Predict Churning Customers

**- Deeksha Rokle(**[**drokle1@student.gsu.edu**](mailto:drokle1@student.gsu.edu)**)**

**-Shayela Alam([salam14@student.gsu.edu](mailto:salam14@student.gsu.edu) )**

Outline

* Problem Statement…………………. Page 3
* Data Source………………………... Page 3
* Data Preparation…………………… Page 4
* Data Modelling……………………. Page 5
* Evaluation…………………………... Page 8
* Recommendation……………………. Page 11

**Problem statement**

The bank wants to predict what is the reason behind the increasing number of customers leaving their credit card service. The bank wants to be able to analyze and predict which customers are leaving so they can proactively address this issue and provide the customers with better service.

In our project, we explored the bank demographic dataset to find correlations between various variables. We made effort to uncover insightful discoveries throughout this project regarding the bank’s credit card limits, demographic correlation, and other factors that will help us predict the issue.

**Data Source**

We acquired our dataset from Kaggle; the dataset consists of 10,000 customers providing the following information. Attrition flag (whether they are existing or attrited customers), age, gender, number of dependents, education level, marital status, Income, card category, months on book, number of products held by the customer, months active, contacts count, credit limit, total revolving balance, average open to buy, total transaction amount, total transection count with in the last 12 month, and average card utilization count.

Here is the link to the dataset: <https://www.kaggle.com/sakshigoyal7/credit-card-customers>

Using Jupyter notebook we will show the relationship between attrition flag and the relevant columns. We will also show which column has the highest correlation with the attrition flag and predict the reason customers might be leaving the credit card company.

**Data Preparation**

We have first uploaded the dataset into NumPy. After reviewing the data, using the drop function we have removed any rows that have N/A value. Yet, we still had some rows with “unknown” values, which we have removed. Also, we have removed any column that is not needed for our prediction.

Now we had 16 columns and 7081 rows in data including the outcome variable. As too many predictors will make the model complex and may affect the accuracy of prediction. We formed a correlation matrix between variables in the form of heatmap and studied that. Highcollinearity between variables should be avoided. The range of collinearity lies between -1 to 1. Lesser the number, the less the collinearity between variables. We observed that two variables Total\_trans\_Amt-Total Trans ct, Customer Age-Months on Book have high collinearity that is .81. We eliminated one variable from each pair.

A picture containing table

Description automatically generated

**Data Modelling**

Decision Tree Classifier: We considered decision tree model, we also used it for feature selection as well. The following is coefficients of decision tree model.

Graphical user interface, table

Description automatically generated with medium confidence

Chart

Description automatically generated

As we can see with decision tree, we are getting good accuracy but a recall value is also present. In our problem we cannot afford to have recall as this will meant we misclassified existing customers to those who already left, so we decided to drop less imp features and fit again a model.

After applying decision tree classifier, we got feature importance as one of the results. So, we applied the same to our data and got important predictors for our model. Through this method we eliminated few more predictors of very less importance.

Chart, bar chart

Description automatically generated

We eliminated Gender, Dependent count, Education Level, Months\_on\_book, Marital Status, Card Category columns as we can see in the graph above, they have less importance.

Our goal of selecting predictors with high importance is completed now and we can move forward with applying data model. Our outcome variable is a binary value logistics regression suits beat in this case. For applying the model, we applied train test split method from sklearn. Here we will not be considering outcome variable in input, so it’s important to remove the outcome variable from Train, Test X and add it to Train Test y. For implementing model, we will have to convert our data frames to NumPy arrays.

We applied logistic regression on 7 predictors, created a list and bar chart of model coefficients.

Table

Description automatically generated

Here features are the predictors present and coefficients what we get after model is applied. The following bar graph represents a compairision betweeen predictors in terms of coefficients.

Chart

Description automatically generated with medium confidence

The following are the key takeaways from the model:

* Total transaction count is a measure of a total number of transactions made by the card is one of the most important factors to determine customer attrition. The more transactions occur on card, the more chances of customers staying with the bank.
* Total relationship counts which represent the total number of bank products customers are using; also, has good involvement in customer attrition.
* Total revolving balance gives view of balance in credit card of respective customer, having some balance in card contributes to customer attrition.
* Customer attrition also depends on the income category of the person, for example, more the customer earns, the more the chances of maintaining a credit card.
* Credit limit is a reason to stay in business with the bank.
* Next come in line average utilization ratio which represents average amount of time, card is utilized, it also has minimal share in customer attrition.
* Average open to buy represents customers open to buy credit line which has negligible influence on customer attrition.

Now that we have modeled our data with training data and outcome variable. We would want to test our model by predicting the outcome with the help of test data and outcome variable. Following the same procedure as above converting data frame to NumPy arrays. We used to predict function on Test data and calculate predicted outcome.

**Evaluation**

Now that we fit data into a machine learning model to verify its righteousness, we perform evaluation of model. There are various metrices available to evaluate machine learning model. From regression based metrices we picked MSE (Mean square error) and RMSE (Root mean square error). MSE is average of squared differences between the predicted output and the true output. RMSE is root of same value. MSE value in case of our model is: 0.15949188426252647. RMSE value is 0.39936435026492595. RMSE value should be 0 to 1. The lesser the value better the model.

Text

Description automatically generated

Also we applied classification evaluation matrices that is confusion matrix, accuracy, precision and recall. The latter three can be derived from confusion matrix as well.

Chart

Description automatically generated

From confusion matrix we can derive the following conclusion:

* True Positive: when you predict an observation belongs to a class and it does belong to that class. In our model 1189 is the value of true positives.
* True Negative: when you predict an observation does not belong to a class and it does not belong to that class. In our model this value is 2.
* False Positive: when you predict an observation belongs to a class when it does not. In this model the value for above metrices is 226.
* False Negative: when you predict an observation does not belong to a class when in fact it does. In the model above this value is 0.

Accuracy:

A picture containing text

Description automatically generated

Accuracy = TP+TN/(TP+TN+FN+FP) =2+1189/1417=.8405

Precision:

A picture containing diagram

Description automatically generated

Precision = 1189/1191= 0.9983

Recall:

A picture containing diagram

Description automatically generated

Recall = 1189/1189=1

AUC/ROC values: The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

* If AUC=1, the model correctly predicts all positives as positives and all negatives as negative. If AUC=0, then model gets confused between positives as negative and negatives as positives.
* If 0.5<AUC<1, then here is a high chance that the classifier will be able to distinguish the positive class values from the negative class values.
* If AUC<=0.5, the classifier is predicting random class or constant class for all the data points.

In above model we plotted auc/roc curve :

Chart, line chart

Description automatically generated

The AUC value for logistic regression is .8 which is on better side.

**Recommendation:**

After analyzing the data and prediction which we got with the help of model, we came to conclusion that Total transaction count, Total relationship counts, Total revolving balance and average utilization ratio are main predictors. If we summarize in terms of business problem customers with high transaction count, high relationship count, or high revolving balance will not leave bank’s credit card services.

Our top priority in the business problem was to identify customers who are getting churned. Even if we predicted non-churning customers as churned, it won't harm the business. But predicting churning customers as Non-churning will do. So recall (TP/TP+FN) need to be higher, which we tried to achieve in the model.

In conclusion, the bank need to priotize the customers with lower Total transaction count, Total relationship counts, Total revolving balance and average utilization ratio over the year. Provide them with different offers to keep them with the company.