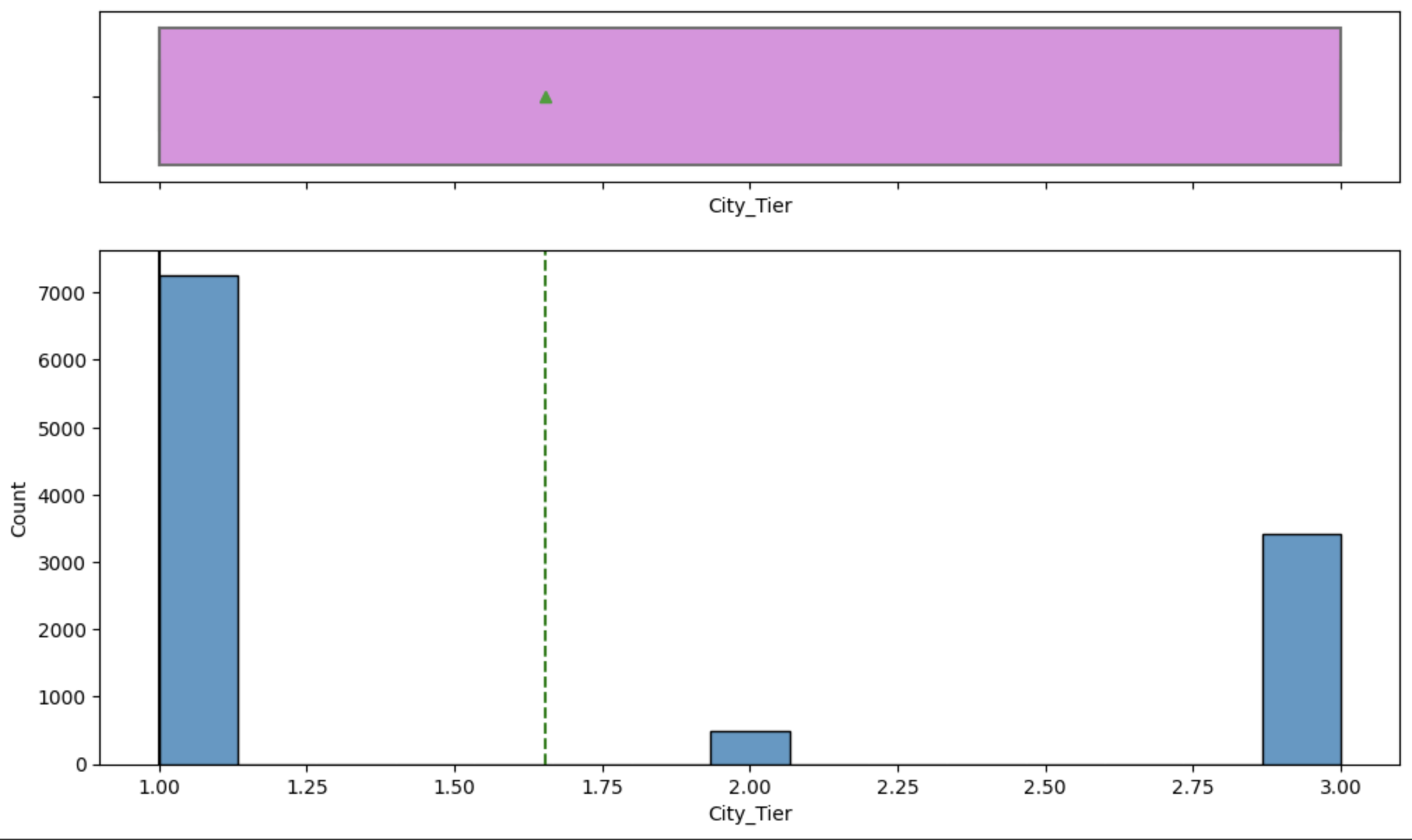
**Exploratory Data Analysis**

**Key Takeaways**

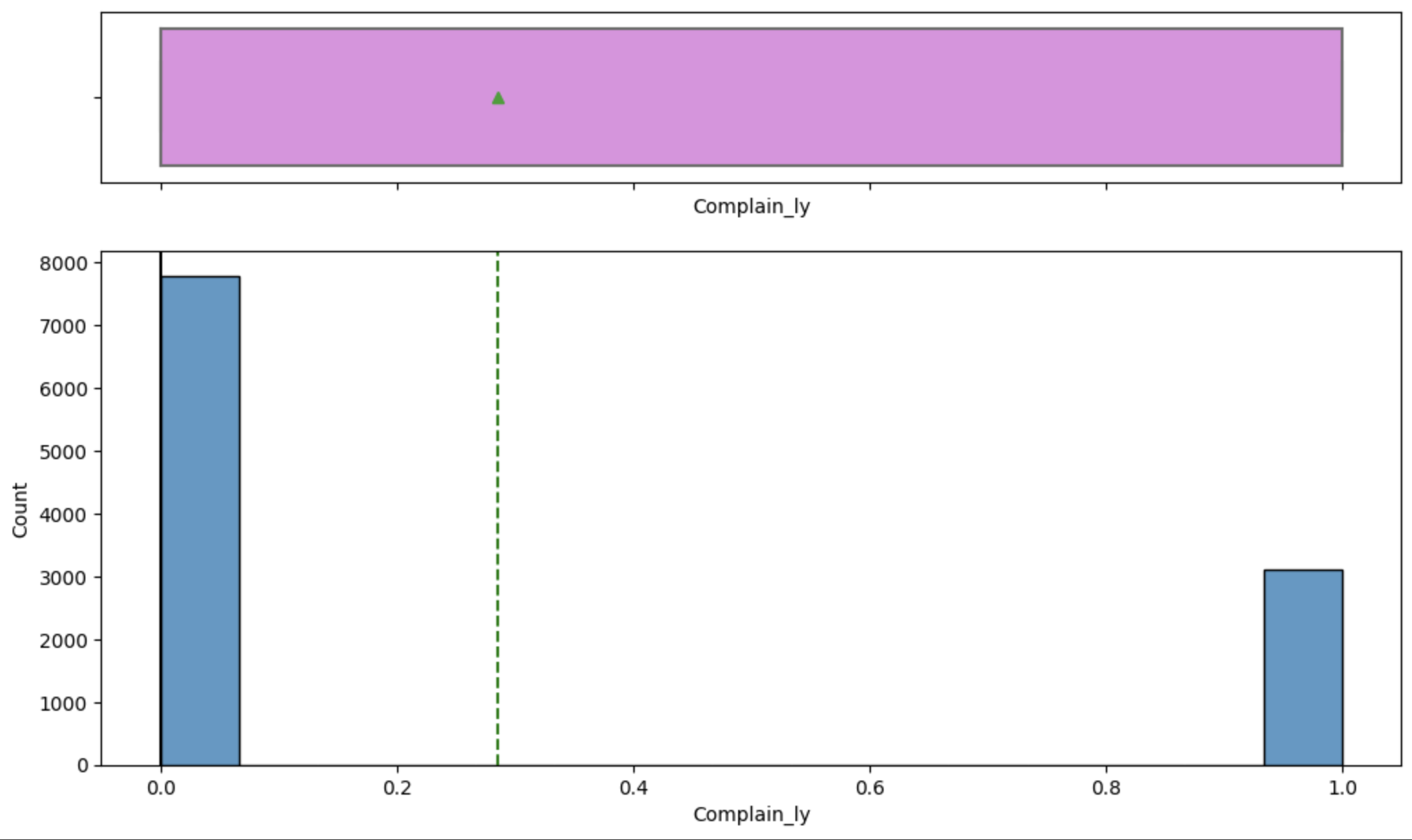
1. Accounts with shorter tenure have higher churn rates.
2. Accounts that pay with Cash or Digital Wallet have higher churn rates.
3. Gender doesn’t affect churn rate much.
4. Accounts with more customers have higher churn rates.
5. ‘Regular Plus’ accounts have the highest churn rate.
6. Accounts with people labled as ‘Single’ have higher churn rates.
7. Accounts with lower rev\_growth\_yoy have higher churn rates.
8. Accounts with higher coupon usage have higher churn rates.
9. Most accounts prefer to use Mobile devices.



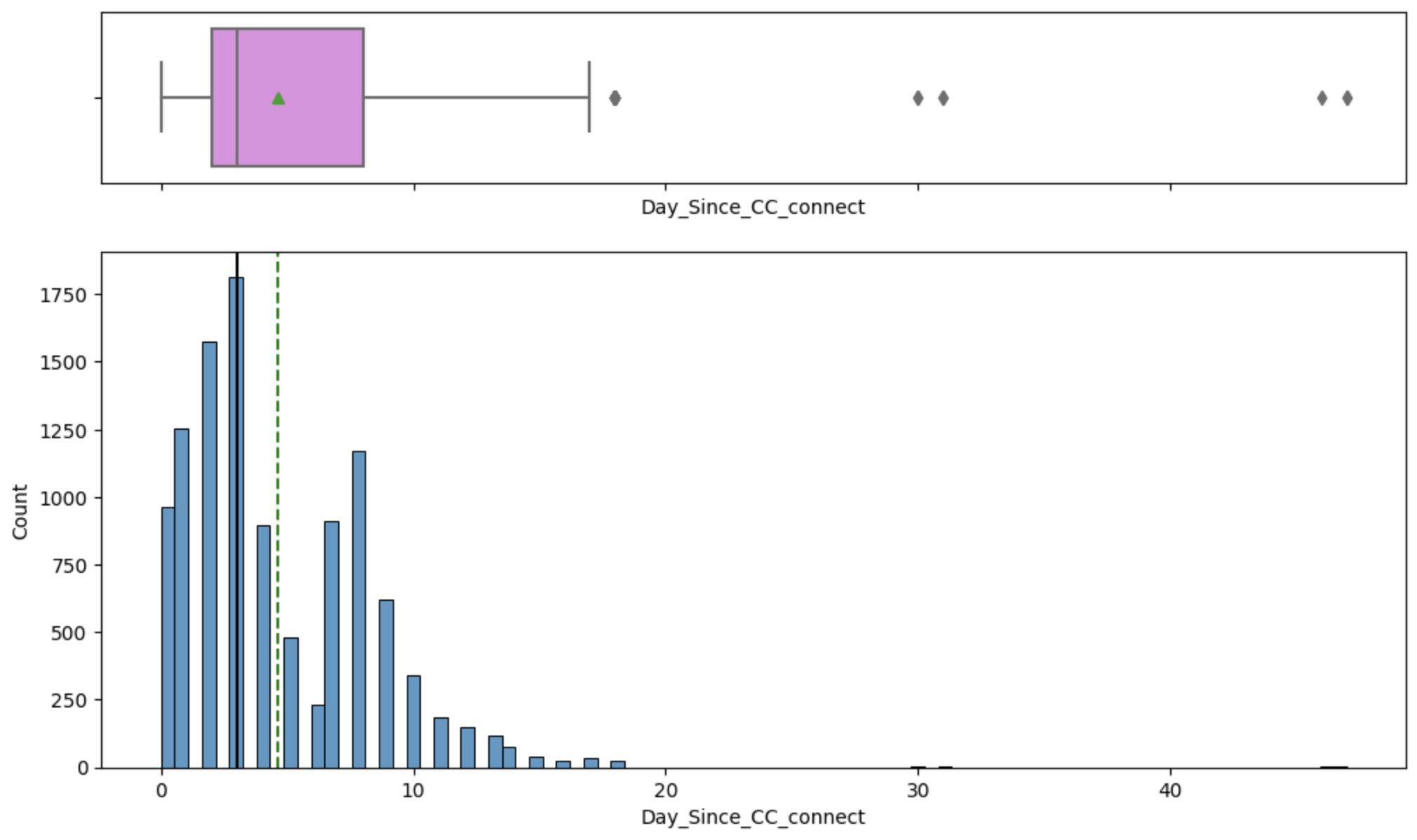
* About 2000 customers churned (about 18%)



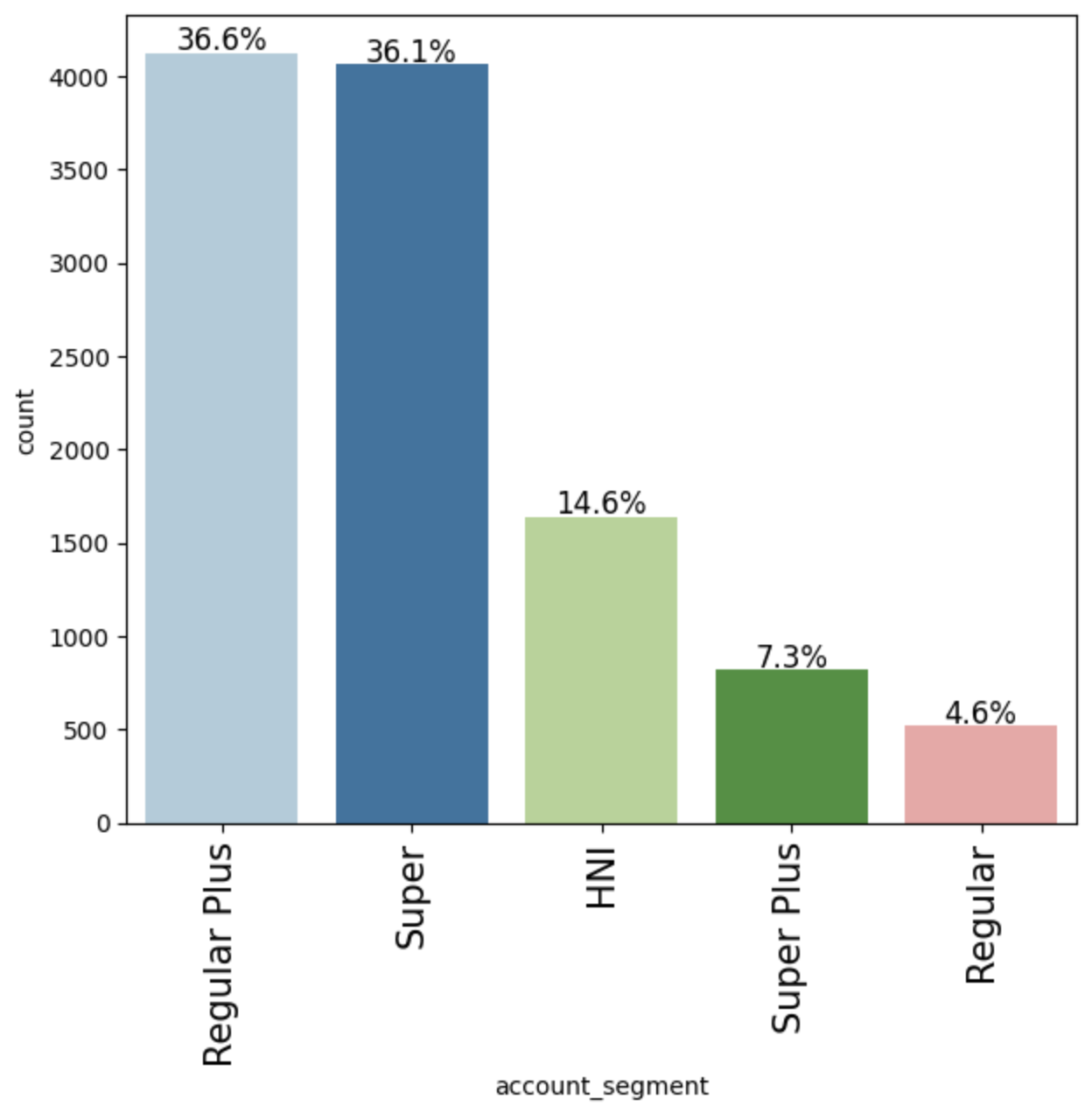
* Most accounts are located in Tier 1 cities.



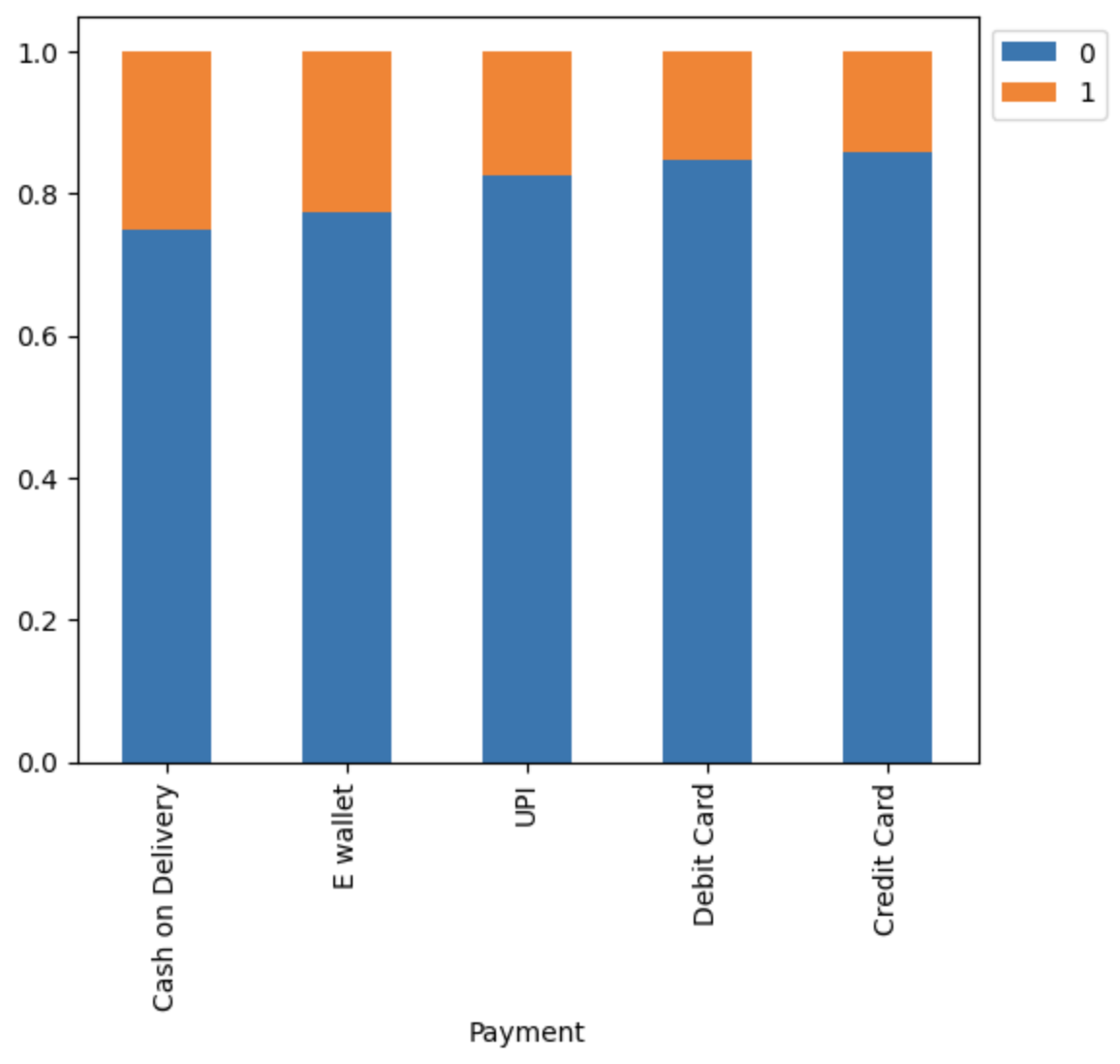
* About 3000 accounts issued a complaint within the past 12 months (about 27%)



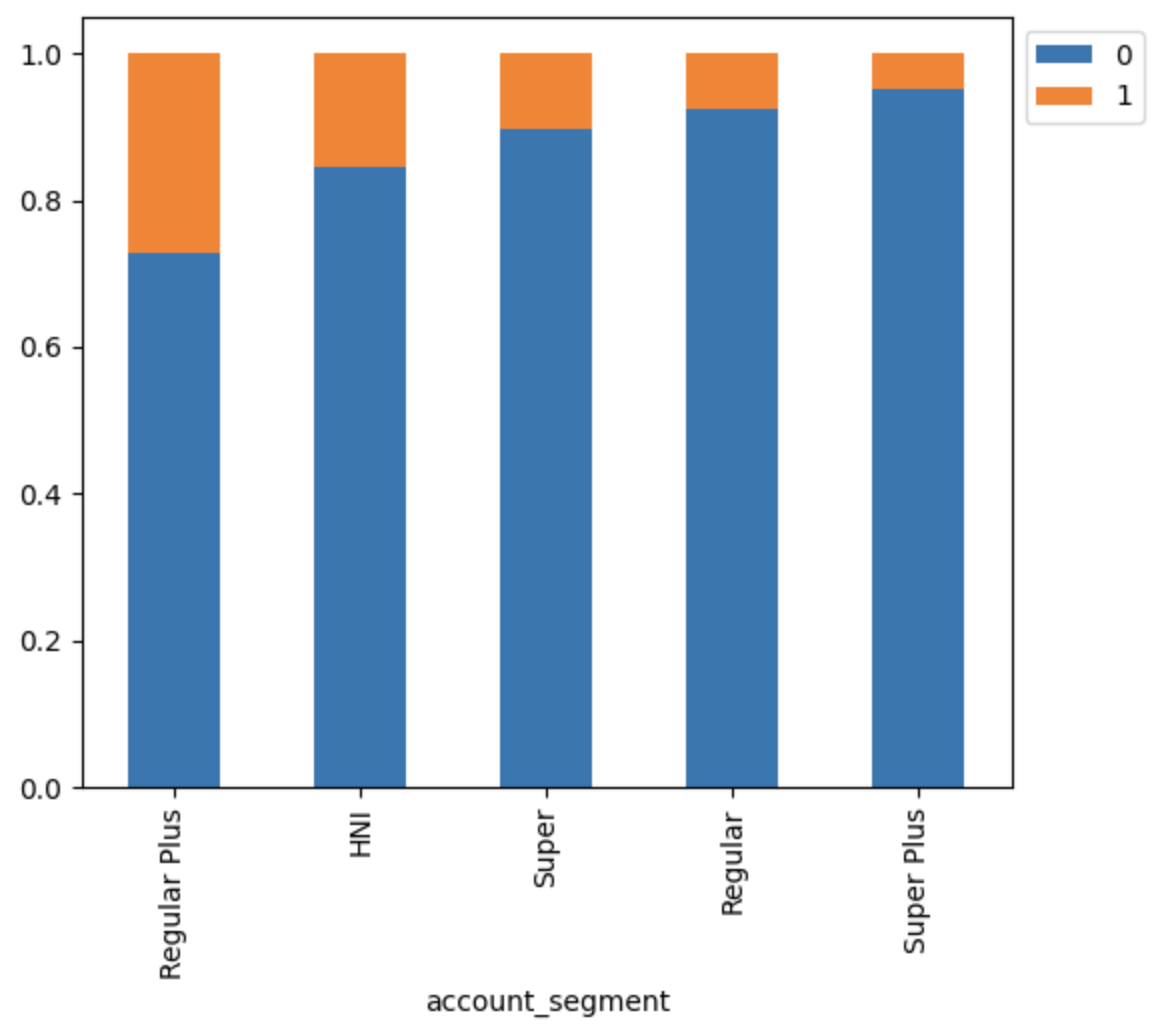
* Most accounts have contacted Customer Care within the past 10 days



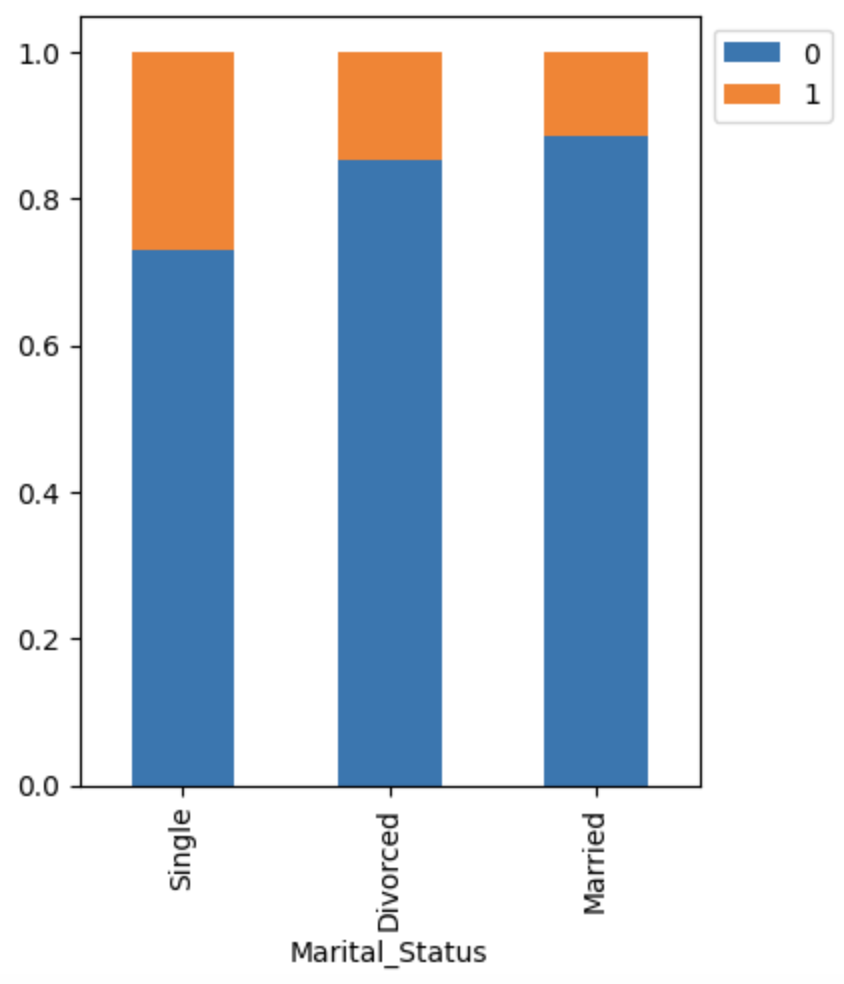
* Most accounts are either ‘Regular Plus’ or ‘Super’



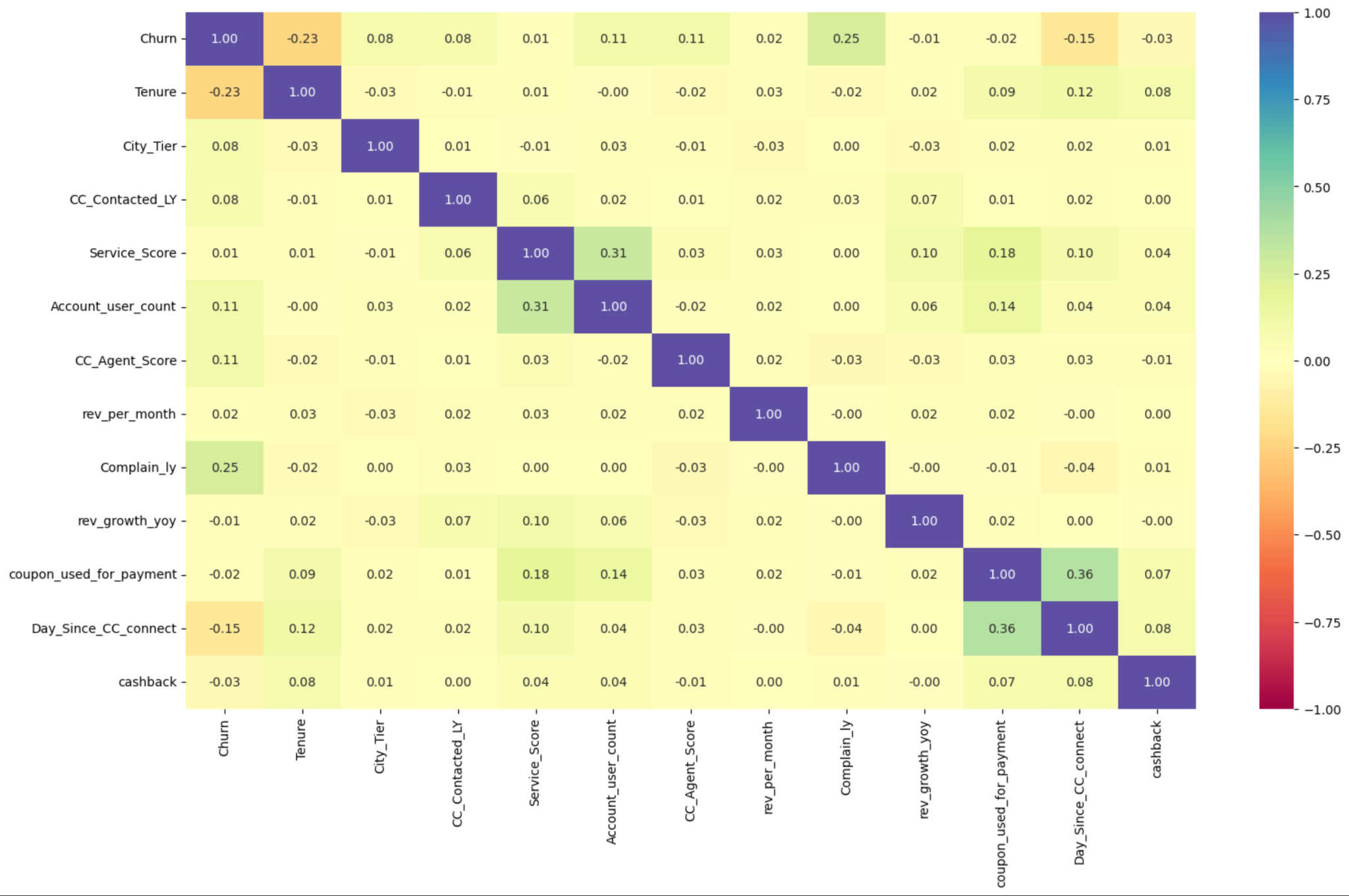
* Accounts that pay via ‘Cash’ or ‘E wallet’ are more likely to churn



* ‘Regular Plus’ accounts are most likely to churn



* ‘Single’ customers are more likely to churn



* Variables mostly do not show strong correlation with each other, so multicollinearity will not be an issue.

Model Building

**Strategy**

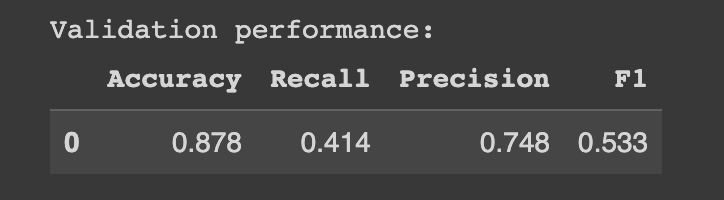
* Split the dataset into training, validation, and test data.
* Impute missing values using KNN Imputer.
* Choose model performance evaluation metric: Recall – (This will allow us to minimize false negatives).
* Build a variety of bagging and boosting classification models
* Test Oversampling and Undersampling.
* Choose the top 3 best-performing models and tune them using RandomizedSearchCV.
* Compare the models and determine the best-performing model.
* Build a pipeline to standardize the model in preparation for fresh data.

**Initial Model Testing**

Using the raw data, we created a logistic regression classification model that returned a Recall score of .407 on Training data, and .414 on Validation data. We used Recall score as our measure of success, because it represents false negatives. Maximizing the Recall score means minimizing false negatives. Since we want to reduce customer churn, sending an offer to a customer who was not intending to attrite (false positive) will not be as costly for the company as if we fail to predict a churner (false negative).

The following is the results from the Logistic Regression model, using the original data:



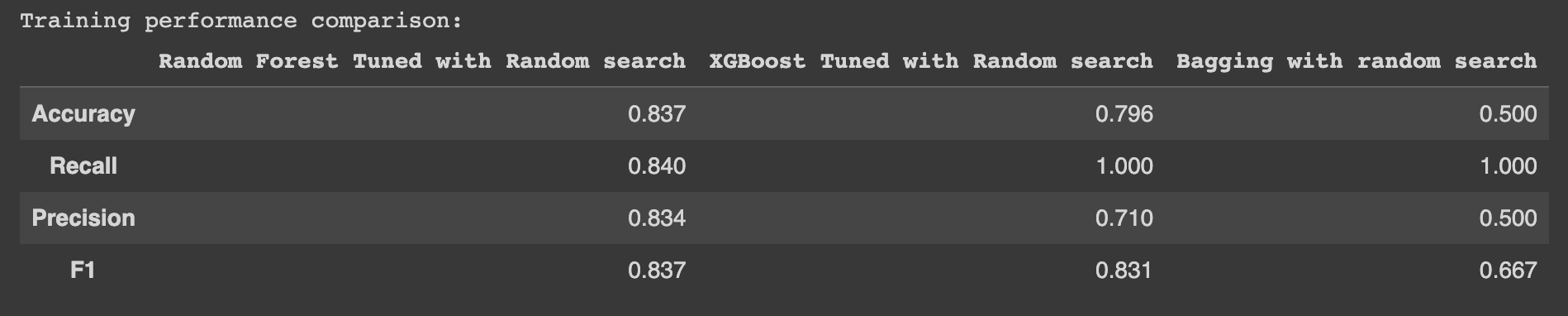
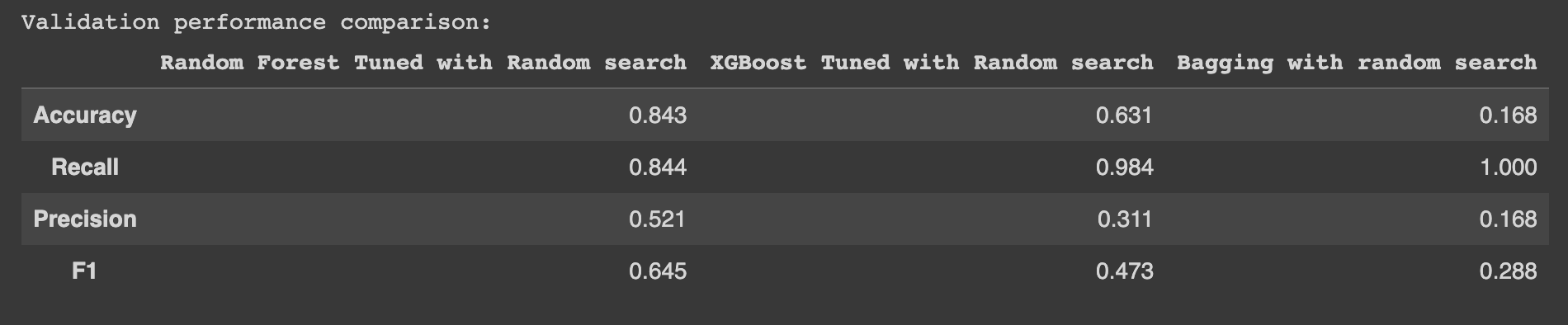


We then tested the Logistic Regression model using oversampled and undersampled data. Undersampled data returned the highest Recall score (.794 Training, .789 Val), which is a 91% increase in performance. We moved forward with the undersampled data.

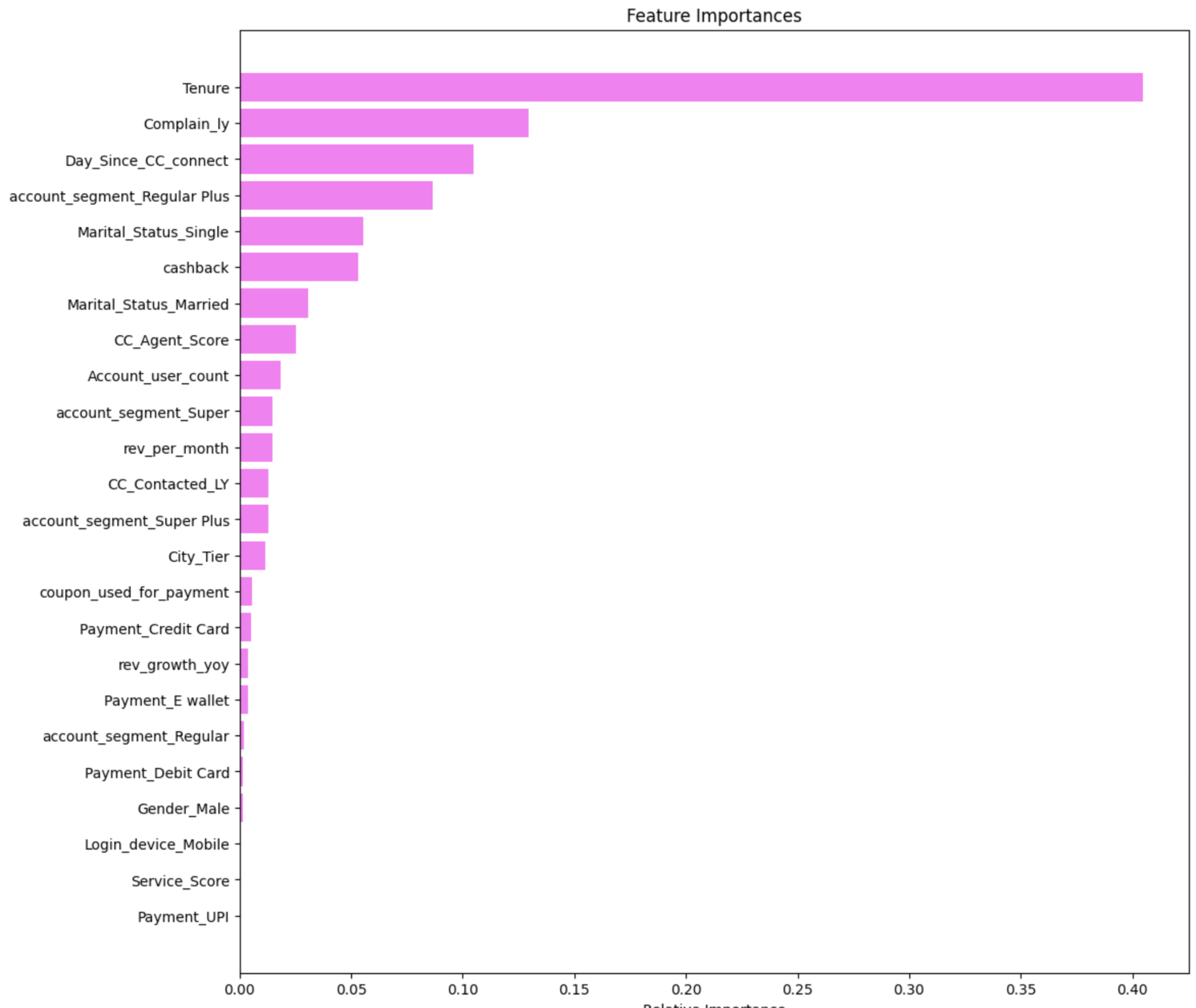
**Model Testing**

We tested 6 different models using the undersampled data. The top 3 models were Random Forest, Bagging, and XGBoost. We then tuned each model’s hyperparameters to optimize results.

**Final Model: Random Forest Classifier**

* Random Forest returned the strongest overall results using undersampled data.
* Below: XGBoost and Bagging have very strong recall, but are overfitting the data.
* 

**Feature Importance**



* Tenure of the account is the most important variable in predicting customer churn.
* Also important features:
  + If the account issued a complaint within the past 12 months.
  + Days since the account contacted Customer Care.
  + If the account is a Regular Plus account.

Business Insights and Recommendations

Our business objective is to decrease customer churn by identifying potential churners and running targeted marketing campaigns. We have identified the top features that indicate a customer is likely to churn:

* Tenure of the account
* If the account issued a complaint within the past 12 months.
* Days since the account contacted Customer Care.
* If the account is a Regular Plus account.

Based on the above features, the following accounts are most likely to attrite, and could be targeted with offers:

* New accounts that were set up within the past 12 months
* Accounts that issued at least one complaint within the past 12 months
* Accounts that have contacted customer care within the past 3 months
* ‘Regular Plus’ accounts