Credit Card Churn Prediction

Project 3 Advanced Machine Learning

Matthew Clark

May 18, 2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Building
- Model Performance Summary
- Model Tuning.
- Appendix

Executive Summary

- The Thera Bank has seen a large decline in their number of credit cards users. To reverse this trend we are analyzing customer data to identify customers who may leave and areas in which the bank can improve to avoid those losses. Based on the findings in our modeling the Data Science team can make several recommendations.

- Our model shows the number of relationship with the bank is an indictor that they may stop using the bank credit card services. Customers that have more than two products with the bank are less likely to leave. It is recommended that to keep credit card customers the bank also aggressively markets additional bank products to credit card customers.

- The amount of contact a customer has with the bank is another indicator that a customer may be leaving the banks credit cards services. With less contact it is more likely you will see attrition, especially when you hit month 4 and later. It is recommended you regularly reach out to customers to encourage and maintain contact.

- A decline in the number of transactions is another indicator that a customer may leave the bank credit card services. It is recommended you closely track the number of transactions and reach out to customers that are using their cards less and provide incentives for more transactions.

Executive Summary (continued)

- The data shows that almost all credit card customers have the Blue card. Less than 2% have Silver, Gold or Platinum. In order to encourage customers to retain their credit cards it may be helpful to reorganize the program so that more customers may earn a higher status level. Having a Gold or Platinum card may encourage them to prioritize the Thera bank cards over others.

- Income and credit limits are not indicators that a customers may be leaving so it is recommended that all customers be incentivized to use and retain their credit card accounts equally.

- The data shows that the majority of customers have income under $40k with limits of less than $4K. Be sure your retentions strategy is focused on customers across all incomes and credit limits.

- The data set provided did not contain information on any rewards offered by the bank for credit card use. However, rewards and incentives are an important part of encouraging credit card use and retention.

Business Problem Overview and Solution Approach

- The Thera Bank has experienced a decline in the number of credit card users. Credit cards are an important line of business for the bank. In order to avoid future losses, the bank seeks to analyze customer data and identify customers that may leave their credit card services and why. This information will allow the bank to improve in the identified areas.
- The Data Science team has been tasked with creating a classification model , using the database of customers, to help the bank improve its services to prevent additional customers from leaving the profitable credit card line of business.
- Our approach to this task will start with processing and analyzing the dataset to ensure accuracy and best modeling. Next, a number of different models will be constructed using original, undersampled, and oversampled data. The output of the different models will be assessed and compared using recall as the primary metric. To select the best model for this task we will tune the models that performed well initially and then compare those results to determine the best and final model for this task.

EDA Results

- The target data Attrition Flag is highly imbalance with 8500 existing customers and 1627 customers that have left.
- Customer age is evenly distributed with a mean age of 46.
- Months on books is evenly distributed with a large spike at the mean of 36 months.
- Credit Limit is right skewed with 50% of customers having a limit under $5000 while some customers have limits well over $30,000. The is understandable considering different incomes and credit histories.
- The Total Revolving balance is slightly left skewed with a mean for $1250. 25% of customers do not carry a balance.
- Avg Open To Buy is right skewed. 50% of customers have less than $3000 available.
- Total Trans Ct and Total Trans Amt are both bimodal with similar distribution. Trans amount has more outliers.
- Total Amt Chng Q4 Q1 and Total Cnt Chng Q4 Q1 are both heavily right skewed.
- Avg Utilization Ratio is heavily right skewed with more than 50% below a 0.2 ratio. Avg Utilization is lower for customers that attrite.

EDA Results (continued)

- 50% of customers have 2 or 3 dependents. Less than 10% have no children.
- 80% of customers have 3 or more relationships with the bank. Customers with fewer relationships are more likely to attrite.
- After 3 months the instances of inactivity drops significantly. When compared to attrition the Cant rate actually decreases with more months of inactivity.
- Around 66% of customers have 2-3 contacts with the bank in the last 12 months. As contacts increase so does the attrition rate.
- Gender shows close to an even split among customer with 5358 Males and 4769 Females.
- Graduate is the most common level of Education.
- Marital Status is fairly evenly split with less than 10% being divorced.
- More than 30% of customers make less that $40K. The remaining customers are split across the other income brackets with the least being above $120K.
- More than 90% of customers have Blue Cards, about 5 percent have Silver Cards, and only 1% have Gold and Platinum Cards.

Data Preprocessing

- No duplicate values were identified.
- The CLIENTNUM column was removed because it had no value.
- The Outliers do appear in a number of columns like Credit Limit(9.717%) and Avg Open To Buy(9.509%).
- However, because the percentages of outliers were relatively low, and the categories could all reasonably be found to contain explainable outliers they were not treated.
- Anomalous values were replaced with NaN in the Income Category.
- The data was then split into training, validation, and testing before further processing in order to avoid any data leakage in the testing data.

Data Preprocessing

- There were missing values in the Education Level and Marital Status Columns. Missing values were imputed using SimpleImputer with the strategy of most frequent.
- One Hot Encoding was used to encode categorical variables for Gender, Education Level, Marital Status, Income Category, and Card Category.
- There were several columns with a large number of unique values. We ran the models with all of the unique values and then we ran the models again after creating bins for those categories using the qcut function. We found the scoring to be nearly identical and determined that creating bins was not necessary to create the best model output.

Model Building

- We used six different types of models to look at original data, oversampled data, and undersampled data.
- The models used were:
    - Bagging Classifier
    - Random Forest Classifier
    - Gradient Boosting Classifier
    - AdaBoost Classifier
    - Decision Tree Classifier
    - XGBoost Classifier

Model Performance Summary (original data)

Training Recall Performance:
Bagging: 0.985655737704918
Random forest: 1.0
GBM: 0.875
Adaboost: 0.826844262295082
dtree: 1.0
XGBoost: 1.0

Validation Recall Performance:
Bagging: 0.8067484662576687
Random forest: 0.8128834355828221
GBM: 0.8588957055214724
Adaboost: 0.852760736196319
dtree: 0.7822085889570553
XGBoost: 0.9079754601226994

Model Performance Summary (oversampled data)

Training Recall Performance:
Bagging: 0.9974504804863699
Random forest: 1.0
GBM: 0.9792116101196313
Adaboost: 0.9645028436948421
dtree: 1.0
XGBoost: 1.0

Validation Recall Performance:
Bagging: 0.8773006134969326
Random forest: 0.8588957055214724
GBM: 0.9079754601226994
Adaboost: 0.8926380368098159
dtree: 0.843558282208589
XGBoost: 0.901840490797546

Model Performance Summary (undersampled data)

Training Recall Performance:
Bagging: 0.9907786885245902
Random forest: 1.0
GBM: 0.9795081967213115
Adaboost: 0.9528688524590164
dtree: 1.0
XGBoost: 1.0

Validation Recall Performance:
Bagging: 0.9171779141104295
Random forest: 0.9263803680981595
GBM: 0.9570552147239264
Adaboost: 0.9601226993865031
dtree: 0.8957055214723927
XGBoost: 0.9601226993865031

Model Tuning

- From the initial 18 models we select the 3 that performed best at tuned them to increase performance before making the final model selection.
- We used RandomizedSearchCV to tune each models hyperparameters.
- The models selected for tuning were:
    - Gradient boosting trained with Undersampled data
    - AdaBoost trained with Undersampled data
    - Gradient boosting trained with Oversampled data

Model Tuning

- Tuned Model Scoring:

AdaBoost with Undersampled Data Tuned			
Training			
Accuracy	Recall	Precision	F1
0.991	0.997	0.985	0.991
Validation			
0.938	0.969	0.731	0.834
Gradient Boosting with Undersampled Data Tuned			
Training			
Accuracy	Recall	Precision	F1
0.978	0.985	0.973	0.979
Validation			
Accuracy	Recall	Precision	F1
0.94	0.957	0.743	0.836
Gradient Boosting with Oversampled Data			
Training			
Accuracy	Recall	Precision	F1
0.975	0.98	0.971	0.975
Validation			
Accuracy	Recall	Precision	F1
0.957	0.911	0.834	0.871

Model Performance Summary (final model)

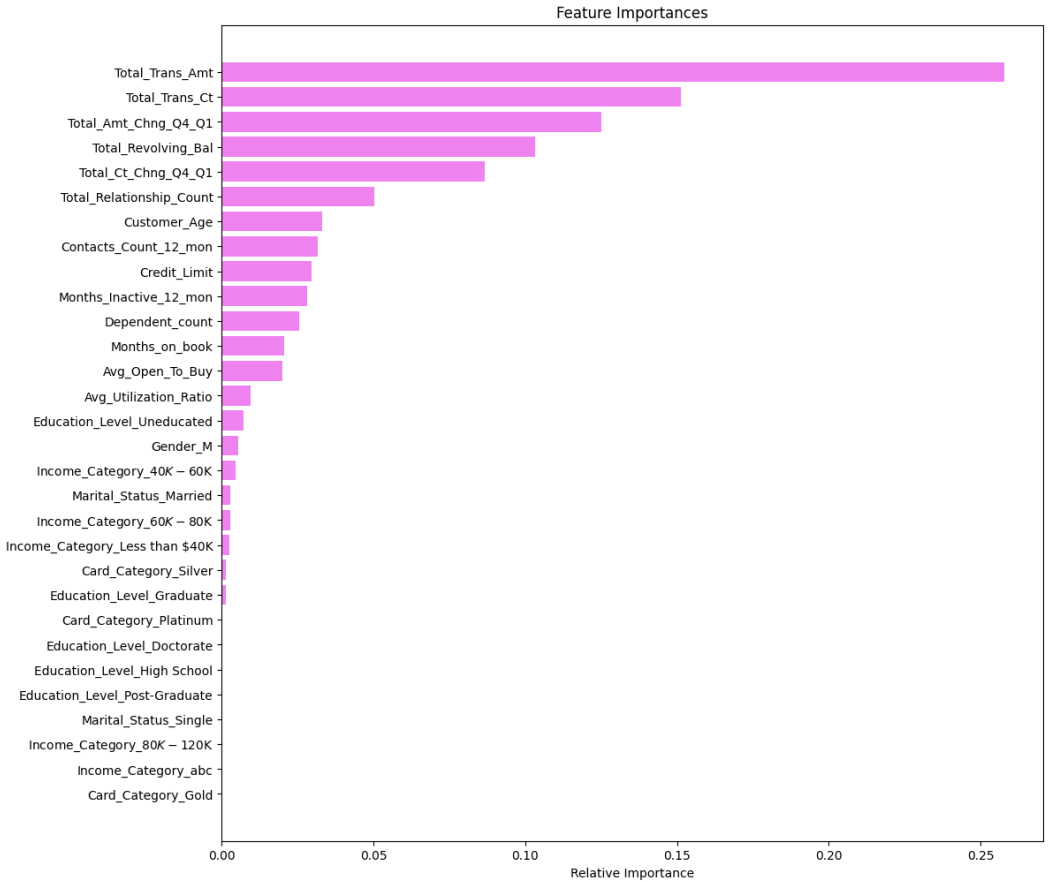
- This model performed the best across training, validation, and after tuning.
- Results of AdaBoost with Undersampled data on test data:

Accuracy	Recall	Precision	F1
0.938	0.966	0.732	0.833

Feature Importance of Final Model

AdaBoosting trained

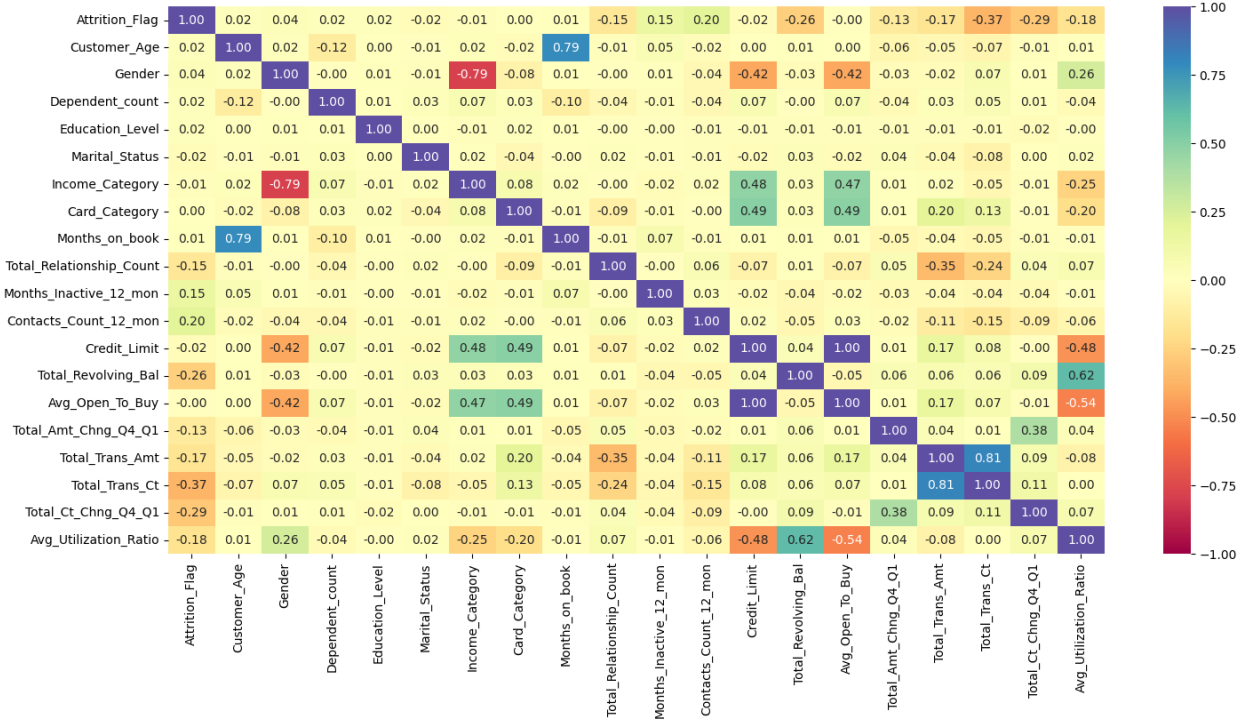
with undersampled data.



APPENDIX

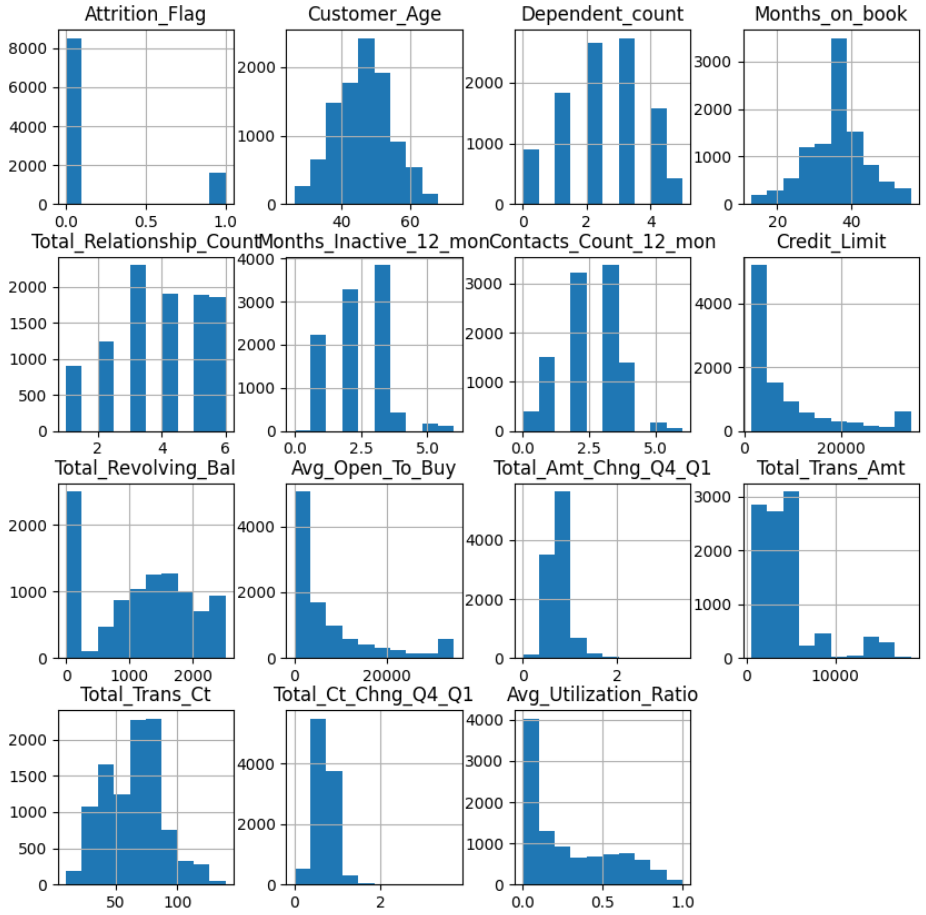
Correlations

- The strongest correlations are between categories with similar data such as Trans_Count and Trans_Amt
- But there are also correlations between Income and Credit Limits and Income and Avg open to buy.



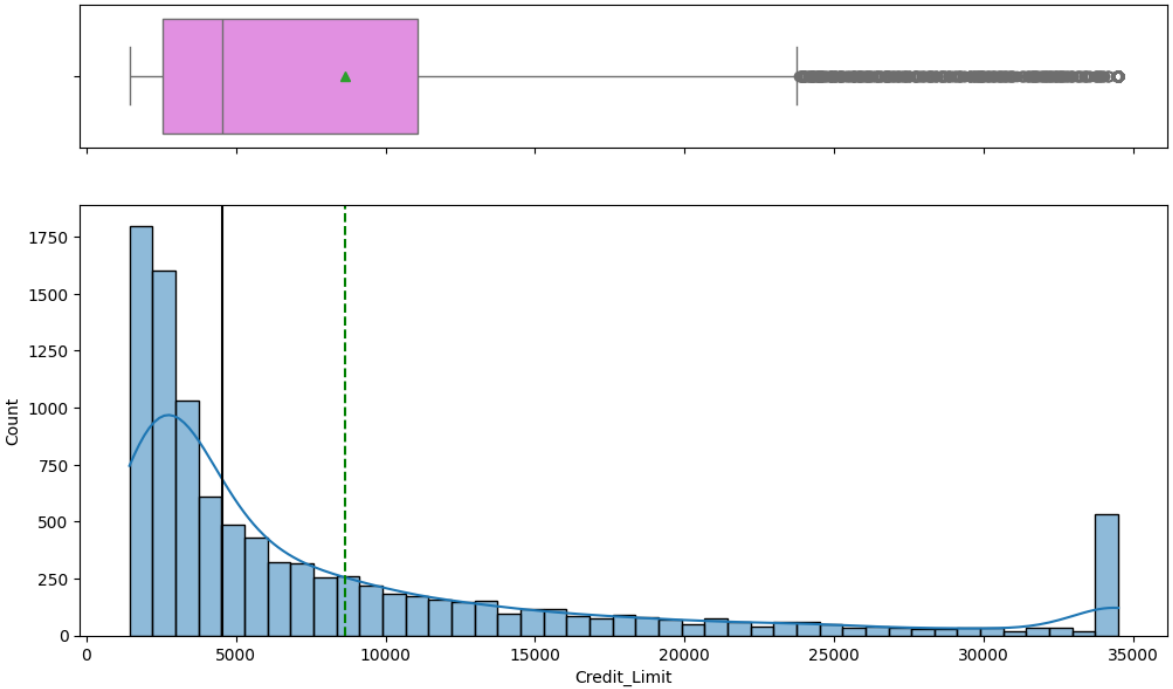
Histograms

- Much of the data is right skewed.
- Most customers have a lower income so it's logical that limits and transaction would be skewed towards lower means.
- Categories less linked to income and limits are more evenly distributed.



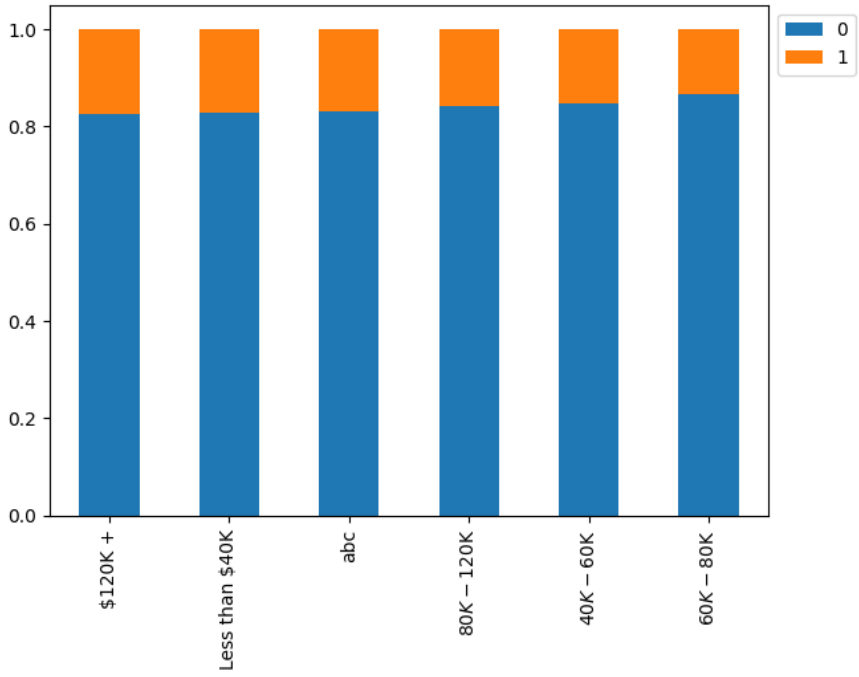
Credit Limit

- The majority of customers have a lower credit limit.
- Credit limit features outliers that make sense in the context of limits based on income.



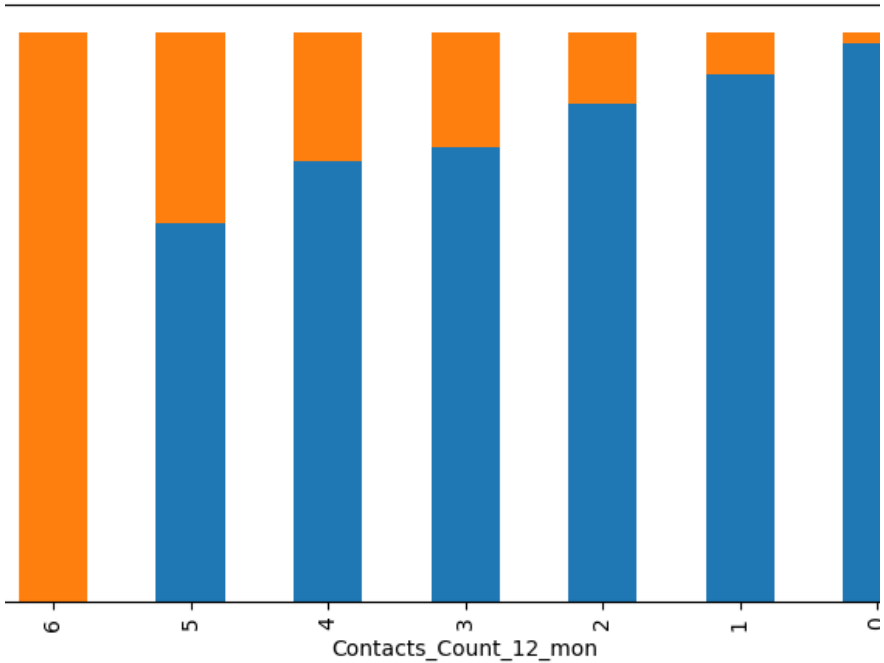
Income vs. Attrition

- Customers making less that $40k and customers making more than $120k have similar rates of attrition.



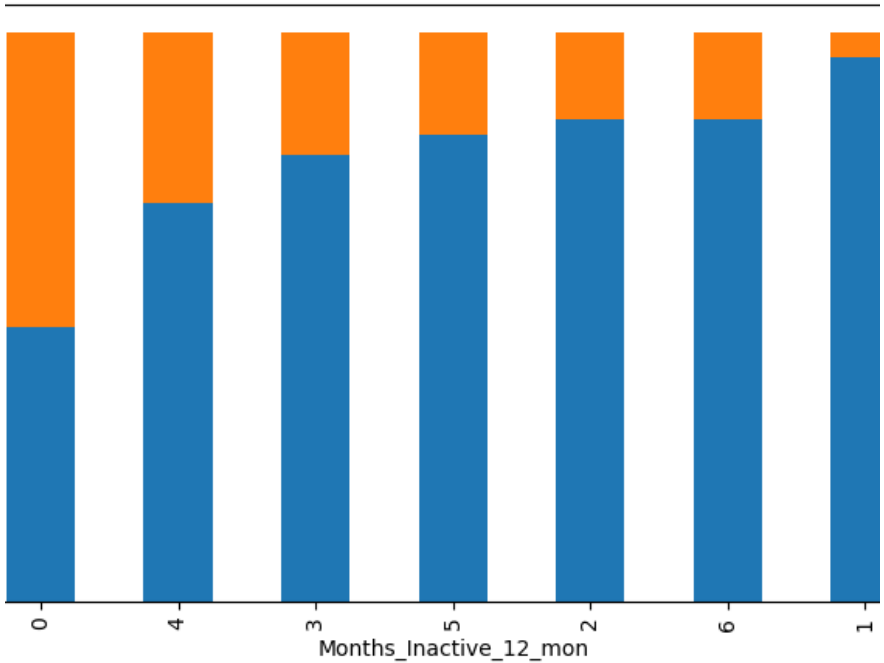
Contact Counts vs. Attrition

- As customers have less contact with the bank the attrition rate increases rapidly.
- At 6 months of no contact the attrition rate becomes 100%.



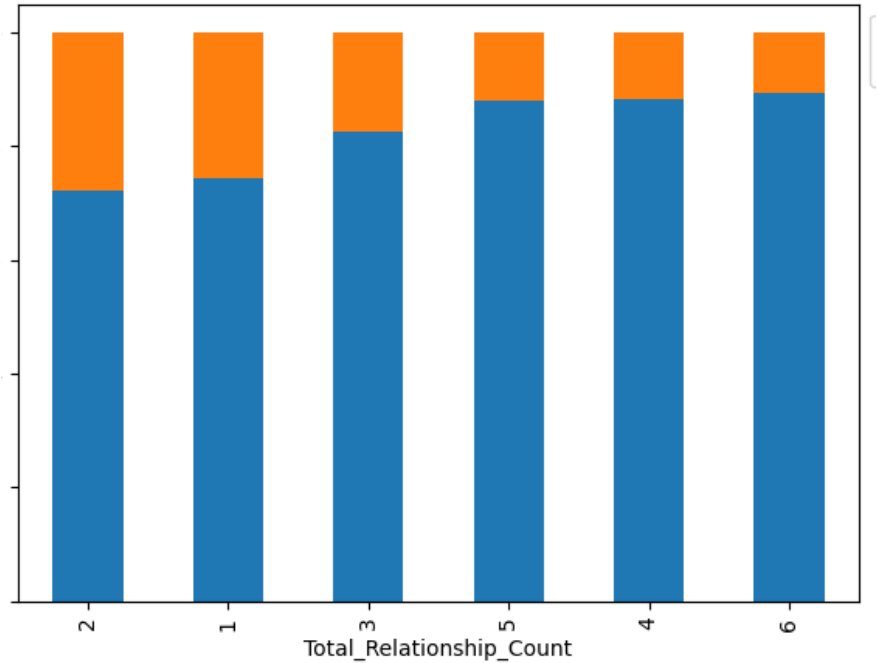
Months Inactive vs. Attrition

- Attrition happens at a high rate when customers are not active in the first month.
- Attrition is lower after the first month on inactivity.
- It rises again in months 3 and 4.
- Attrition is lower in months 5 and 6.



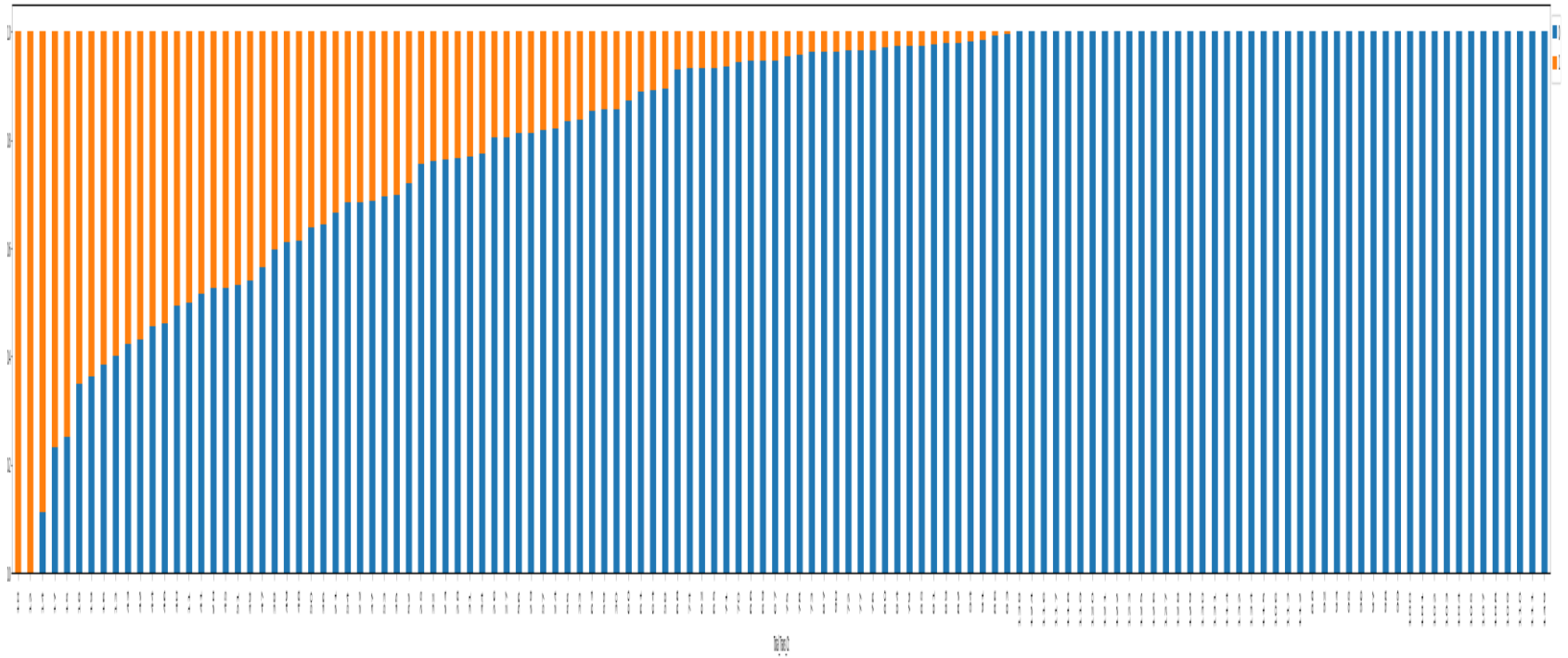
Total Relation Ship Count vs. Attrition

- The more relationships the bank has with customers the less likely they will leave the bank.



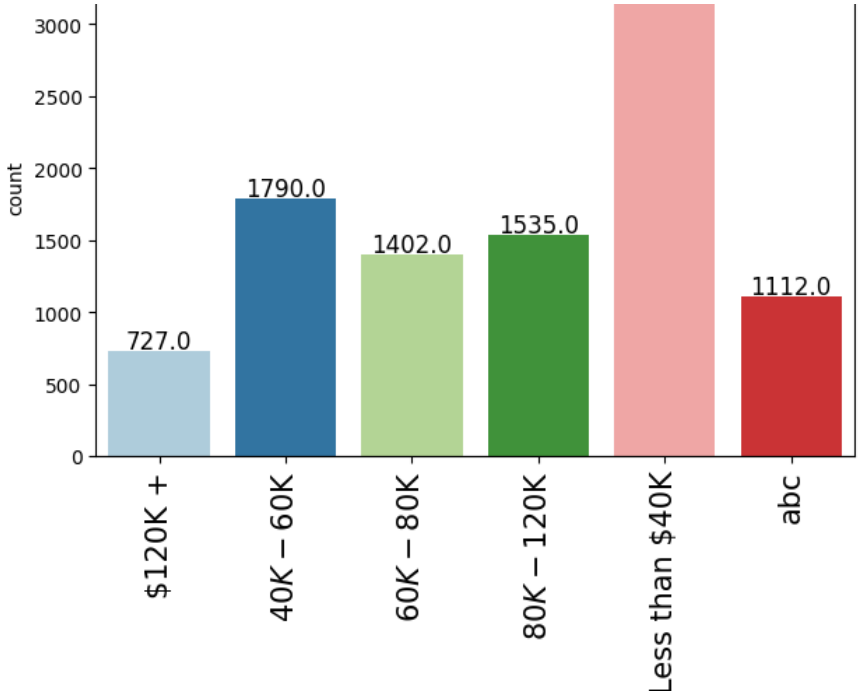
Total Transaction Count vs. Attrition

- As the number of transactions fall the attrition rate increases.



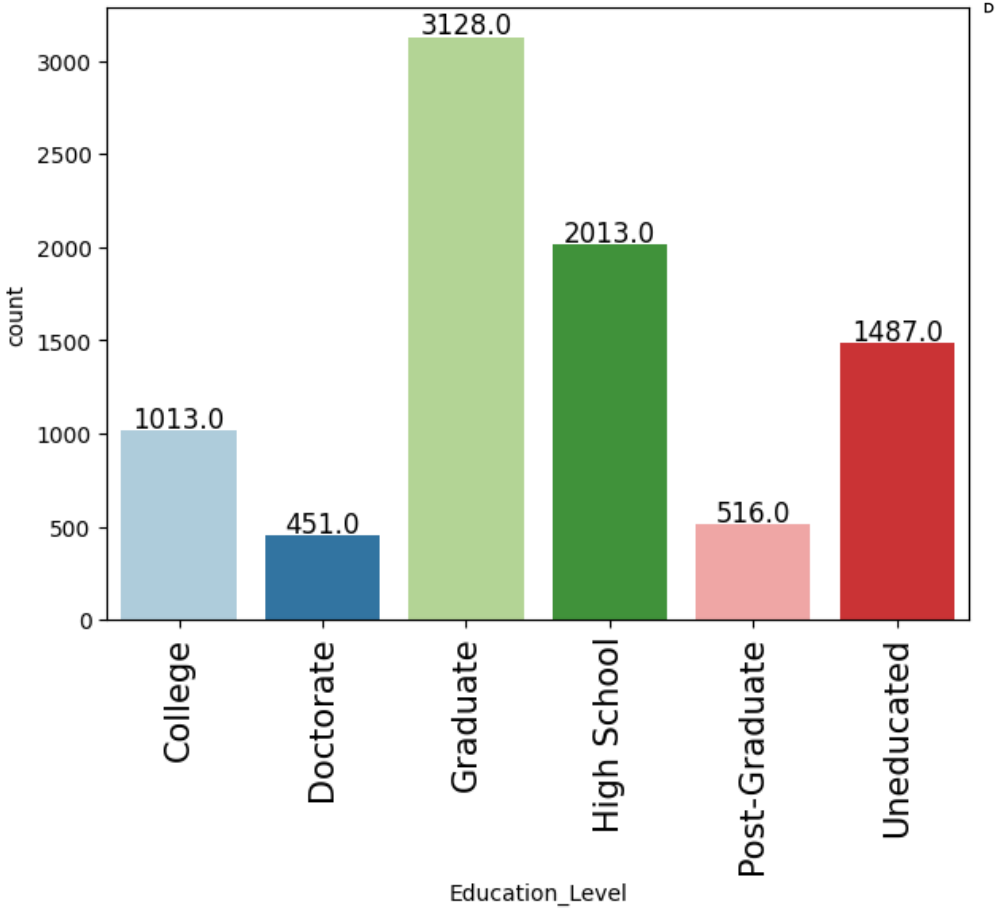
Income Category

- A third of customers are in the less than $40K column
- Income is evenly distributed among the other income levels.



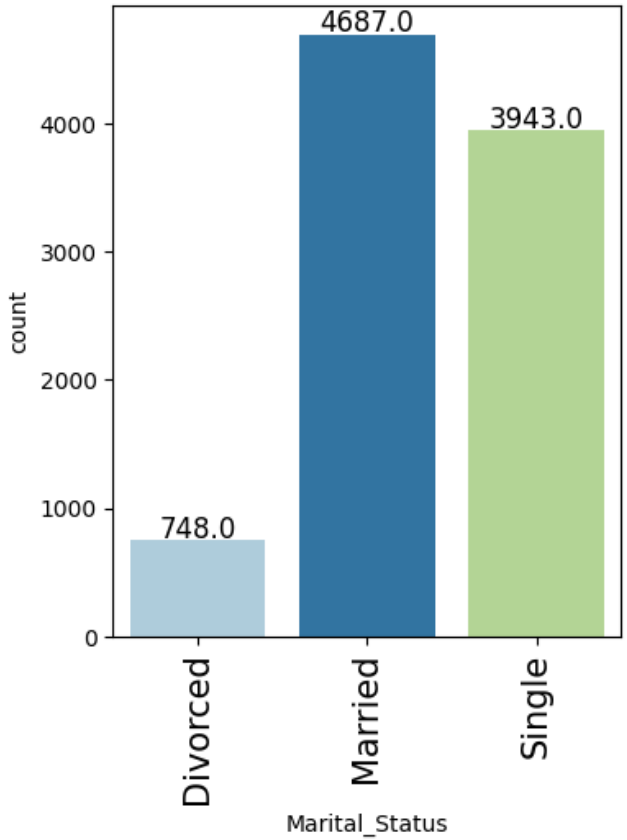
Education Level

- The majority of customers are Graduate Level.
- Education Level had missing values that were imputed with the most frequent strategy for modeling.



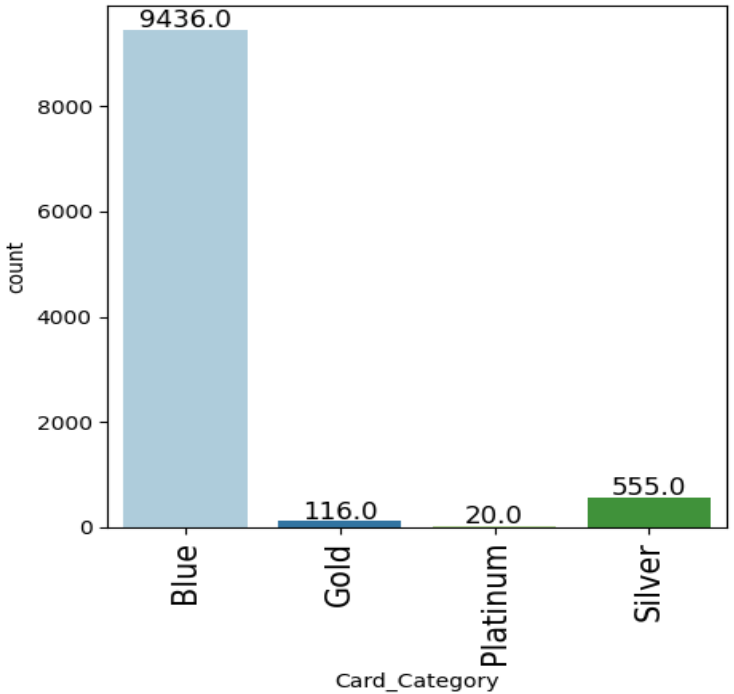
Marital Status

- Marital Status was mostly split between Married and Single with some customers Divorced.
- Marital Status had missing values that were imputed using most frequent strategy for modeling.



Card Category

- A large majority of card holders have Blue Cards.
- Very few customers have Silver, Gold and Platinum rewards.



Great Learning

Happy Learning !