Project 6: CreditCard Users Churn Prediction

BILLY HEIDEL

Introduction

We have obtained 21 variables from 10,127 customers.

Objectives:

- Explore and visualize the dataset.
- Build a classification model to predict if the customer is going to churn or not.
- Optimize the model using appropriate techniques.
- Generate a set of insights and recommendations that will help the bank.

Variables Analyzed

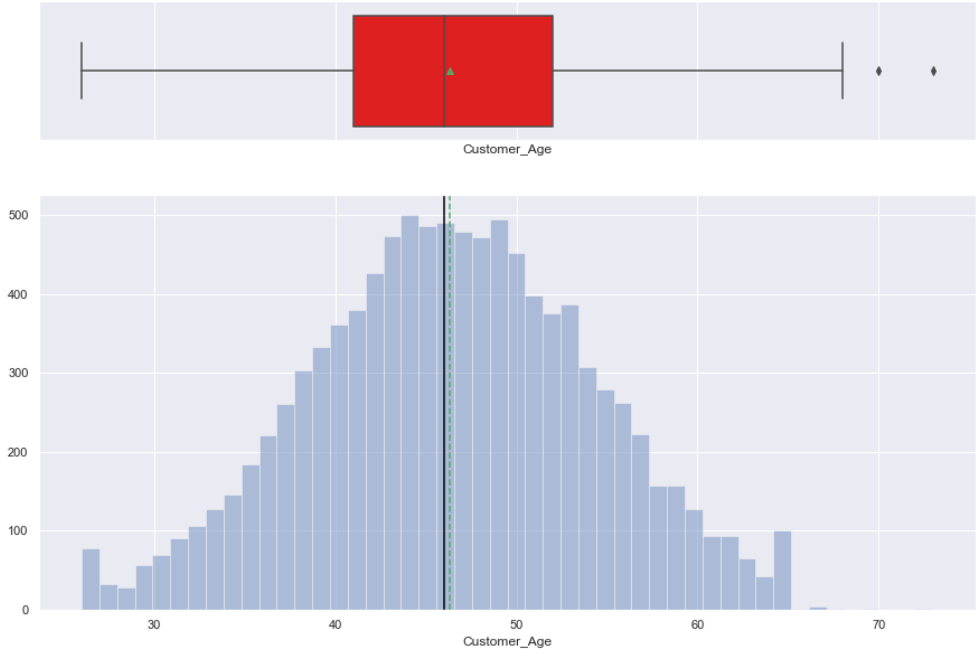
- **CLIENTNUM:** Client number. Unique identifier for the customer holding the account.
- **Attrition_Flag:** Internal event (customer activity) variable - if the account is closed then 1 else 0.
- **Customer_Age:** Age in Years.
- **Gender:** Gender of the account holder.
- **Dependent_count:** Number of dependents.
- **Education_Level:** Educational Qualification of the account holder.
- **Marital_Status:** Marital Status of the account holder.
- **Income_Category:** Annual Income Category of the account holder.
- **Card_Category:** Type of Card.
- **Months_on_book:** Period of relationship with the bank.
- **Total_Relationship_Count:** Total no. of products held by the customer.

Variables Analyzed

- **Months_Inactive_12_mon:** No. of months inactive in the last 12 months.
- **Contacts_Count_12_mon:** No. of Contacts in the last 12 months.
- **Credit_Limit:** Credit Limit on the Credit Card.
- **Total_Revolving_Bal:** Total Revolving Balance on the Credit Card.
- **Avg_Open_To_Buy:** Open to Buy Credit Line (Average of last 12 months).
- **Total_Amt_Chng_Q4_Q1:** Change in Transaction Amount (Q4 over Q1).
- **Total_Trans_Amt:** Total Transaction Amount (Last 12 months).
- **Total_Trans_Ct:** Total Transaction Count (Last 12 months).
- **Total_Ct_Chng_Q4_Q1:** Change in Transaction Count (Q4 over Q1).
- **Avg_Utilization_Ratio:** Average Card Utilization Ratio.

Exploratory Data Analysis (EDA)

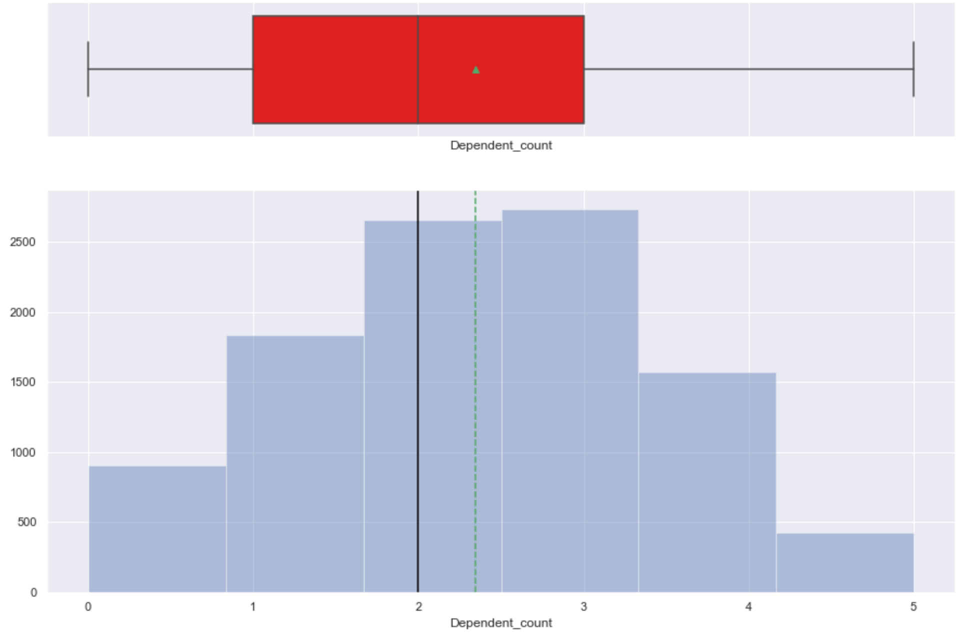
Customer_Age



Observations:

- Ages range from 26-73 with a median of 46 years.
- The distribution of age is very close to normal, except for small peaks at 26 and 65.
- There are two outliers: 70 and 73.

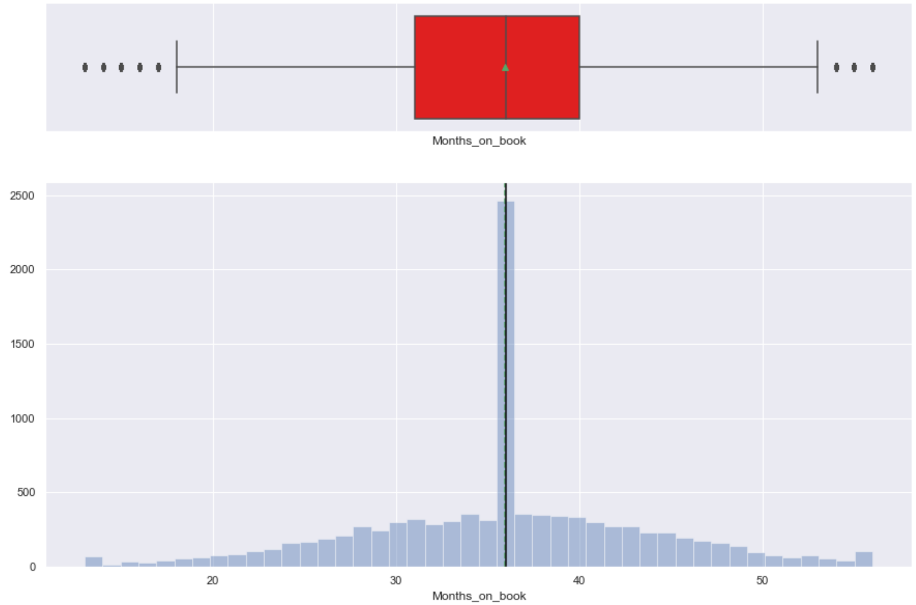
Dependent_count



Observations:

- Number of dependents ranges from 0 to 5 with a median of 2 and mean 2.35.
- Distribution looks fairly close to a normal around a mean of 2.5.
  - However, there is a right skew to the data despite a mode of 3.
- There are no outliers.

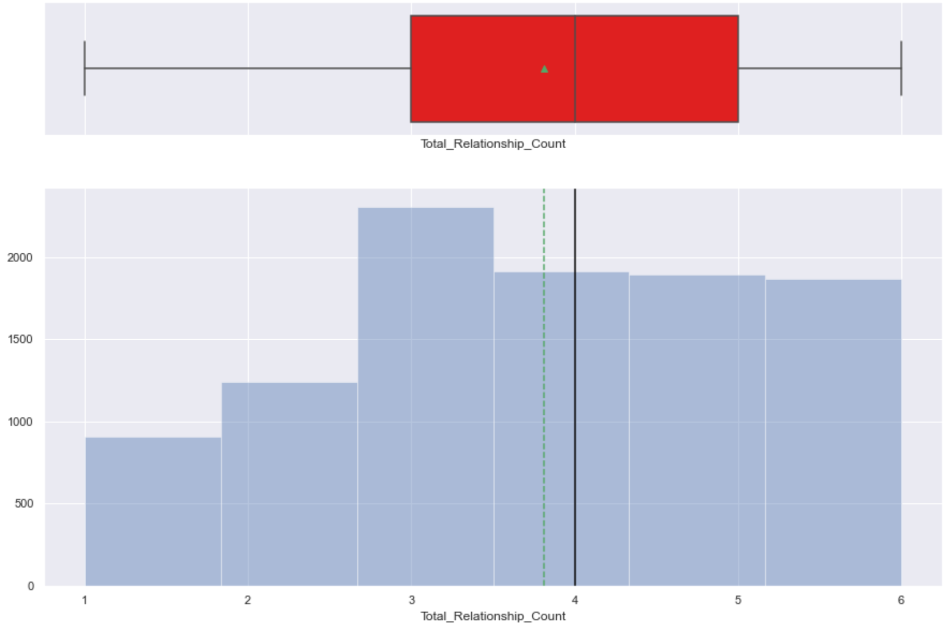
Months_on_book



Observations:

- Ranges from 13 to 56 months with mean and median of 36 months.
- Distribution looks somewhat normal, except for the mode of 36 which is much more frequent than any other number of months on book.
    - Did something happen that caused a lot of customers to join the bank 36 months ago? This is worth investigating.
- There are small peaks at 13 and 56 months.
- There are a few outliers on both sides.

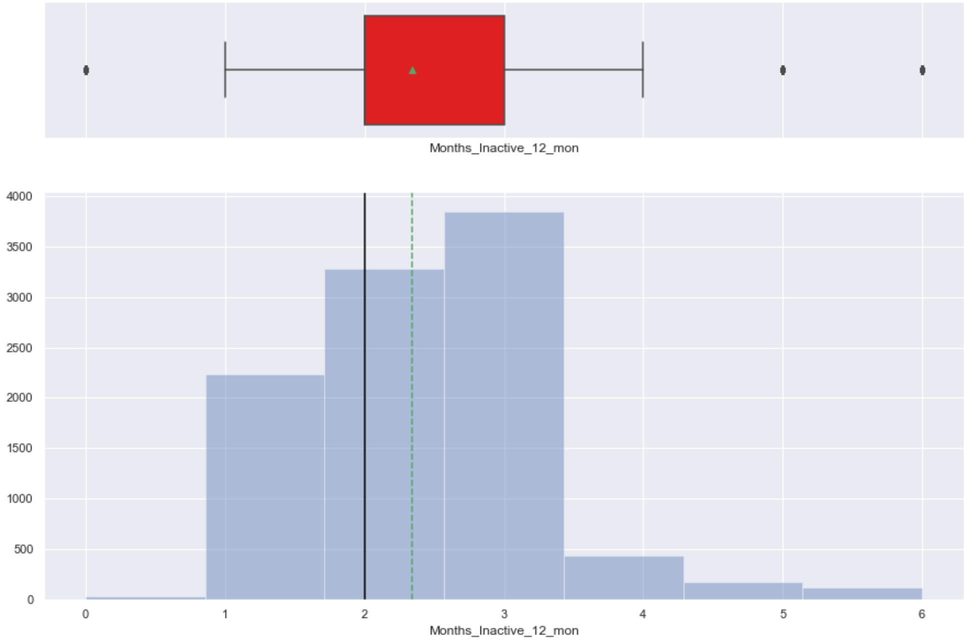
Total_Relationship_Count



Observations:

- Ranges from 1 to 6 products with a median of 4 products and mean of 3.81.
- More customers buy a larger number of products until the mode of 3, then the number of customers with 4 products is less, and the number of customers with 5 or 6 products is almost equal to those with 4 products.
- There are no outliers.

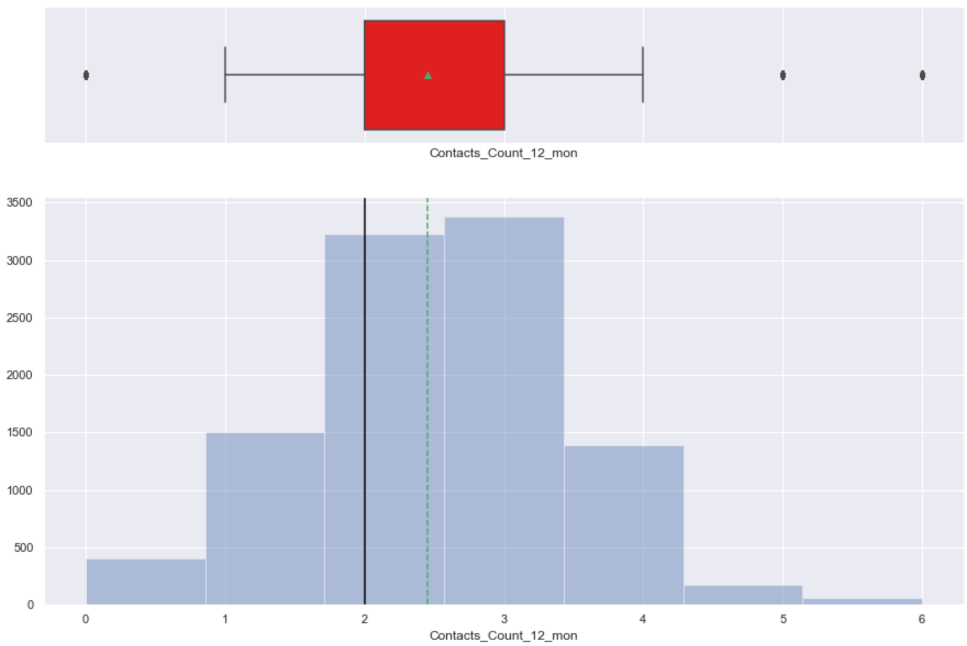
Months_Inactive_12_mon



Observations:

- Ranges from 0 to 6 months inactive in the last year with a median of 2 months.
- Data is right skewed.
- The vast majority of customers were inactive for 1-3 months in the past year.
- Almost no customers were active for all 12 months in the past year.
- There are some outliers present.

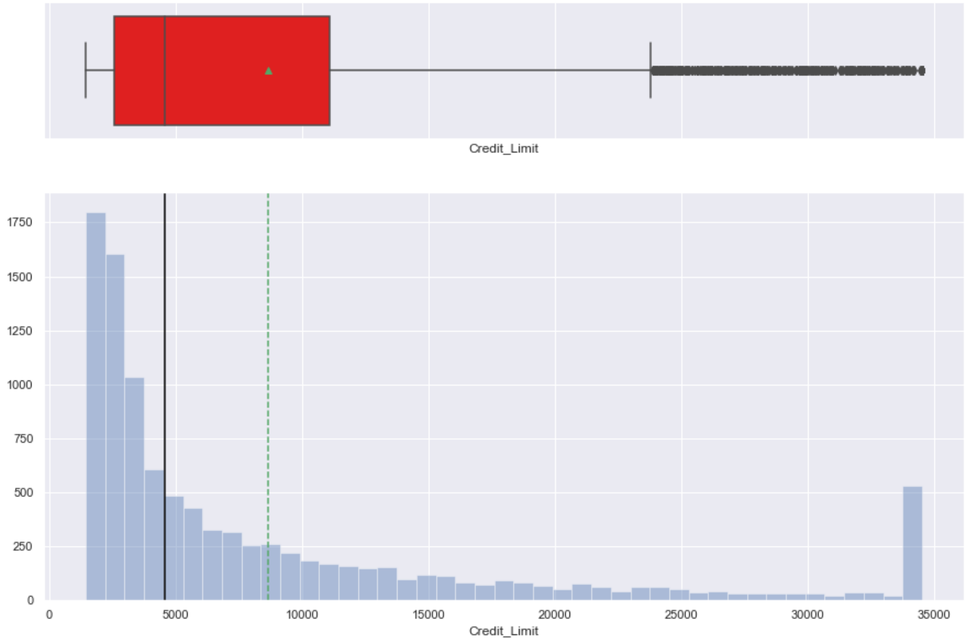
Contacts_Count_12_mon



Observations:

- Ranges from 0 to 6 contacts with a median of 2 and mean 2.34.
- Data is right skewed.
- Most customers had either 2 or 3 contacts in the last 12 months. The amount of customers decreases sharply as the number of contacts gets farther away from 2 and 3.
- Some outliers are present.

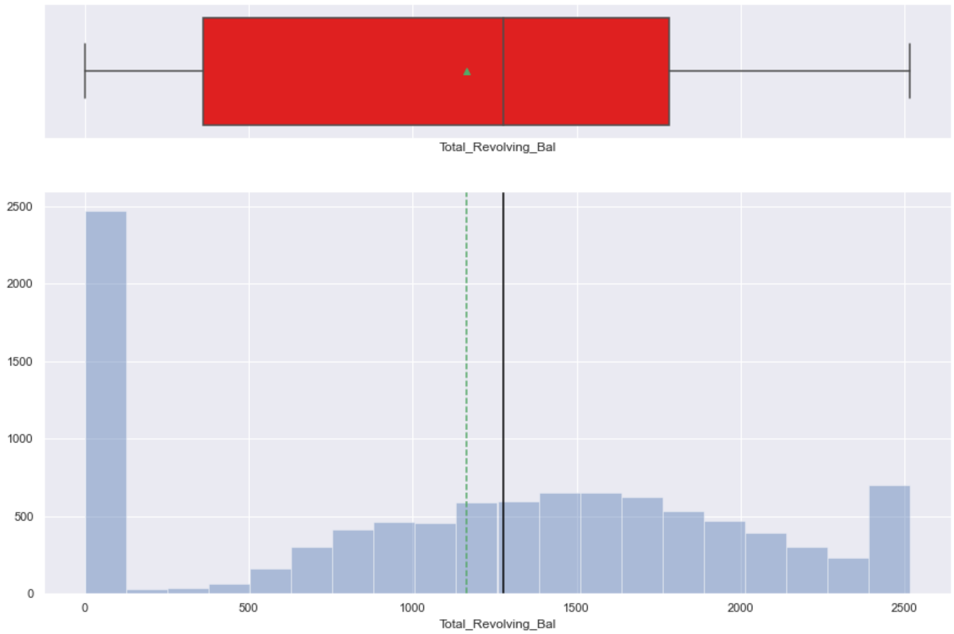
Credit_Limit



Observations:

- Ranges from $1,438.30 to $34,516.00 with a median of $4,549.00 and average of $8,631.95.
- Data is right skewed with many outliers on the far side.
- There is a peak at the maximum value of $35,516.
    - Perhaps this is the maximum credit limit Thera Bank will allow.

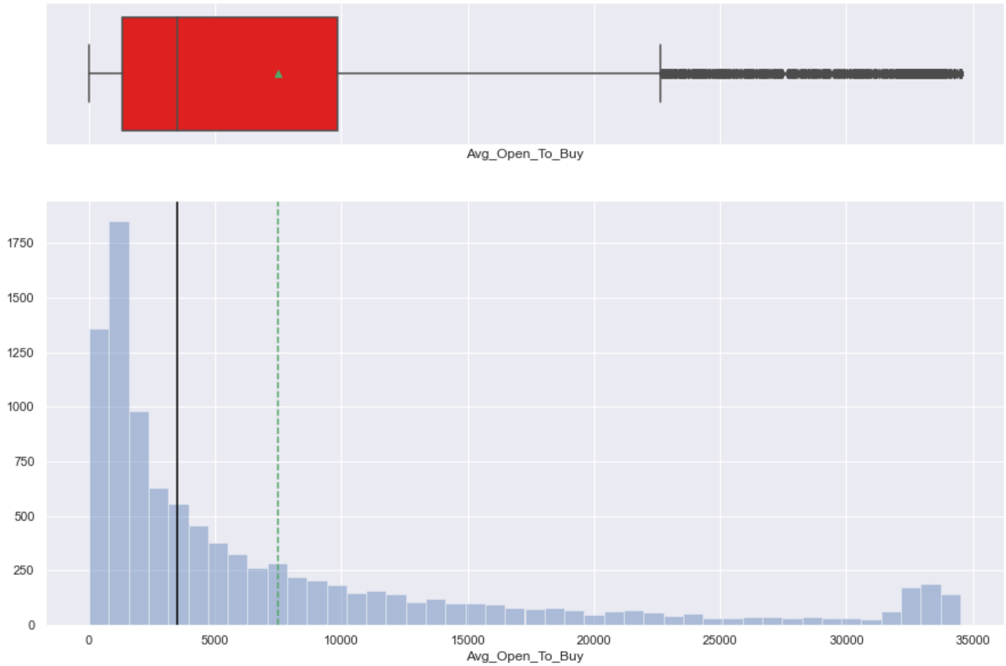
Total_Revolving_Bal



Observations:

- Ranges from $0 to $2,517 with a median of $1,276 and mean of $1,162.81.
- Data is left skewed with no outliers.
- There are peaks at $0, $2,517, and somewhere around $1,500.
  - More customers have 0 revolving balance than any other amount.

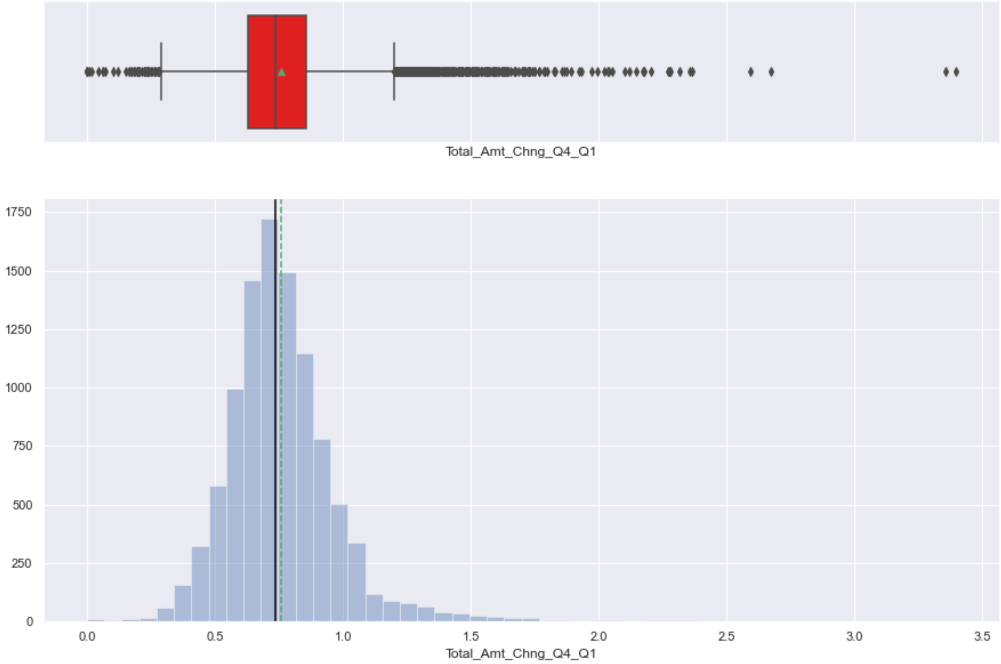
Avg_Open_To_Buy



Observations:

- Ranges from $3 to $34,516 with a median of $3,474 and mean of $7,469.14.
- Very right skewed with many outliers on the far end.
- There is a jump in the frequency of values from about $31,000 to our max value of $34,516.
  - This is likely related to the peak of customers with the max credit limit of $34,516 as seen earlier.
- **This variable is just Credit_Limit - Total_Revolving_Bal. It provides no new information, so I removed it.**

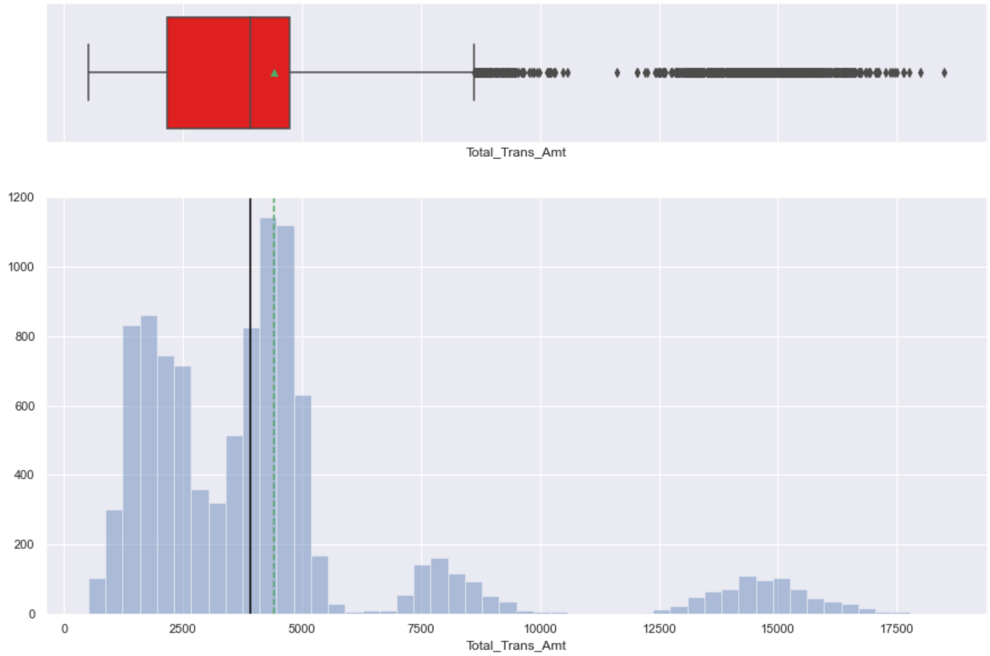
Total_Amt_Chng_Q4_Q1



Observations:

- Ranges from 0 to 3.397 with a median of 0.736 and mean of 0.760.
  - This means that the average transaction amount has decreased from quarter 1 to quarter 4.
- Distribution seems close to normal with outliers on both sides.
  - There are more extreme outliers on the far side.

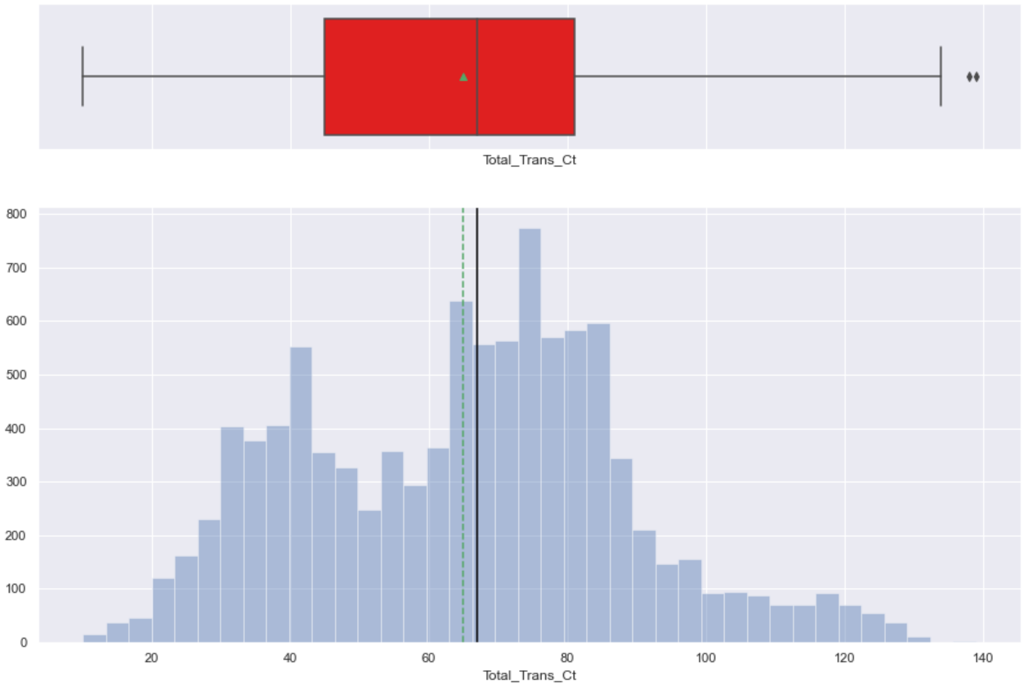
Total_Trans_Amt



Observations:

- Ranges from $510 to $18,484 with a median of $3,899 and mean of $4,404.09.
- There are many extreme outliers on the far end.
- **There appears to be 4 different distributions in this graph. We will investigate this further.**

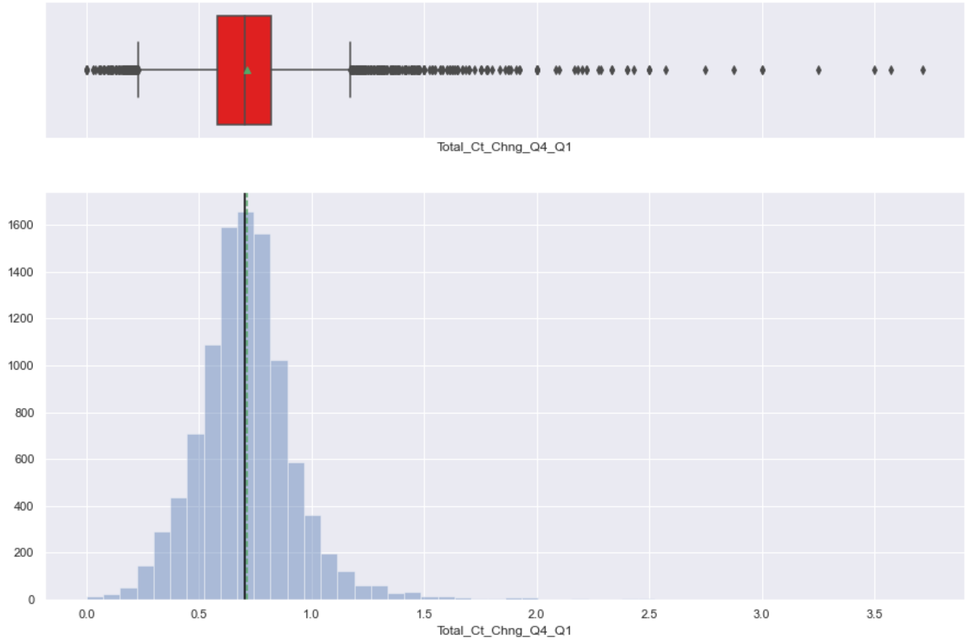
Total_Trans_Ct



Observations:

- Ranges from 10 to 139 with a median of 67 and mean of 64.89.
- There is only a small amount of outliers.
- There seems to be 2-3 distributions in this plot. This will be investigated further.
  - This is likely because total transaction count and total transaction amount are related.

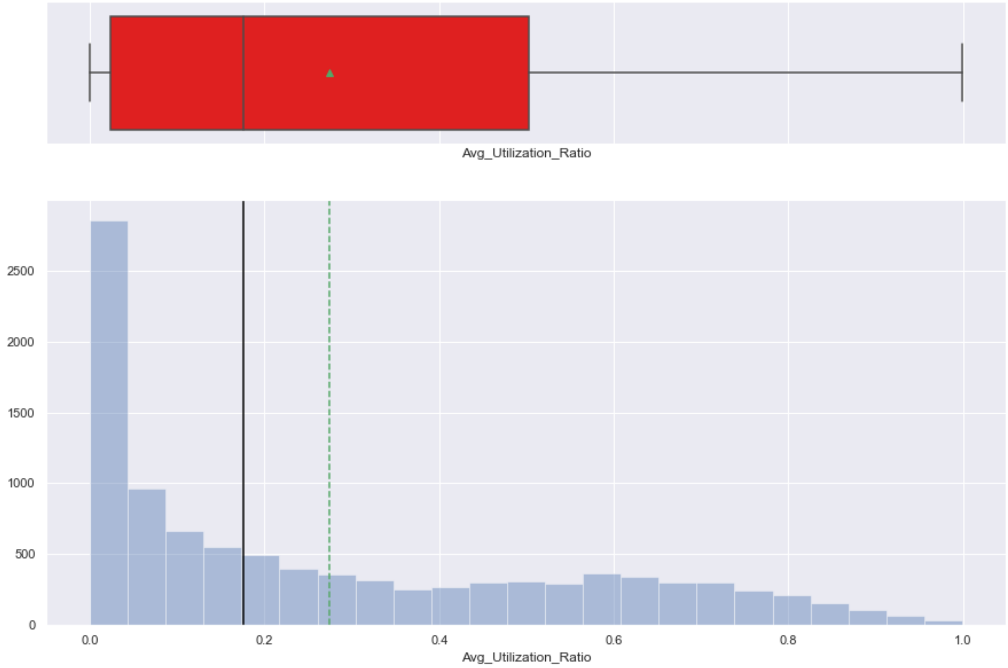
Total_Ct_Chng_Q4_Q1



Observations:

- Ranges from 0 to 3.714 with median of 0.702 and mean of 0.712.
  - This means that the average number of transactions has decreased from quarter 1 to quarter 4.
- Distribution seems close to normal with outliers on both sides.
  - There are more extreme outliers on the far side.

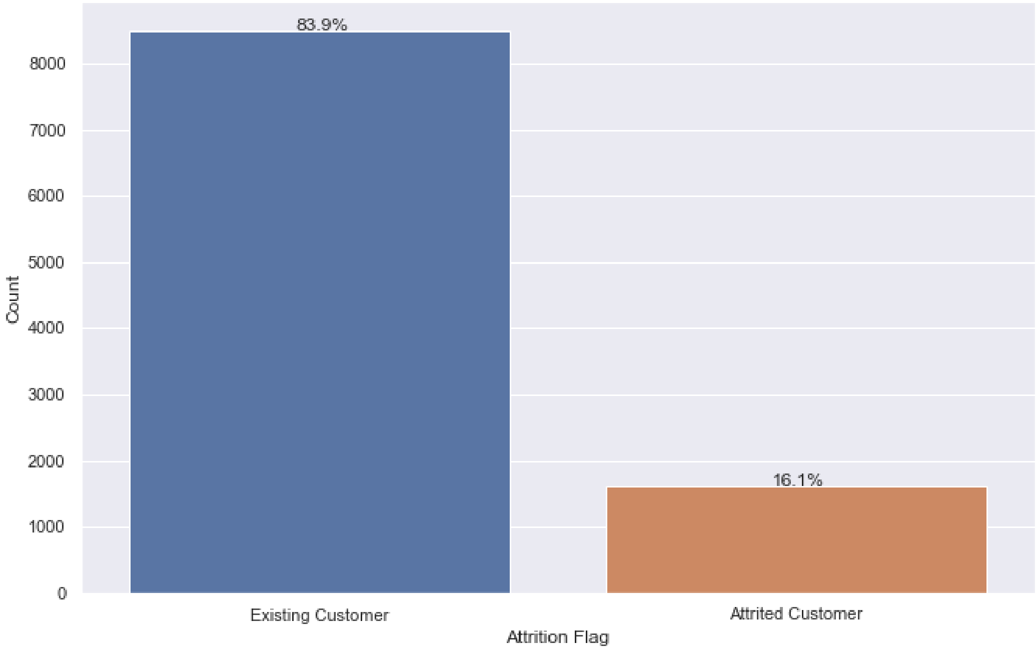
Avg_Utilization_Ratio



Observations:

- Ranges from 0 to 0.999 with a median of 0.176 and mean of 0.275.
- Data is right skewed with no outliers.
- Possibly a combination of 2 distributions because of the peak around 0.6.
- **This variable is just Total_Revolving_Bal divided by Credit_Limit. It provides no new information, so it will be removed.**

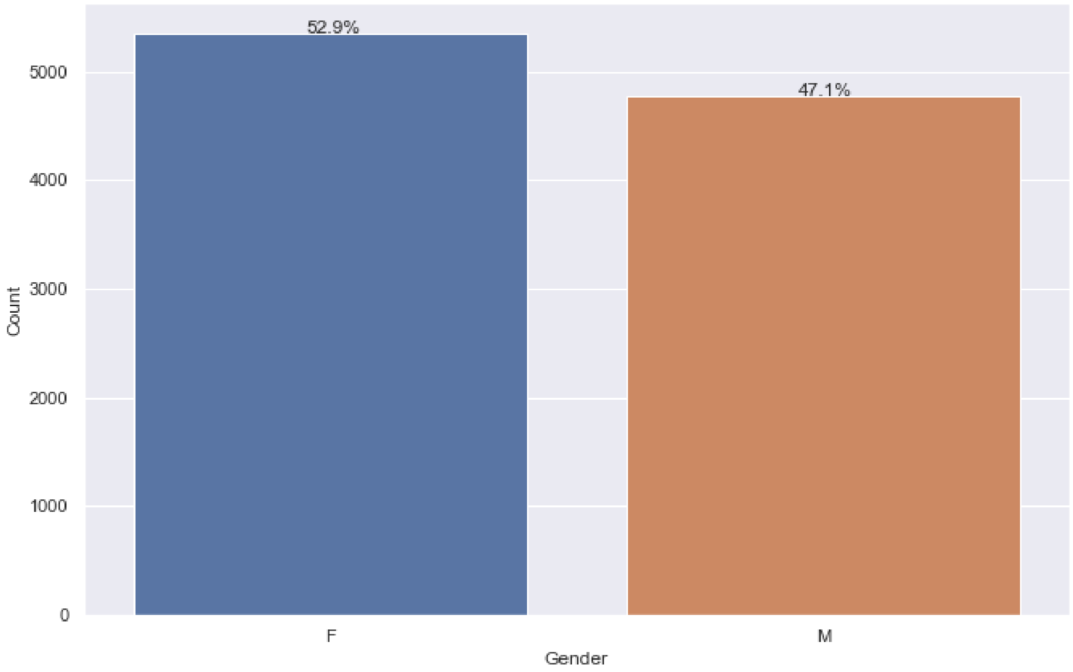
Attrition_Flag



Observation:

- 16.1% of customers have attritted.

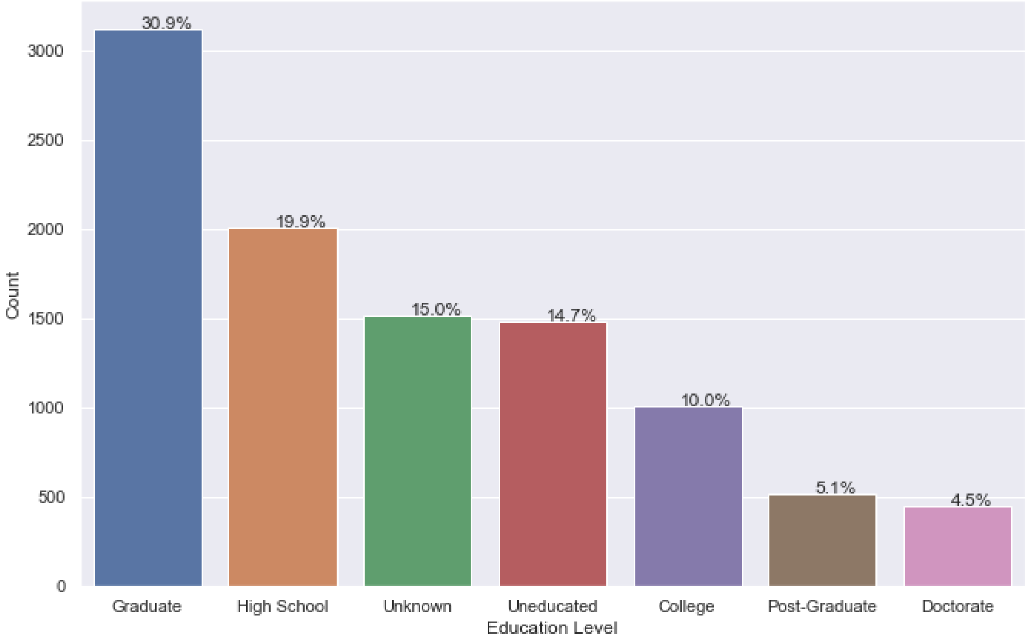
Gender



Observation:

- Slightly more customers are female (52.9%) than male (47.1%).

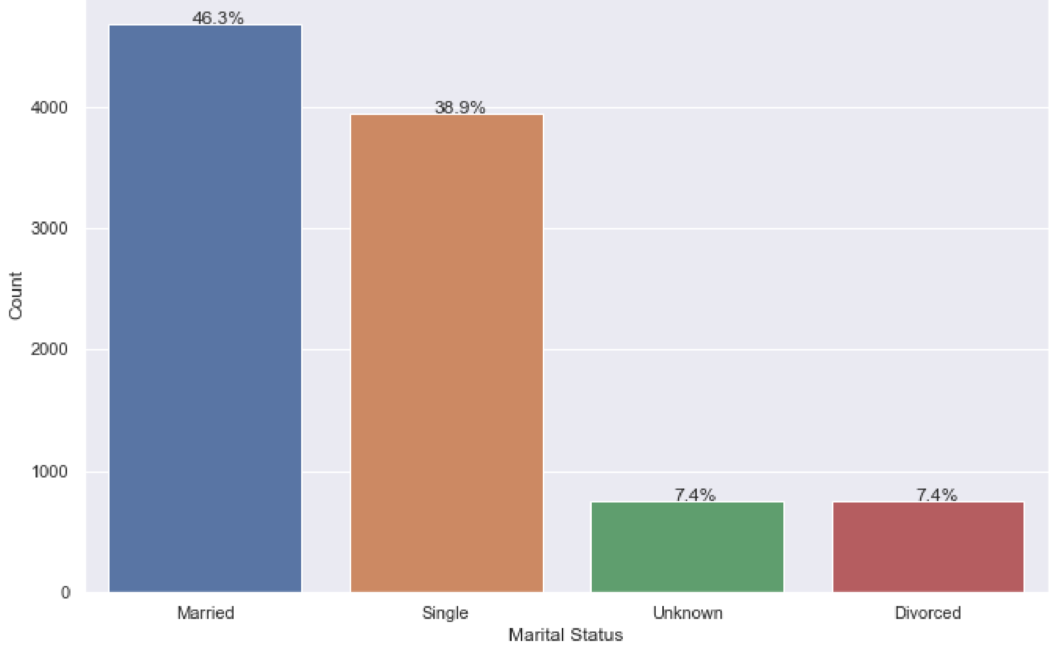
Education_Level



Observations:

- Education level is not uniformly distributed.
- Many more customers are Graduates than any other group.
- After this, it is generally less common for customers to be more educated.
- Fifteen percent of customers have unknown education level.

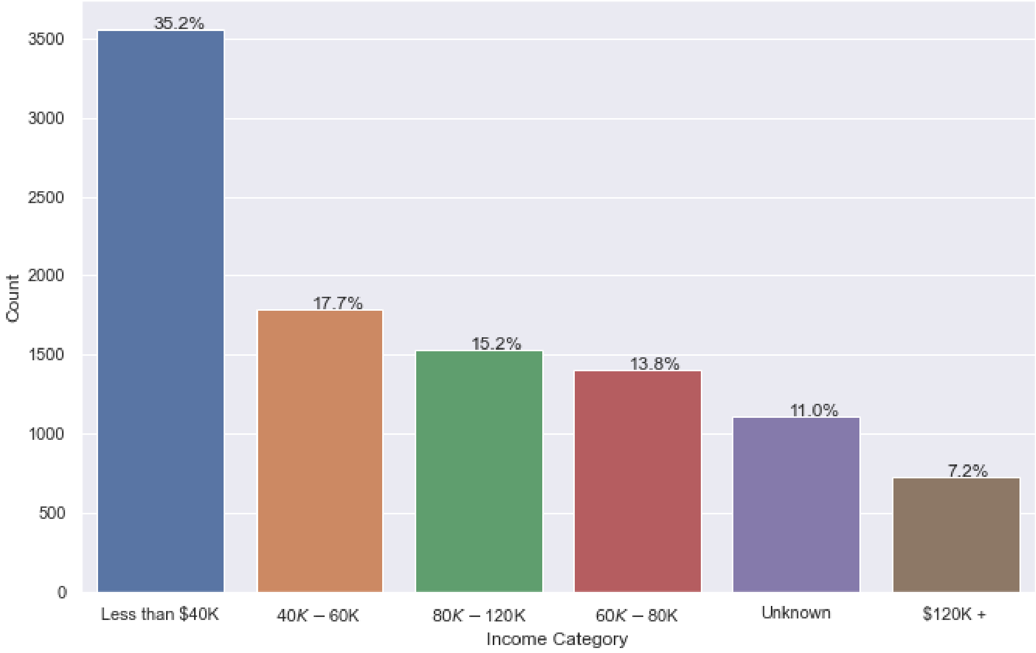
Marital_Status



Observations:

- Almost half (46.3%) of the customers are married.
- Over a third (38.9%) of customers are single.
- 7.4% of customers are divorced.
- 7.4% of customers have unknown marital status.

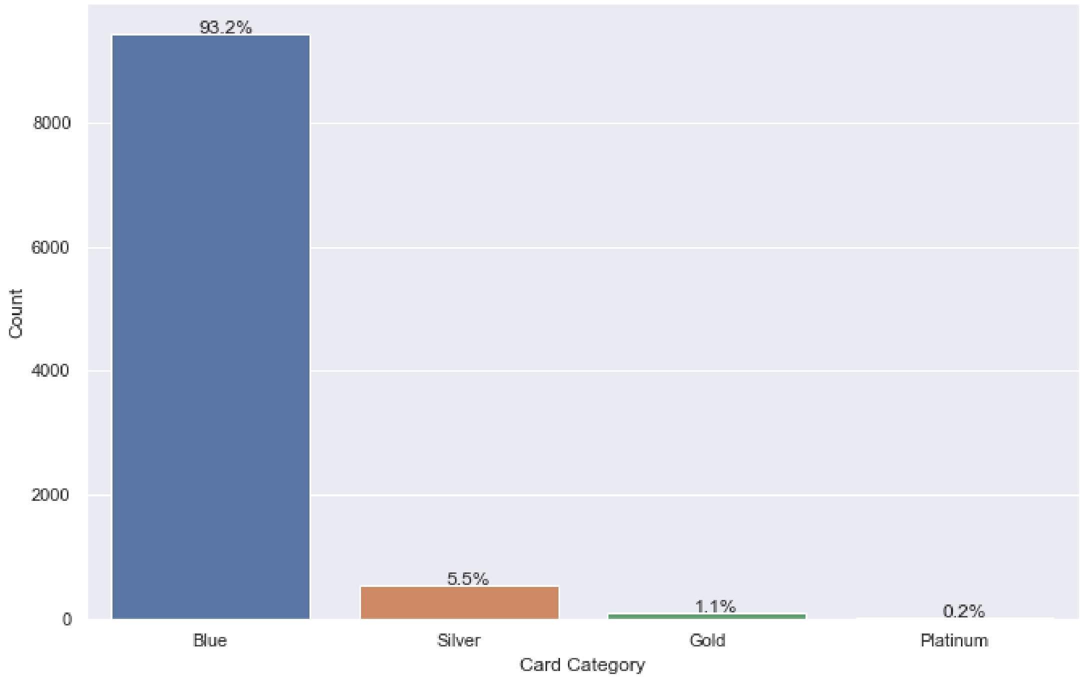
Income_Category



Observations:

- Customers are less likely to have a higher income.
- Over 1/3 (35.2%) of customers make less than 40 thousand dollars per year.
- 11.0% of customers have unknown income.

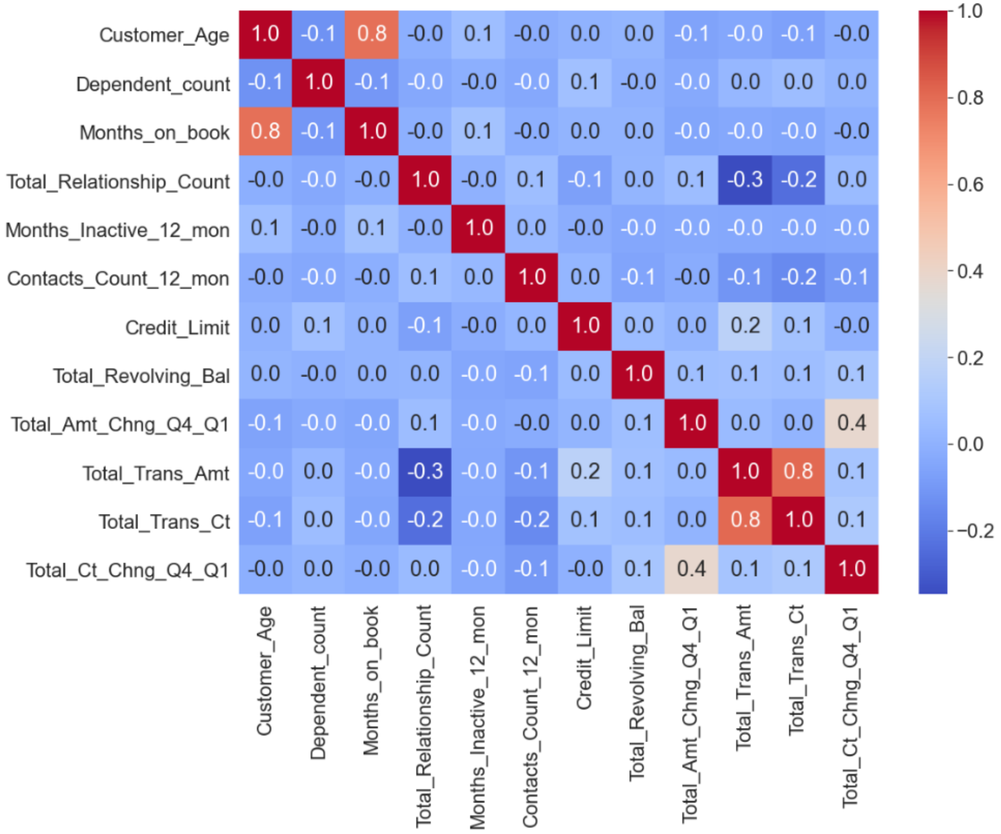
Card_Category



Observations:

- The higher the category of card, the less likely the customer is to possess that card. Almost all customers (93.2%) have a blue card.
- Only 0.2% of customers have the platinum card.

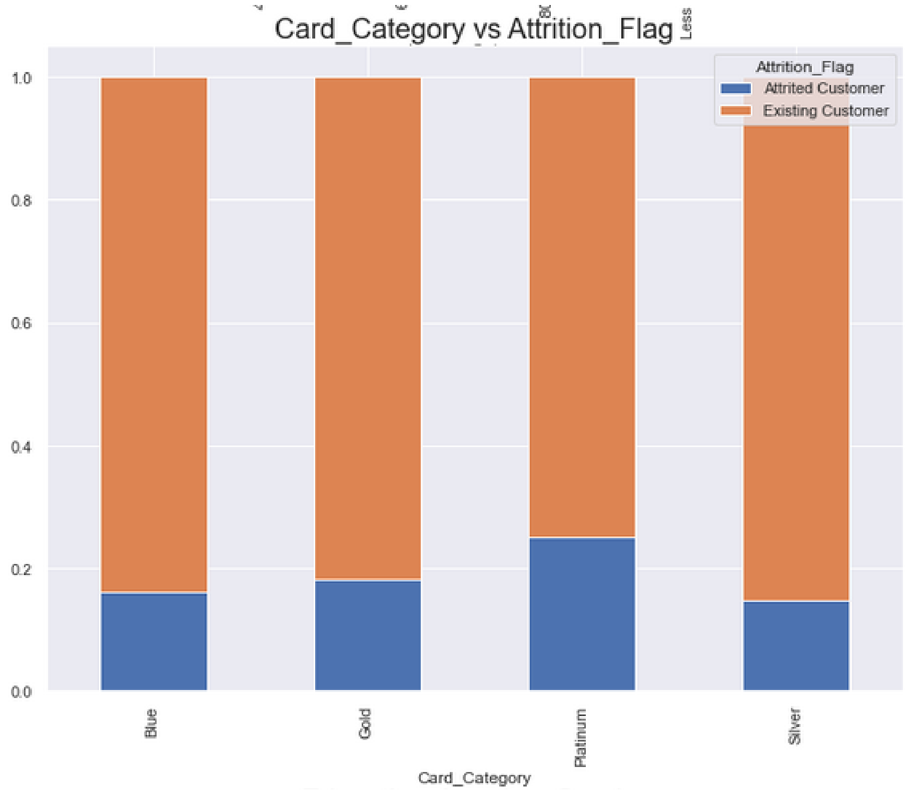
Relationships Between Numerical Variables



Observations:

- Customer_Age and Months_on_book have an unusually strong positive linear relationship.
  - The relationship is clear in the pairplot. Only where months_on_book = 36 is the relationship thrown off.
  - This must be investigated further.
- There is a very strong positive linear relationship between Total_Trans_Amt and Total_Trans_Ct.
- There is a weak (but existent) positive linear relationship between Total_Amt_Chng_Q4_Q1 and Total_Ct_Chng_Q4_Q1.
- **There are 3-4 mixed Gaussians in the Total_Trans_Amt variable and 2-3 mixed Gaussians in the Total_Trans_Ct variable.**
  - Pairplots for these variables have the data clearly grouped into different sections.
  - **The pairplot with the target variable (Attrition_Flag) as hue shows that these Gaussians are related to customer attrition.**

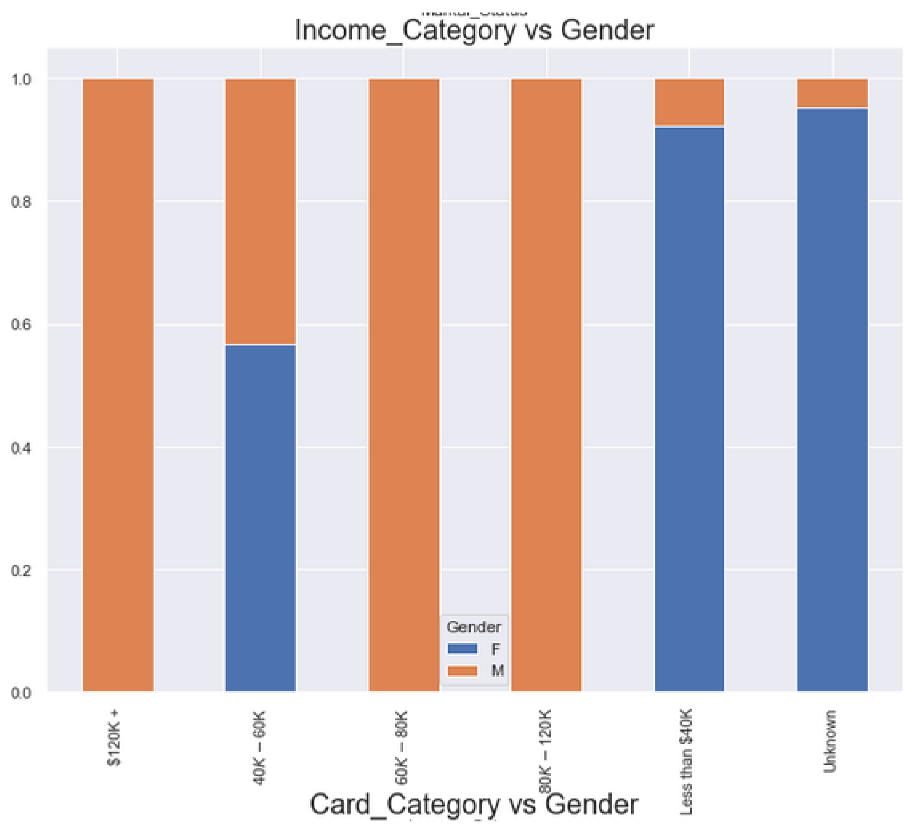
Relationships Between Categorical Variables



Observations:

- Customers with a higher card category have a higher likelihood of attrition.
  - The exception to this is that customers with the blue card are very slightly more likely to attrit than those with the silver card.
- Attrition_Flag's relationships with the other categorical variables are weaker, but still existent.
  - A slightly higher proportion of females attritted than males.
  - Customers with more education have a slightly higher attrition rate.
  - Single customers were slightly more likely to attrit than divorced or married customers.
  - Customers with income from 60K-80K were the least likely to attrit. As income increases or decreases from this amount, the likelihood of attrition increases.

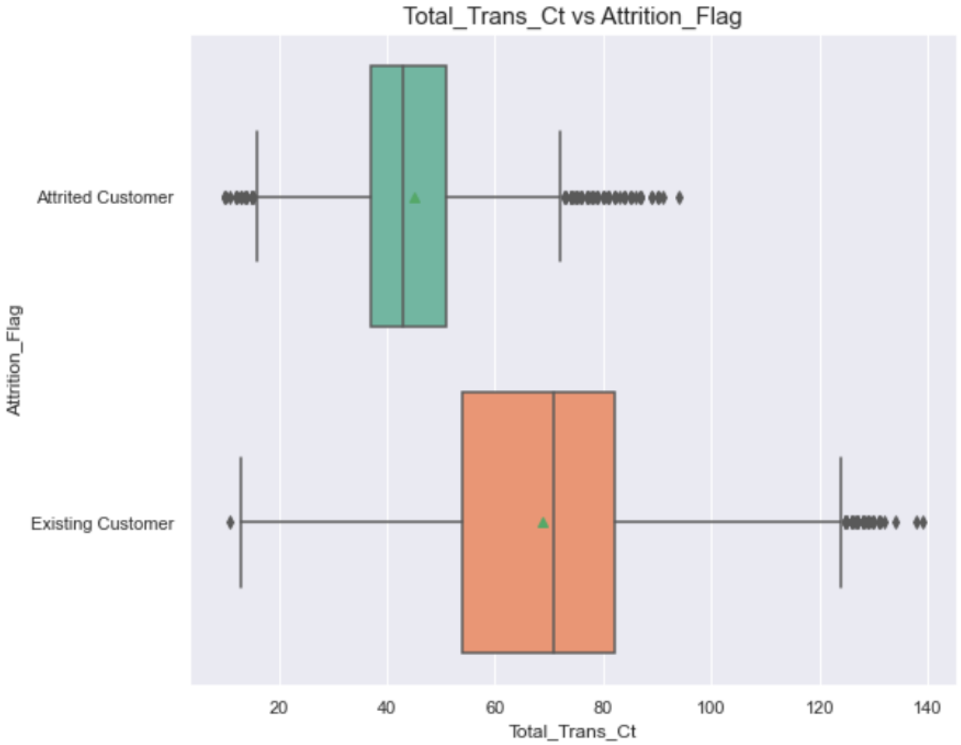
Relationships Between Categorical Variables



Observations:

- No females make more than 60K in a year, and very few men make less than 40K.
  - Exceptions to this could be in the unknown income, which is dominated by females.
- Female are less likely to own a card of a higher category, except for the Platinum card.
- No customers with the platinum card have the College level of education.
  - Also, a larger proportion of customers with the platinum card have a higher education level than college or are uneducated.
- Single customers are more likely to have a higher level card, whereas married and divorced customers are more likely to have a lower level card.
- Customers with higher income are more likely to own a higher level card.
  - Those who own the platinum card have a higher proportion of unknown income than any other card category.

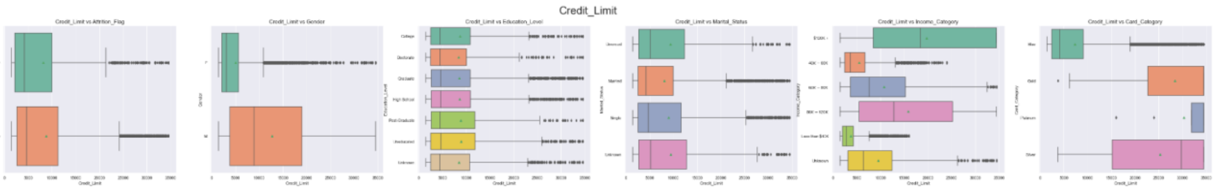
Relationships Between Numerical and Categorical Variables



Observations:

- Customers who attritted... hold a smaller total number of products than existing customers.
  - have a larger number of inactive months on average.
  - had a larger average number of contacts in the last year.
  - have a slightly smaller average credit limit.
  - have a much lower average revolving balance.
  - have a much lower average total transaction amount and count.
  - have a lower average change in transaction amount and count from Q1 to Q4.

Relationships Between Numerical and Categorical Variables



Observations:

- Those with a higher income have more dependents on average.
- Customers with a lower card category on average hold a higher number of products.
- Customers with a higher card category have a higher average total transaction amount and count.
- Credit_Limit is related to all categorical variables.
  - Males have a much higher average credit limit than females.
  - Those with more education have a slightly higher average credit limit.
  - The exception to this is those who are uneducated. They have the highest average credit limit of all.
  - Married customers have the lowest average credit limit, while those divorced or of unknown marital status have the highest average credit limits.
  - Those with more income have a much higher average credit limit.
  - Those with a higher level card category have a higher average credit limit.
  - The average credit limit for those with the clue card are much lower than that for those with other card types.

Insights from EDA

1. The relationships between Attrition_Flag and other variables are as follows:

- Customers who attritted...
  - hold a smaller total number of products than existing customers.
  - have a larger number of inactive months on average.
  - had a larger average number of contacts in the last year.
  - have a slightly smaller average credit limit.
  - have a much lower average revolving balance.
  - have a much lower average total transaction amount and count.
  - have a lower average change in transaction amount and count from Q1 to Q4.
- Customers with a higher card category have a higher likelihood of attrition.
  - The exception to this is that customers with the blue card are very slightly more likely to attrit than those with the silver card.
- Attrition_Flag's relationships with the other categorical variables are weaker, but still existent.
  - A slightly higher proportion of females attritted than males.
  - Customers with more education have a slightly higher attrition rate.
  - Single customers were slightly more likely to attrit than divorced or married customers.
  - Customers with income from 60K-80K were the least likely to attrit. As income increases or decreases from this amount, the likelihood of attrition increases.

Insights from EDA

2. I would like to know more about how this dataset was handled.
   - Small bumps at the extreme values of some numerical variables such as Customer_Age and Total_Revolving_Bal suggest that outliers were capped at these values. Is this in fact the case?
   - Credit_Limit has a bump as well, but on only the far end. I would like to confirm that this is because Thera Bank does not provide credit greater than 34,516.
3. There is a strikingly large number of customers who have had a relationship for the bank for exactly 36 months.
   - Is this because of some sort of outreach, imputation of missing values as 36, or some other reason?
4. There are multiple distributions in the Total_Trans_Amt and Total_Trans_Ct variable histograms.
   - Bivariate analysis was not able to adequately explain this. Perhaps there is a missing variable from the dataset that could help separate the data into these different distributions.
5. Credit_Limit was related to all numerical variables. Are these variables used to determine credit limit?
   - If so, this variable might not be useful in model-building.
6. There is an unusually strong linear relationship between Customer_Age and Months_on_book.
   - A relationship between the two makes sense, but the pairplot between them is almost perfectly usual. This is worth investigating in further detail.
7. The average total amount and count of transactions have decreased over the past year, even among customers that have not attritted.
   - This is a concerning trend, and something Thera Bank must focus on changing.

Building and Tuning Models

Data Pre-Processing

Transforming the Data

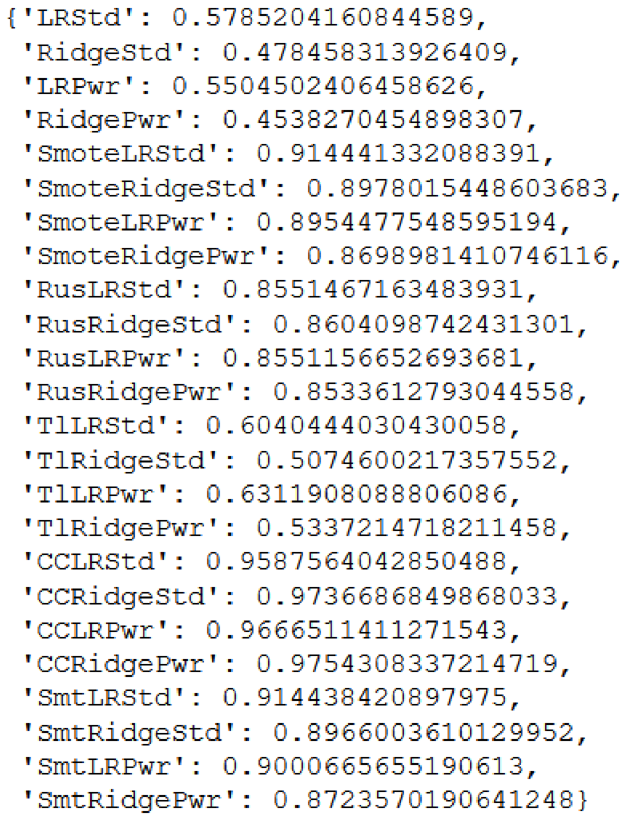
Logistic models will be fit to data transformed by both the Standard Scaler and Power Transformation, and we will compare the results. These transformations will help with outliers as well, which all of our models except logistic regression are robust to.

Which metric do we optimize?

In this case, we want to minimize false negatives. Predicting that a customer will not leave the bank when they actually will leave is more costly than predicting a customer will leave when they actually will stay. Therefore, the performance metric that we will optimize is recall.

Logistic Regression Models

Results of Logistic Regressions Models with Cross-Validation

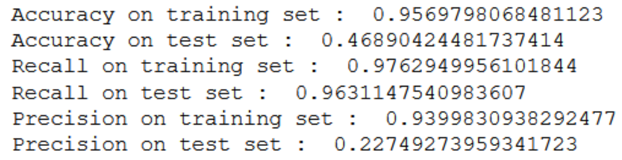


```
{'LRStd': 0.5785204160844589,
 'RidgeStd': 0.478458313926409,
 'LRPwr': 0.5504502406458626,
 'RidgePwr': 0.4538270454898307,
 'SmoteLRStd': 0.914441332088391,
 'SmoteRidgeStd': 0.8978015448603683,
 'SmoteLRPwr': 0.8954477548595194,
 'SmoteRidgePwr': 0.8698981410746116,
 'RusLRStd': 0.8551467163483931,
 'RusRidgeStd': 0.8604098742431301,
 'RusLRPwr': 0.8551156652693681,
 'RusRidgePwr': 0.8533612793044558,
 'TlLRStd': 0.6040444030430058,
 'TlRidgeStd': 0.5074600217357552,
 'TlLRPwr': 0.6311908088806086,
 'TlRidgePwr': 0.5337214718211458,
 'CCLRStd': 0.9587564042850488,
 'CCRidgeStd': 0.9736686849868033,
 'CCLRPwr': 0.9666511411271543,
 'CCRidgePwr': 0.9754308337214719,
 'SmtLRStd': 0.914438420897975,
 'SmtRidgeStd': 0.8966003610129952,
 'SmtLRPwr': 0.9000665655190613,
 'SmtRidgePwr': 0.8723570190641248}
```

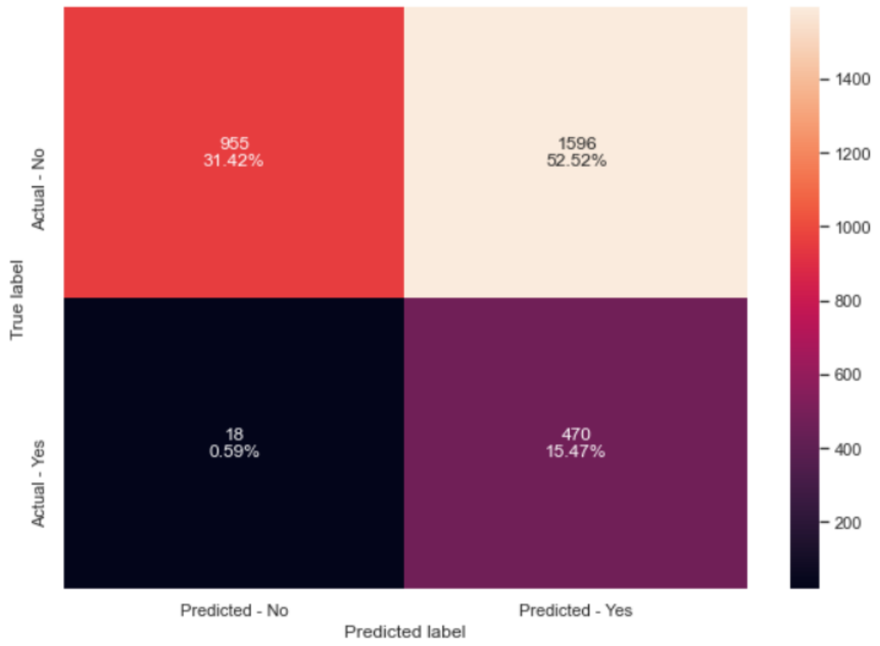
The following techniques were used to build logistic regression models:

- Transformations
  - Standard Scaler
  - Power Transformer
- Oversampling
  - SMOTE
- Undersampling
  - Random Undersampling
  - TomekLinks
  - Cluster Centroids
- SMOTETomek over and undersampling
- Ridge Regularization

Best Logistic Regression Model



```
Accuracy on training set :  0.9569798068481123
Accuracy on test set :  0.46890424481737414
Recall on training set :  0.9762949956101844
Recall on test set :  0.9631147540983607
Precision on training set :  0.9399830938292477
Precision on test set :  0.22749273959341723
```

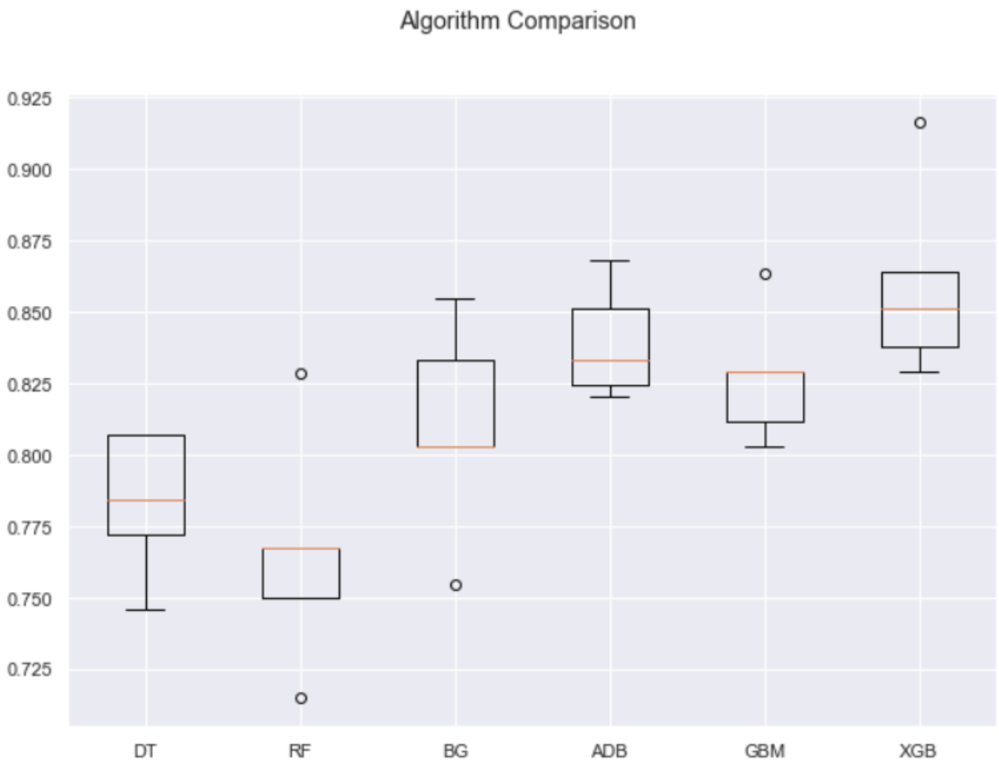


Our best logistic regression model by Cross-Validation Score is Ridge Regression with a Power Transformation and Cluster Centroids undersampling.

Overall, while recall on both the train and test sets are extremely high, other performance metrics are much lower and suggest overfitting. This certainly seems to be the case, because almost 70% of test data was predicted to default. This would be a major problem when put into practice, so **this linear regression model should not be put into production".**

Boosting and Bagging Models

Initial Cross-Validation

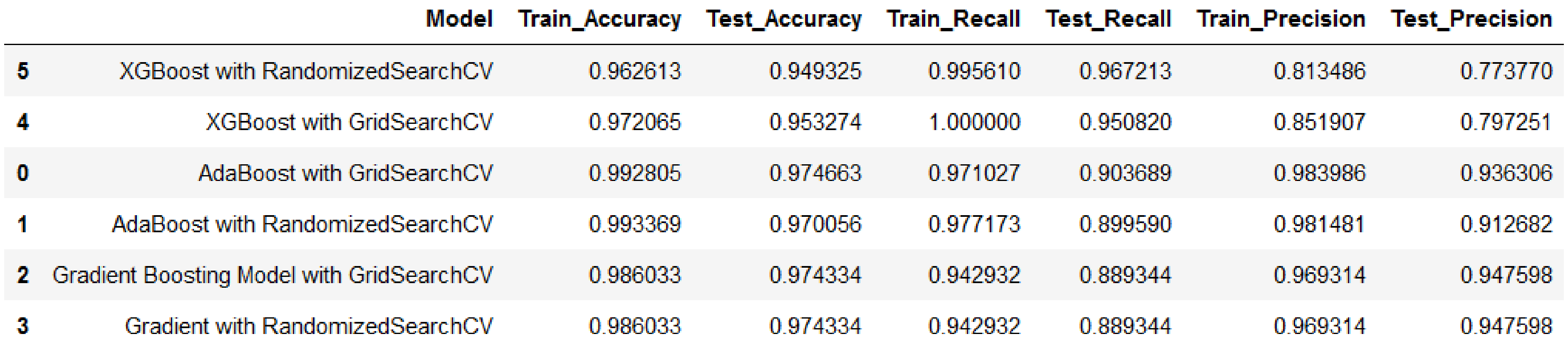


The models that we are testing are:

1. Decision Tree
2. Random Forest
3. Bagging Classifier
4. AdaBoost Classifier
5. Gradient Boosting Classifier
6. XGBoost Classifier

The three best models based on average cross-validation scores are our 3 boosting models: Adaboost, Gradient Boosting, and XGBoost. These are the 3 models that we will proceed forward with for hyperparameter tuning.

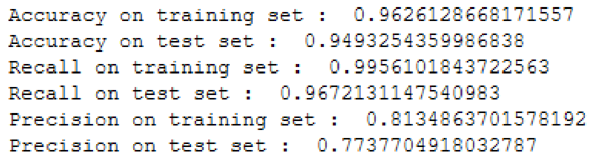
Results of Hyperparameter Tuning



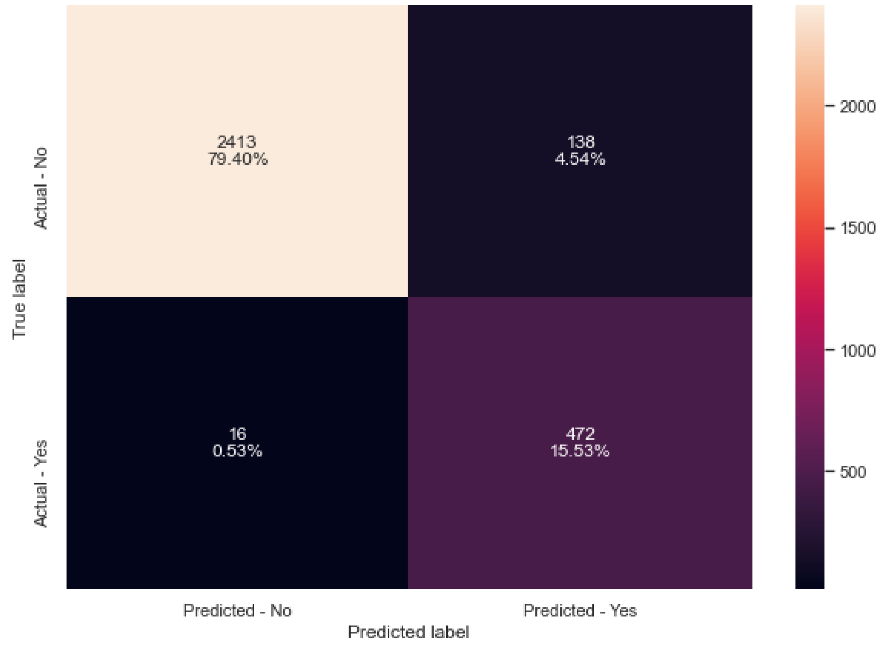
	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision
5	XGBoost with RandomizedSearchCV	0.962613	0.949325	0.995610	0.967213	0.813486	0.773770
4	XGBoost with GridSearchCV	0.972065	0.953274	1.000000	0.950820	0.851907	0.797251
0	AdaBoost with GridSearchCV	0.992805	0.974663	0.971027	0.903689	0.983986	0.936306
1	AdaBoost with RandomizedSearchCV	0.993369	0.970056	0.977173	0.899590	0.981481	0.912682
2	Gradient Boosting Model with GridSearchCV	0.986033	0.974334	0.942932	0.889344	0.969314	0.947598
3	Gradient with RandomizedSearchCV	0.986033	0.974334	0.942932	0.889344	0.969314	0.947598

Overall, hyperparameter tuning definitely helped increase the performances of all models. On this set, GridSearch generally performed slightly better than RandomSearch, but this was because (except for gradient boosting) the same parameters were used for both. RandomSearch, however, was able to take minutes off of the computing time, so it may be worth the tradeoff, especially if RandomSearch is used to search a larger hyperparameter space than GridSearch for similar or less computation time.

The Final Model



```
Accuracy on training set :  0.9626128668171557
Accuracy on test set :  0.9493254359986838
Recall on training set :  0.9956101843722563
Recall on test set :  0.9672131147540983
Precision on training set :  0.8134863701578192
Precision on test set :  0.7737704918032787
```

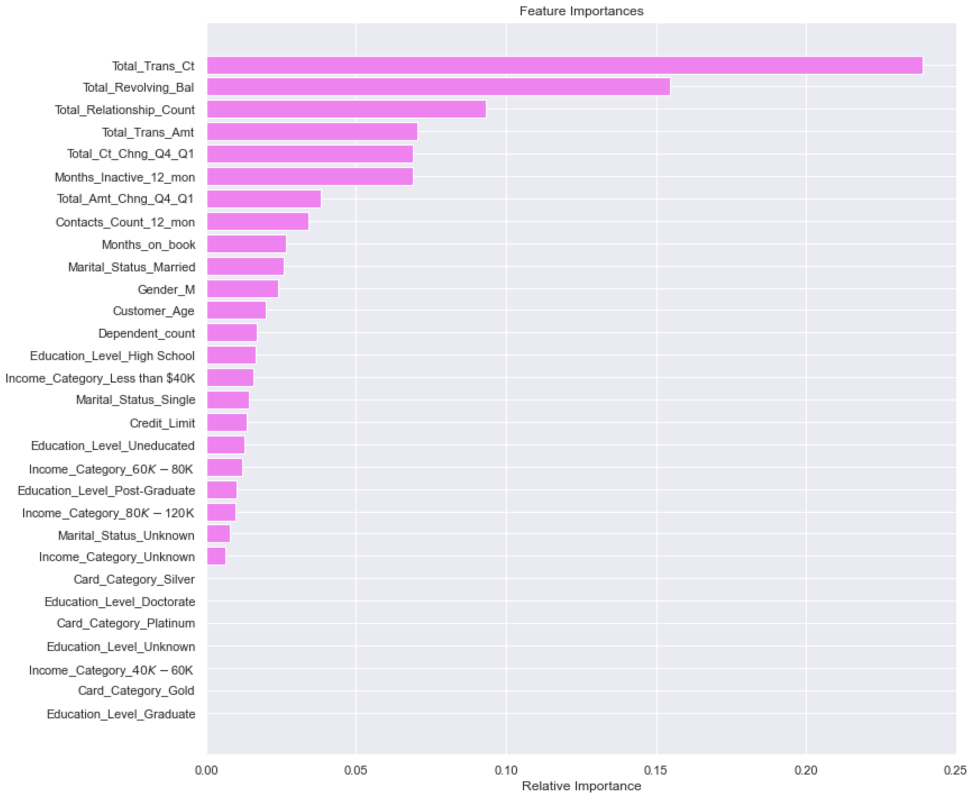


The XGBoost model tuned with RandomizedSearch has the best recall, is strong in all performance metrics, and does not overfit.

Therefore, **the final model model to be put into production will be XGBoost tuned with RandomizedSearch.**

- This model predicts that 20.07% of customers will attrit.
- It is good enough to put into production!

Feature Importances of Final Model



Total transaction count is by far the most important feature in this dataset. EDA made it clear that customers with a lower amount of transactions in the past year were more likely to attrit.

Total Revolving Balance, Total Relationship Count, Total Transaction Amount, change in transaction count from Q1 to Q4, and number of months inactive in the last year are the next most important features.

Insights and Recommendations

1. Put the model we have developed into production.
   - It has performed very well and be confidently used to predict customer attrition.

2. Gather data on more variables.
   - The histogram of Total Transaction Amount consisted of multiple distributions that could not be explained by the current variables. Finding a variable to explain these distributions could provide much insight into customers and their spending habits.

3. Almost all customers own a card from the Blue Category. There is much room for these other card categories to grow.
   - The bivariate analysis provides insights into the relationships between Card_Category and other variables if Thera bank wishes to grow the other card categories.

4. Develop a better understanding of this dataset.
   - A lot of curious questions about the behavior of the data, such as bumps at extreme values, missing value imputation, and an unusually large amount of customers having a relationship with the bank of exactly 36 months suggests that this data was handled by someone before this current analysis. Some relationships between variables also raised interest, such as that between months_on_book and customer age, so more domain expertise may be useful in better understanding this data set.

5. Provide incentives to increase spending. Customers are using their credit cards less frequently.
   - This could be a bigger problem for Thera Bank than even attrition. Perhaps some rewards programs or something similar will encourage customers to use their credit cards more.