Introduction

In this project, we will analyze the credit card customers churn dataset. This dataset is about the various characteristics of the bank's credit card customers including customers’ age, gender, number of dependents, education level, and marital status, as well as their historical behavior data, and whether they have churned or not. The goal of this report is to help the bank reduce churn by investigating which user characteristics are more likely to lead to churn. To analyze such binary classification problems, three models will be introduced including logistic regression, decision trees, and random forests to answer three main questions: What are the three most critical characteristics that affect credit card customer churn? Which model performs best in predicting customers’ churning? And What suggestions can we provide to help bank managers reduce customer churn?

Data Cleaning

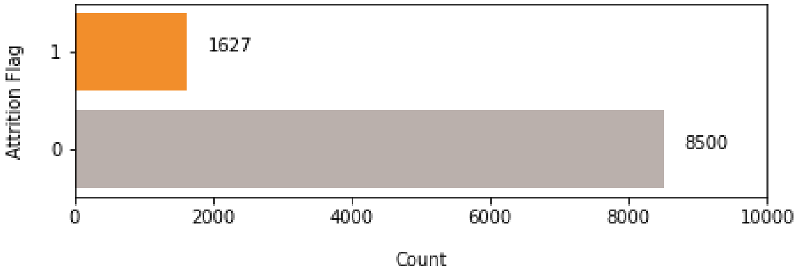
The “Bank Churners” dataset selected from Kaggle contains 10,127 rows and 20 useful variables (8 categorical variables and 12 numerical variables). First, we removed ClientNum and the last two useless variables. The target variable Attrition_Flag is then encoded as 1 and 0, where 1 means that the customer has churned and 0 means that the customer has not churned. This dataset does not contain any missing values. By examining the distributions of all categorical variables, the ‘Blue’ category in the variable Card_Category accounts for the majority, so the other categories are merged into the ‘Others’ category.

By checking the distributions of all numerical variables, the variables Credit_Limit, Avg_Open_To_Buy, and Total_Trans_Amt are right-skewed and therefore need to be converted to their logarithms from. After the logarithmic transformation, the outliers of all numerical variables were detected by using a range of ±3 standard deviations from the mean. The results shows that the variable Customer_Age, Total_Amt_Chng_Q4_Q1, Total_Ct_Chng_Q4_Q1 and Log_Avg_Open_To_Buy contain some outliers, and we used capping and flooring method to limit these outliers to within ±3 standard deviations from the mean.

Exploratory Data Analysis

Our target variable Attrition_Flag contains two values, namely 1 and 0, meaning whether the credit card customers have churned or not. And about 16% of the customers have churned, while the remaining 84% have not.

Figure 1: Count of Churned or Not



By plotting the distribution of churned or not in each group under each categorical variable, we can see that most variables, including Gender, Education_Level, Marital_Status, Income_Category and Card_Category, are not very discriminative for whether a customer has churned (See Appendix 1-5). According to Appendix 6, the number of inactive months in a 12-month period shows a significant distinction between churned or not. The number ‘0’ indicates that users who were active each month had the highest percentage of churn. This may be an error caused by too small sample size of the ‘0’ class. For other classes, the percentage of user churn is higher as the number of months of user inactivity increases. Among them, the proportion of users who have not been active for 4 months is as high as 30%. Interestingly, users with inactive months of 5 and 6 had a lower percentage of churn. According to Appendix 7, the higher the number of contacts in 12 months, the higher the percentage of churn.

According to the boxplots of all numerical variables grouped by churned or not, we can see that the variables Customer_Age, Dependent_count, Months_on_book are not very discriminative for customers who churned or not, because churned users and non-churned users have the same distribution (See Appendix 8). According to Appendix 9, the distribution of total relationship count among users who churned is lower than those who did not churn, but the difference is not significant. A similar distribution appears in the distribution of the variable Total_Amt_chng_Q4_Q1. It means that the total amount change does not make much difference to user churn (Appendix 10). The variables Total_Revolving_Bal, Total_Trans_Ct, Total_Ct_Chng_Q4_Q1, Avg_Utilization_Ratio and Log_Total_Trans_Amt have a greater discrimination for whether the user have churned or not. The value of these variables is lower for churned customers than for non-churned customers (See Appendix 11-15). Finally, according to Appendix 16-17, the variables Log_Credit_Limit and Log_Avg_Open_To_Buy are not very discriminative for whether the customer is in default or not.

According to the heatmap of all variables (Appendix 18), the target variable Attrition_Flag has a slight negative correlation with Total_Trans_Ct and Total_Ct_Chng_Q4_Q1, and a slight positive correlation with Contacts_Count_12_mon. In addition, there are strong positive correlations among some independent variables, and the variable Avg_Utilization_Ratio exhibits strong positive or negative correlations with multiple variables. This leads to a potential risk of multicollinearity, which needs to be addressed in subsequent analysis.

Modeling

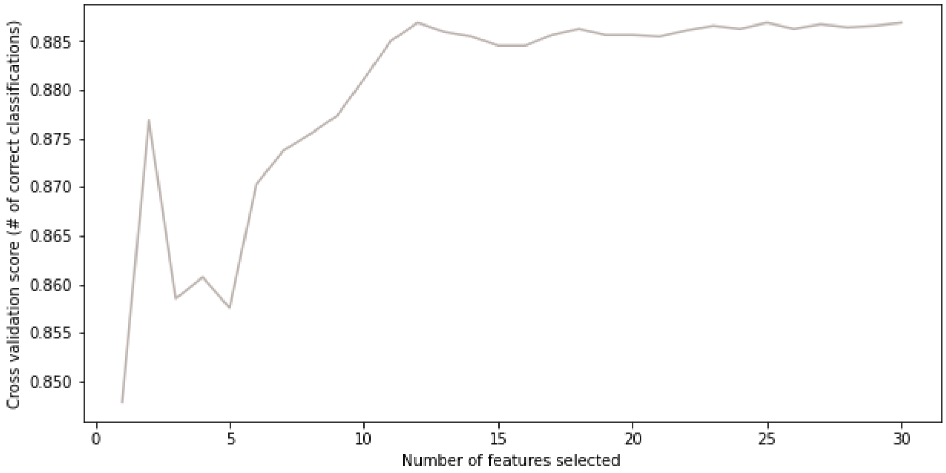
Logistic Regression Model

The logistic regression model is suitable for solving binary classification problems. We used logistic regression to generate the predictive model for this binary classification problem. Before that, we used the One Hot Encode Method to convert all categorical variables into N-1 variables. And we randomly selected 75% of the data as the training set and 25% as the test set and normalized all variables to have values between 0 and 1. Our target variable is unbalanced (16% of customers have churned), which will cause our model to be biased towards predicting the majority class. So, we used SMOTE to oversample the minority class so that the proportions of 1 and 0 are equal.

Then, we used the training set to fit our first logistic regression model, and the results (See Appendix 19) show that the 3 most significant variables related to whether the customers have churned are Total_Trans_Ct, Log_Total_Trans_Amt, and Total_Relationship_Count (with p-value < 0.001). Among them, Total_Trans_Ct is negatively correlated with the target variable, indicating that customers with more total transactions have a lower probability of churn. Interestingly, the Log_Total_Trans_Amt is positively correlated with the target variable, meaning that the higher the transaction amount of the customer, the greater the likelihood of churn. However, many variables are correlated with each other, therefore Model 1 has a multicollinearity problem.

Since some variables are uncorrelated with the target variable, and the model have the risk of multicollinearity, we performed feature selection using the Recursive Feature Elimination with Cross-Validation (RFECV) method with the recall rate as the evaluation criterion (We want to reduce the false negatives). The results show that 12 of the 30 variables were retained, and as the variables increase, the model's recall rate for the training set is also increasing (Figure 2).

Figure 2: Results of Cross-Validation Recursive Feature Elimination

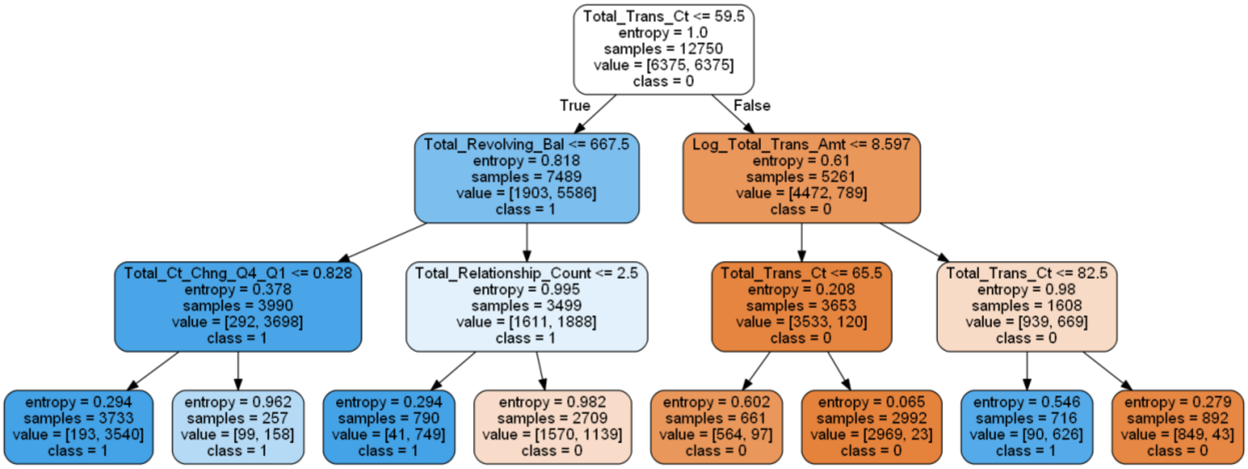


Then we checked for multicollinearity among variables using VIF and 7 variables were retained to build our model 2 (Appendix 20). Compared with model 1, the VIFs of all variables in model 2 are less than 10, indicating that the problem of multicollinearity has been solved. However, the recall rate of the model's prediction on the test set is lower than those of model 1. Hence the performance of logistic regression is mediocre.

Decision Tree Model

Another available classification prediction model is the decision tree. Decision tree model has the advantage of being easy to interpret and can fit nonlinear relationships between variables. First, we used the training set to train a basic decision tree model as a reference model. Then we used Cross-Validation Random Search and Grid Search methods to perform hyperparameter tuning to find the optimal model parameters. After performing the Random Search and the Grid Search, the maximum depth of the optimal model is 80; the minimum samples split is 2, and the minimum samples leaf is 8. The recall rate of the decision tree model on the test set is 90.17%, meaning that the decision tree model has a higher recall rate than the logistic regression models. It is also shown that the decision tree model can reduce the false negatives of the prediction result. According to the feature importance table (Appendix 21), the most important 3 variables are Total_Trans_Ct, Log_Total_Trans_Amt, and Total_Revolving_Bal. The importance of these three variables accounted for 66.70% of the total. By limiting the maximum depth of the tree to 3, we got a graph of the decision tree model.

Figure 3: Decision Tree with Depth of three



According to Figure 3, customers with a lower total transaction count and a lower total revolving balance are more likely to churn. While customers with a higher total transaction amount are more likely to churn. So, the result is the same as that of the logistic regression model.

Random Forest Model

The random forest model, by bagging multiple decision trees and voting on the results, can also be used for this binary classification problem. Random forest models have lower variance than decision tree models and can reduce the risk of overfitting.

First, we used the training set to train a basic random forest model as a reference model. Then we also used the Random Search and the Grid Search methods to perform hyperparameter tuning to find the optimal model parameters. By limiting the number of trees to 500, maximum depth of the trees to 80, minimum samples leaf to 1 and minimum samples split to 2, the prediction recall rate of the random forest model on the test set is 85.75% (Table 1). This is lower than the recall rate of the decision tree model, which means that the random forest model has more false negatives in the prediction results of the test set, so the performance of the random forest model is mediocre.

Table 1: Summary of All Models and the Recall Rate

Model	**Recall**
Logistic Regression	86.24%
Logistic Regression 2	80.59%
Decision Tree	87.47%
Decision Tree GridSearch	90.17%
Random Forest	85.26%
Random Forest GridSearch	85.75%

According to Appendix 22, the feature importance obtained by the random forest model is very close to that of the decision tree and the first 3 most important variables are the same as those of the decision tree. And the importance of the 3 variables Total_Trans_Ct, Log_Total_Trans_Amt, and Total_Revolving_Bal accounts for 37.85%.

Figure 4: Confusion Matrix of Decision Tree Model with Grid Search

1977	148	Not Churned	Actual
40	367	Churned	
Not Churned	Churned		
Predicted			

In summary, the best model is the decision tree model with grid search, which has the highest recall rate. It means that it has the lowest false negatives for the prediction results of the test set. Using this model to predict whether a user has churn can effectively reduce the loss of churning customers.

Conclusion

Through the above analysis, the churn of customers can be predicted by the three most important features: the total transactions of user, the total transaction amount, and the total revolving balance. And the lower the values of total transaction count and total revolving balance, the more likely customers will churn. But the customers with a higher total transaction amount are more likely to churn. Based on the conclusion, we recommend that the bank use these variables as indicators for daily monitoring of customer churn. Once a customer's three values reach a certain threshold, the credit card manager should contact the customer and ask him/her to change his/her mind about potential churn.

References

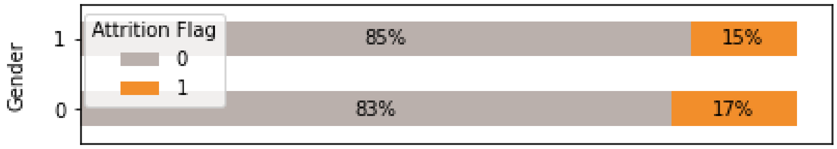
Agresti, A. (2003, March 31). Categorical Data Analysis, 2nd Edition. Wiley.com. Retrieved November 15, 2021.

Goyal, S. (2020, November 19). *Credit Card customers*. Kaggle. https://www.kaggle.com/sakshigoyal7/credit-card-customers

Z., Adrians, & Pieter; Zantinge, D. A. (1996, January 1). Data Mining., 15, 2021.

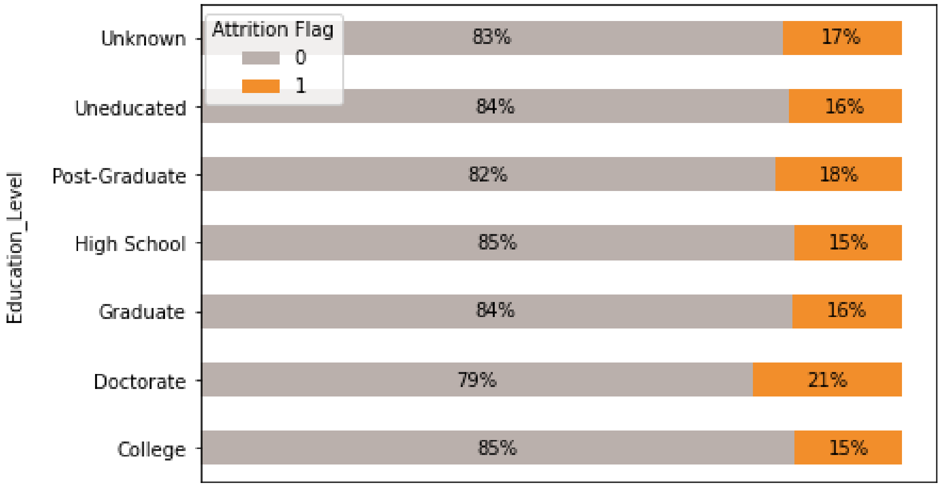
Appendix 1

Percentage of Churned or Not by Gender



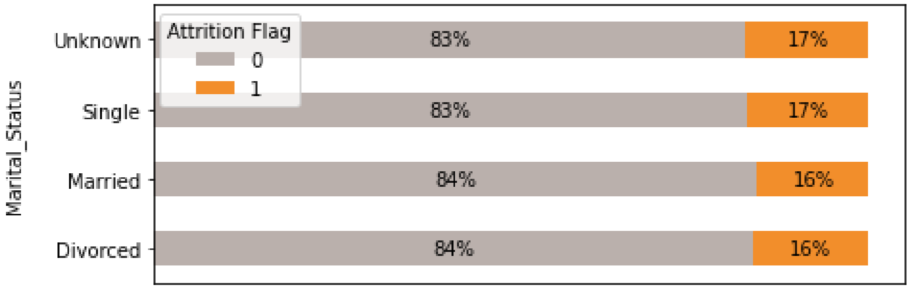
Appendix 2

Percentage of Churned or Not by Education Level



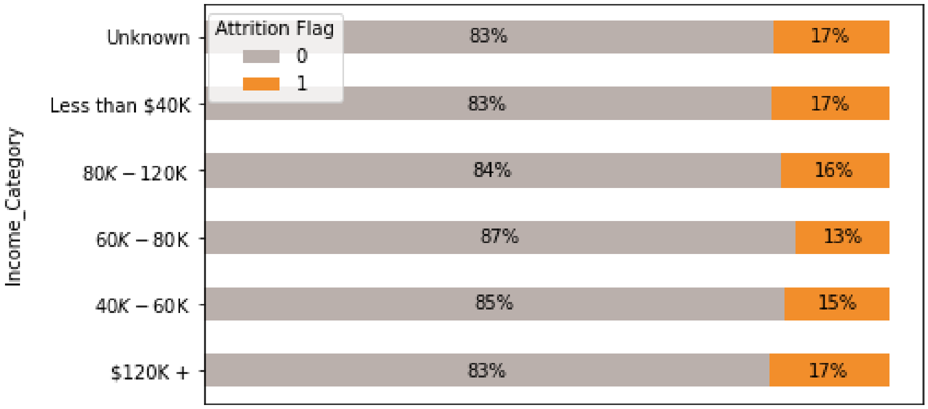
Appendix 3

Percentage of Churned or Not by Marital Status



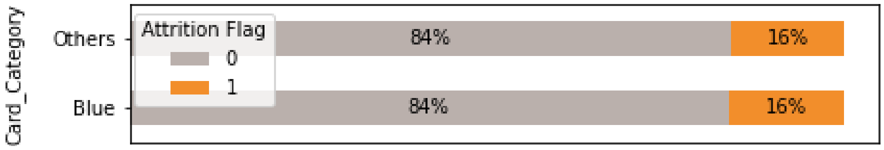
Appendix 4

Percentage of Churned or Not by Income Category



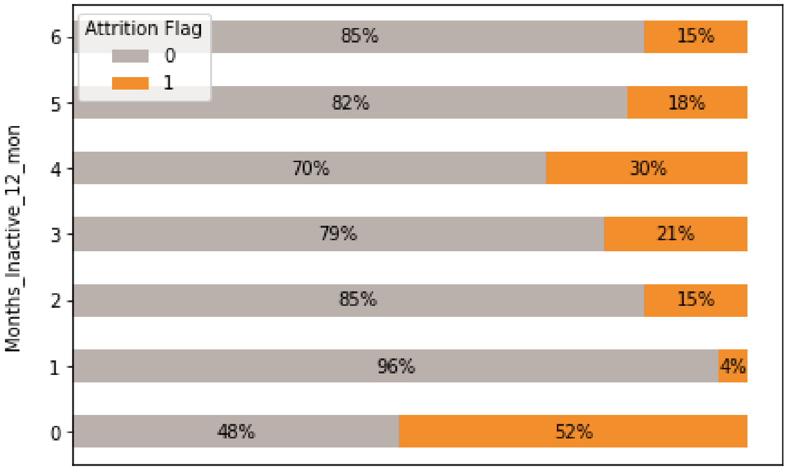
Appendix 5

Percentage of Churned or Not by Card Category



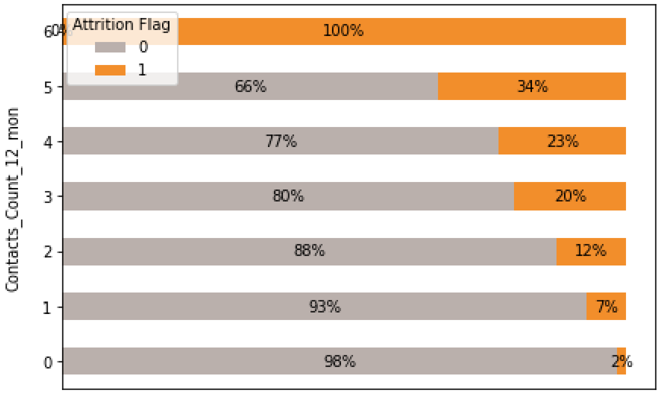
Appendix 6

Percentage of Churned or Not by Months of Inactive in 12 Months



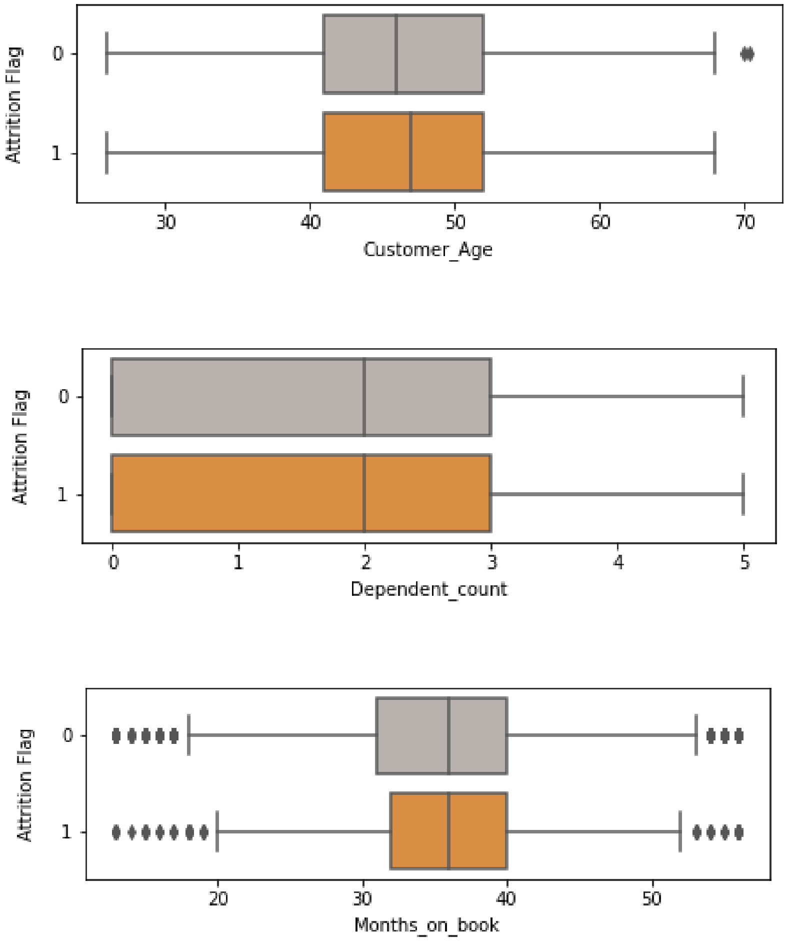
Appendix 7

Percentage of Churned or Not by Contacts Count in 12 Months



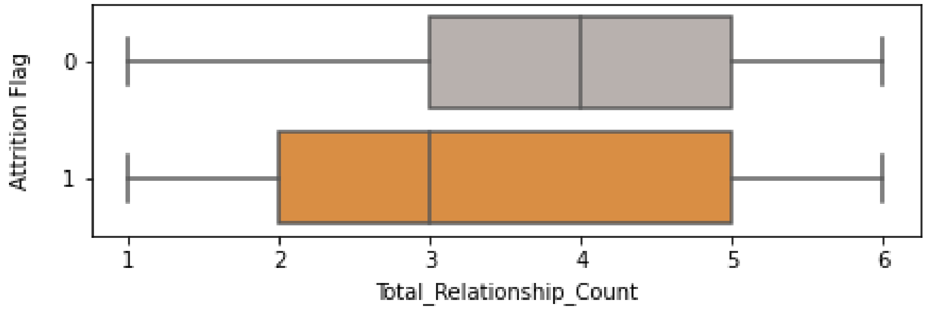
Appendix 8

Boxplots of Age / Dependent Count / Months on Book Grouped by Churned or Not



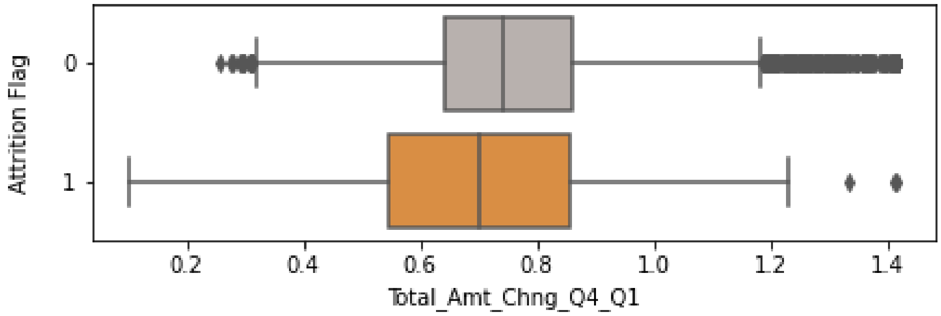
Appendix 9

Boxplots of Total Relationship Count Grouped by Churned or Not



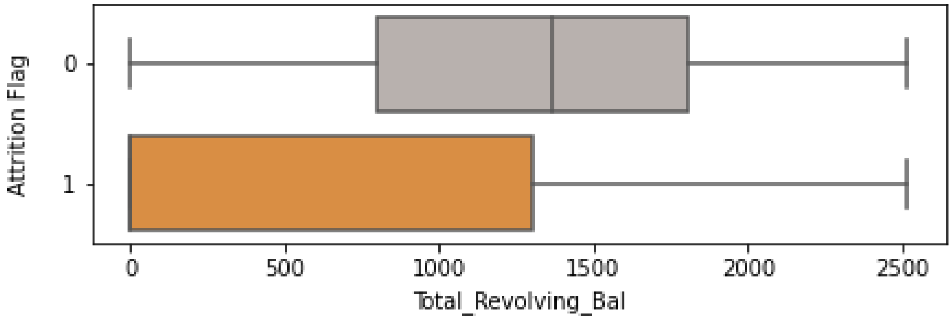
Appendix 10

Boxplots of Total Amount Change Q4 - Q1 Grouped by Churned or Not



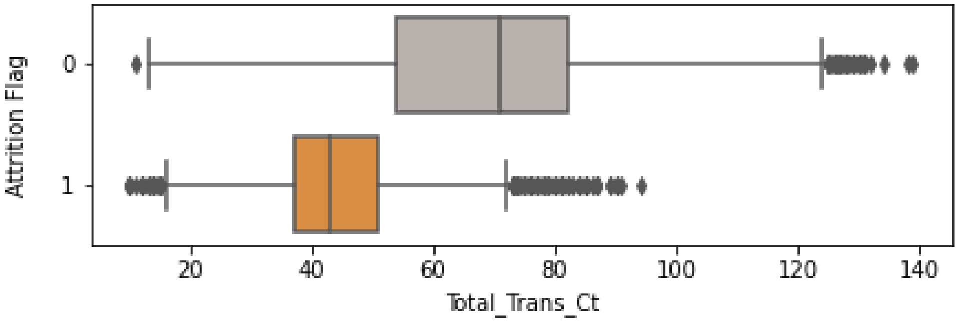
Appendix 11

Boxplots of Total Revolving Balance Grouped by Churned or Not



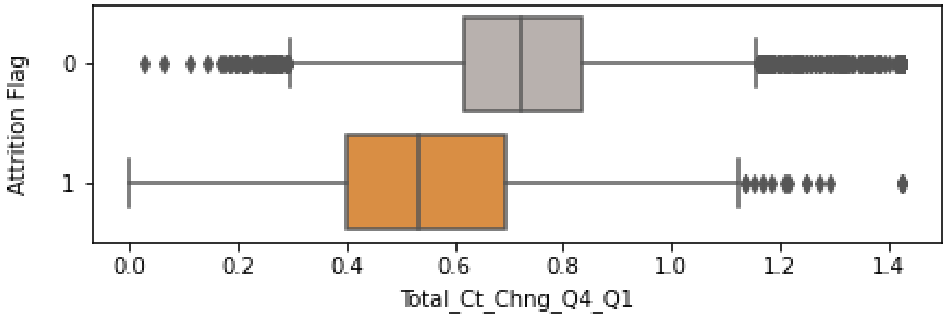
Appendix 12

Boxplots of Total Transaction Count Grouped by Churned or Not



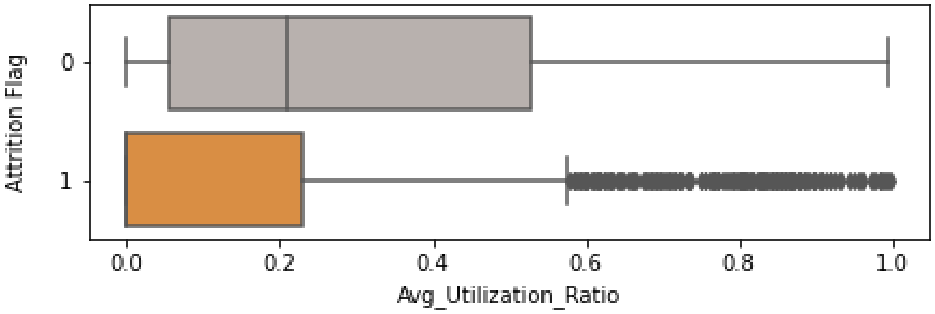
Appendix 13

Boxplots of Total Count Change Q4 - Q1 Grouped by Churned or Not



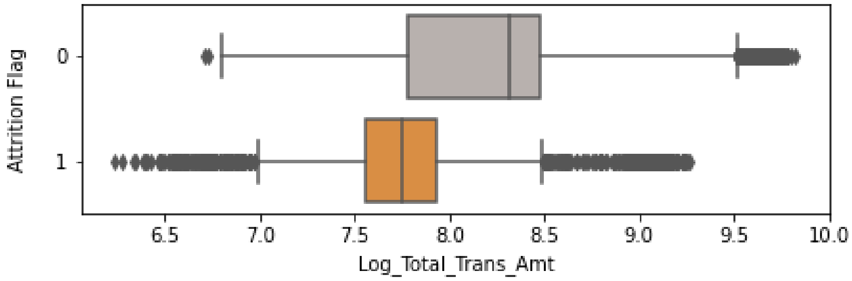
Appendix 14

Boxplots of Average Utilization Ratio Grouped by Churned or Not



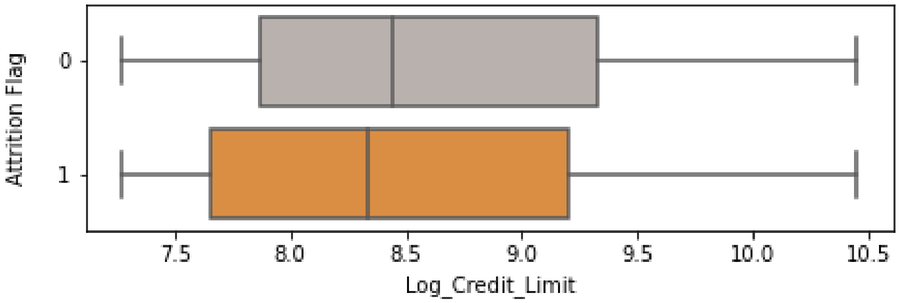
Appendix 15

Boxplots of Log Total Transaction Amount Grouped by Churned or Not



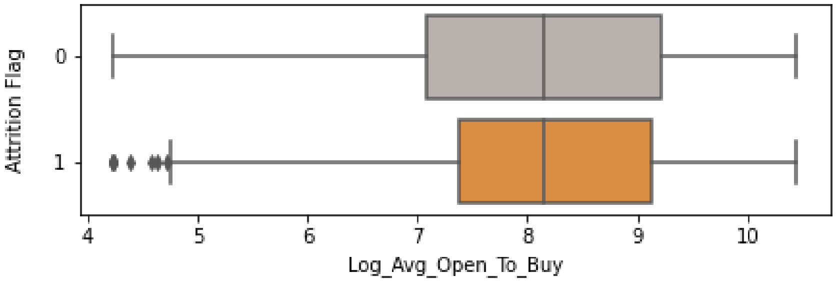
Appendix 16

Boxplots of Log Credit Limit Grouped by Churned or Not



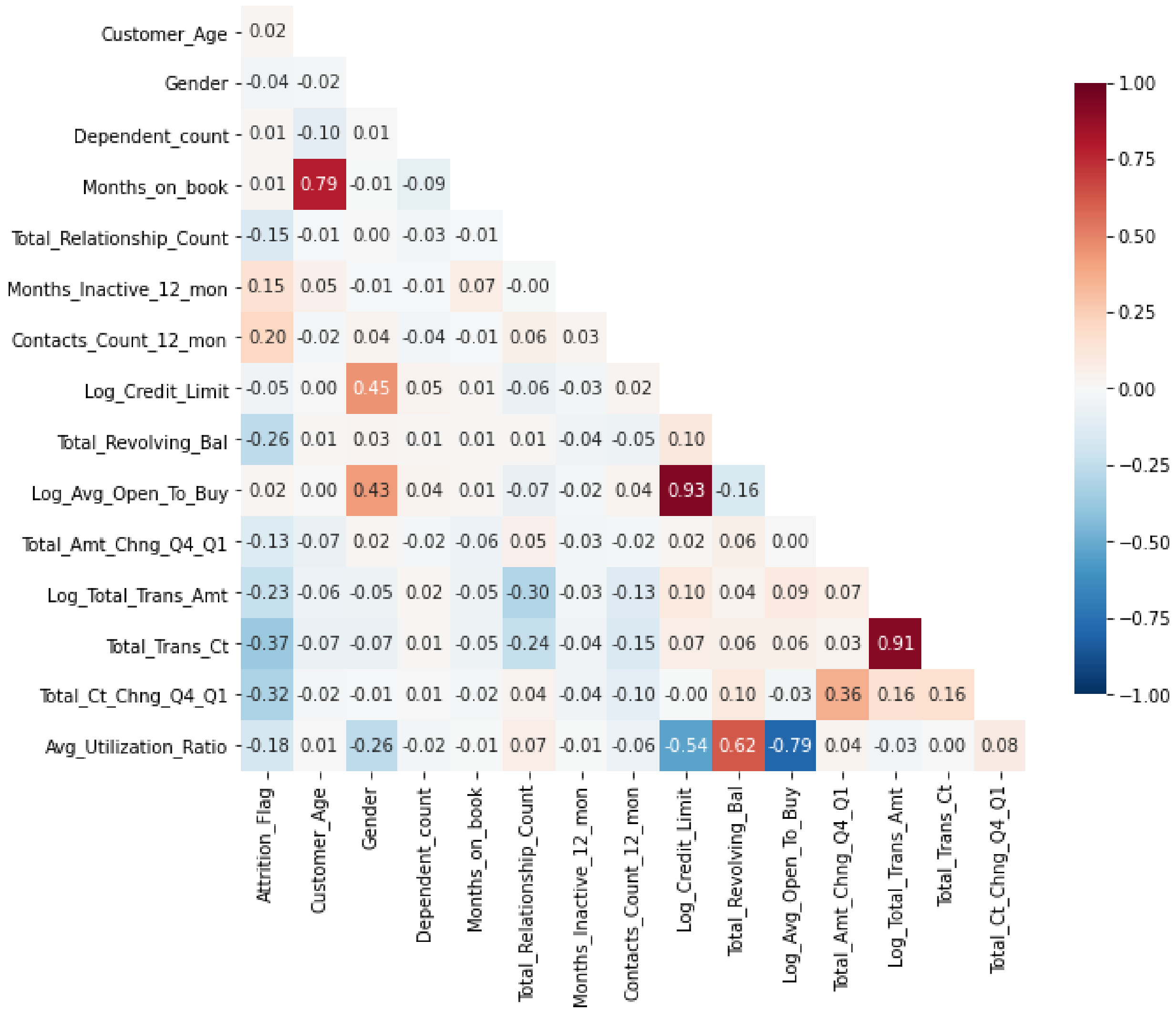
Appendix 17

Boxplots of Log Average Open to Buy Grouped by Churned or Not



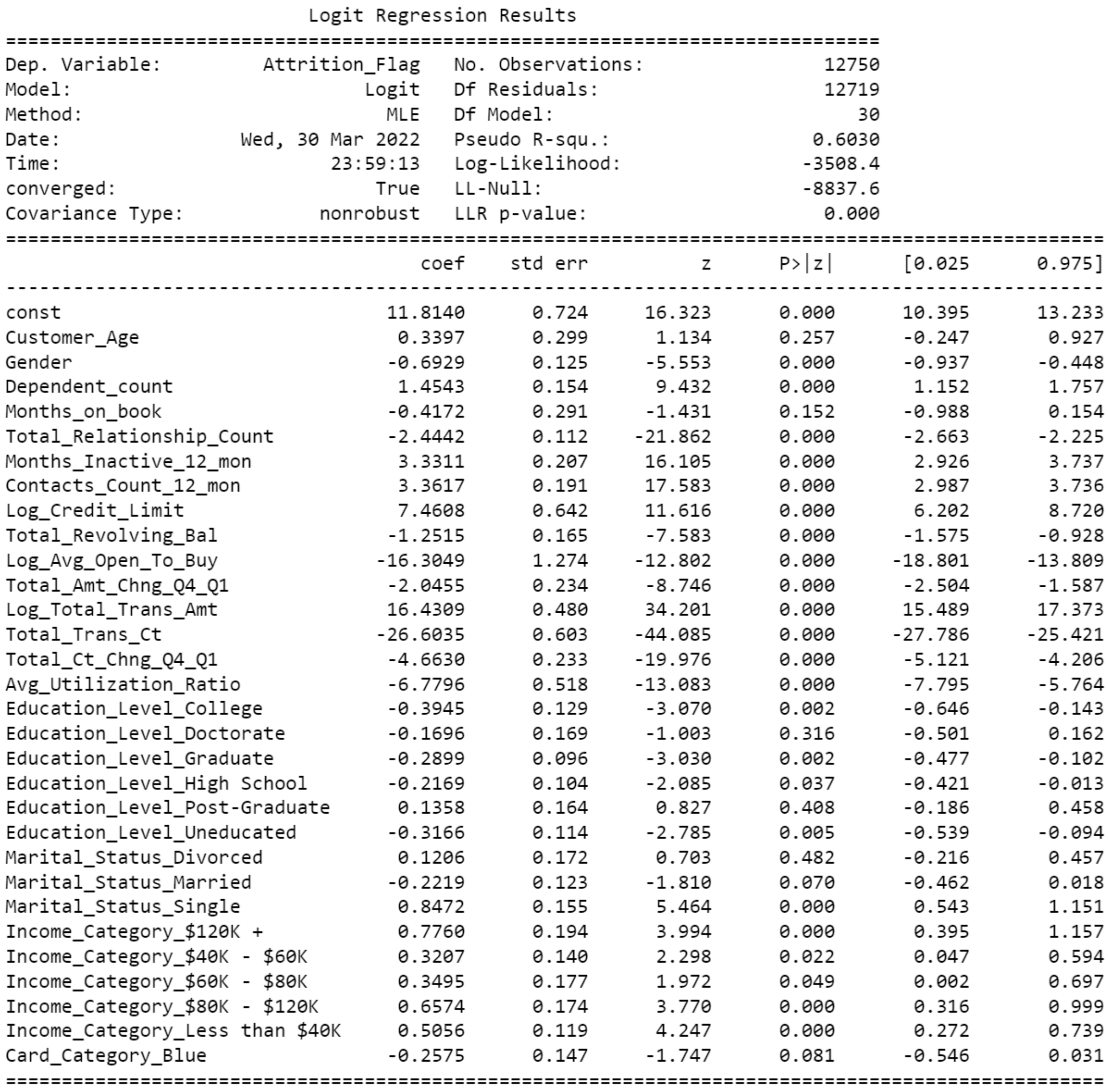
Appendix 18

Heatmap of All Variables



Appendix 19

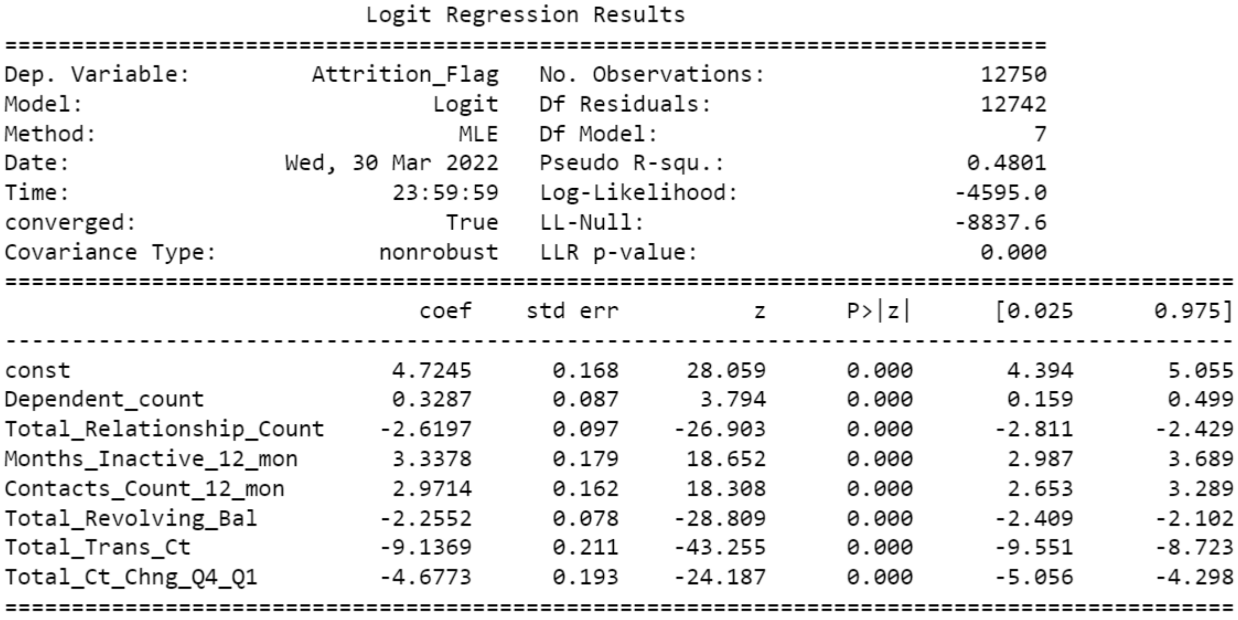
Results of Logistic Regression Model 1



```
                           Logit Regression Results                           
==============================================================================
Dep. Variable:         Attrition_Flag   No. Observations:                12750
Model:                          Logit   Df Residuals:                    12719
Method:                           MLE   Df Model:                           30
Date:                Wed, 30 Mar 2022   Pseudo R-squ.:                  0.6030
Time:                        23:59:13   Log-Likelihood:                -3508.4
converged:                       True   LL-Null:                       -8837.6
Covariance Type:            nonrobust   LLR p-value:                     0.000
==================================================================================================
                                     coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                             11.8140      0.724     16.323      0.000      10.395      13.233
Customer_Age                       0.3397      0.299      1.134      0.257      -0.247       0.927
Gender                            -0.6929      0.125     -5.553      0.000      -0.937      -0.448
Dependent_count                    1.4543      0.154      9.432      0.000       1.152       1.757
Months_on_book                    -0.4172      0.291     -1.431      0.152      -0.988       0.154
Total_Relationship_Count          -2.4442      0.112    -21.862      0.000      -2.663      -2.225
Months_Inactive_12_mon             3.3311      0.207     16.105      0.000       2.926       3.737
Contacts_Count_12_mon              3.3617      0.191     17.583      0.000       2.987       3.736
Log_Credit_Limit                   7.4608      0.642     11.616      0.000       6.202       8.720
Total_Revolving_Bal               -1.2515      0.165     -7.583      0.000      -1.575      -0.928
Log_Avg_Open_To_Buy              -16.3049      1.274    -12.802      0.000     -18.801     -13.809
Total_Amt_Chng_Q4_Q1              -2.0455      0.234     -8.746      0.000      -2.504      -1.587
Log_Total_Trans_Amt               16.4309      0.480     34.201      0.000      15.489      17.373
Total_Trans_Ct                   -26.6035      0.603    -44.085      0.000     -27.786     -25.421
Total_Ct_Chng_Q4_Q1               -4.6630      0.233    -19.976      0.000      -5.121      -4.206
Avg_Utilization_Ratio             -6.7796      0.518    -13.083      0.000      -7.795      -5.764
Education_Level_College           -0.3945      0.129     -3.070      0.002      -0.646      -0.143
Education_Level_Doctorate         -0.1696      0.169     -1.003      0.316      -0.501       0.162
Education_Level_Graduate          -0.2899      0.096     -3.030      0.002      -0.477      -0.102
Education_Level_High School       -0.2169      0.104     -2.085      0.037      -0.421      -0.013
Education_Level_Post-Graduate      0.1358      0.164      0.827      0.408      -0.186       0.458
Education_Level_Uneducated        -0.3166      0.114     -2.785      0.005      -0.539      -0.094
Marital_Status_Divorced            0.1206      0.172      0.703      0.482      -0.216       0.457
Marital_Status_Married            -0.2219      0.123     -1.810      0.070      -0.462       0.018
Marital_Status_Single              0.8472      0.155      5.464      0.000       0.543       1.151
Income_Category_$120K +            0.7760      0.194      3.994      0.000       0.395       1.157
Income_Category_$40K - $60K        0.3207      0.140      2.298      0.022       0.047       0.594
Income_Category_$60K - $80K        0.3495      0.177      1.972      0.049       0.002       0.697
Income_Category_$80K - $120K       0.6574      0.174      3.770      0.000       0.316       0.999
Income_Category_Less than $40K     0.5056      0.119      4.247      0.000       0.272       0.739
Card_Category_Blue                -0.2575      0.147     -1.747      0.081      -0.546       0.031
==================================================================================================
```

Appendix 20

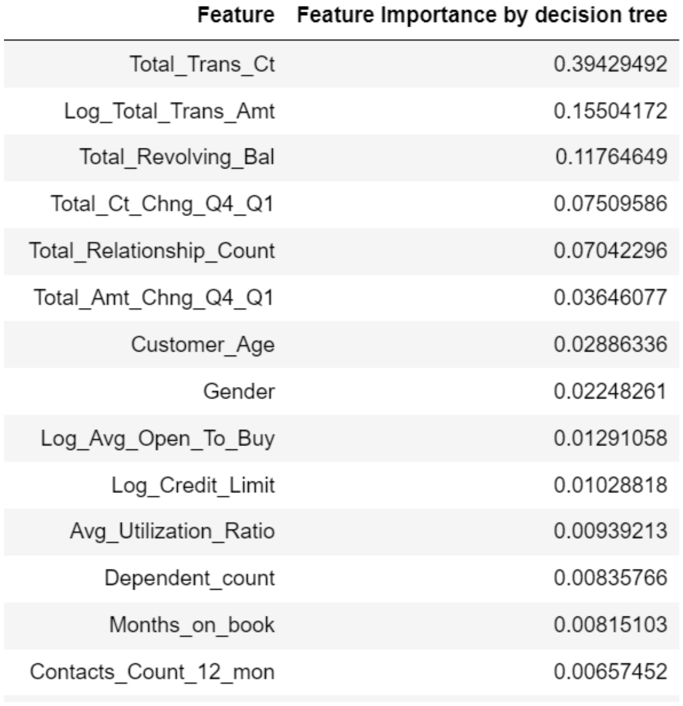
Results of Logistic Regression Model 2



```
                           Logit Regression Results
==============================================================================
Dep. Variable:         Attrition_Flag   No. Observations:                12750
Model:                          Logit   Df Residuals:                    12742
Method:                           MLE   Df Model:                            7
Date:                Wed, 30 Mar 2022   Pseudo R-squ.:                  0.4801
Time:                        23:59:59   Log-Likelihood:                -4595.0
converged:                       True   LL-Null:                       -8837.6
Covariance Type:            nonrobust   LLR p-value:                     0.000
=============================================================================================
                               coef    std err          z      P>|z|      [0.025       0.975]
---------------------------------------------------------------------------------------------
const                        4.7245      0.168     28.059      0.000       4.394        5.055
Dependent_count              0.3287      0.087      3.794      0.000       0.159        0.499
Total_Relationship_Count    -2.6197      0.097    -26.903      0.000      -2.811       -2.429
Months_Inactive_12_mon       3.3378      0.179     18.652      0.000       2.987        3.689
Contacts_Count_12_mon        2.9714      0.162     18.308      0.000       2.653        3.289
Total_Revolving_Bal         -2.2552      0.078    -28.809      0.000      -2.409       -2.102
Total_Trans_Ct              -9.1369      0.211    -43.255      0.000      -9.551       -8.723
Total_Ct_Chng_Q4_Q1         -4.6773      0.193    -24.187      0.000      -5.056       -4.298
=============================================================================================
```

Appendix 21

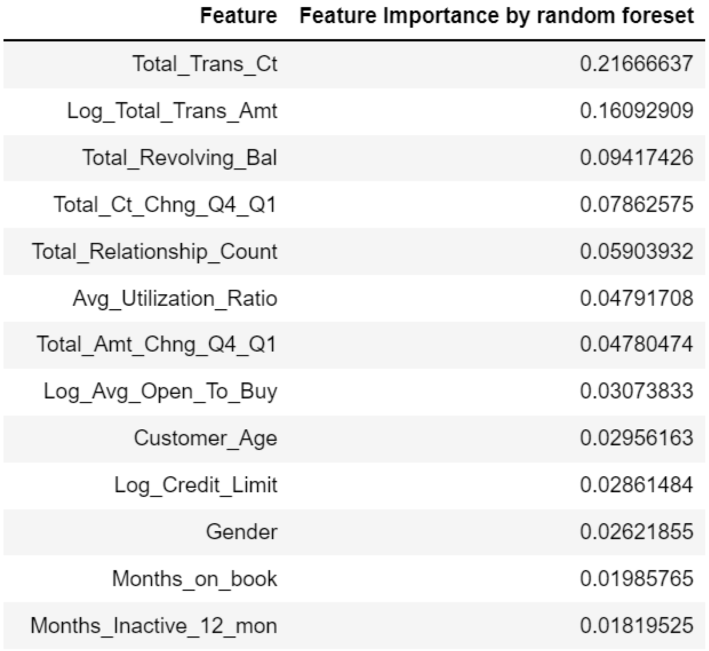
Feature Importance by Decision Tree Model



Feature	Feature Importance by decision tree
Total_Trans_Ct	0.39429492
Log_Total_Trans_Amt	0.15504172
Total_Revolving_Bal	0.11764649
Total_Ct_Chng_Q4_Q1	0.07509586
Total_Relationship_Count	0.07042296
Total_Amt_Chng_Q4_Q1	0.03646077
Customer_Age	0.02886336
Gender	0.02248261
Log_Avg_Open_To_Buy	0.01291058
Log_Credit_Limit	0.01028818
Avg_Utilization_Ratio	0.00939213
Dependent_count	0.00835766
Months_on_book	0.00815103
Contacts_Count_12_mon	0.00657452

Appendix 22

Feature Importance by Random Forest Model



Feature	Feature Importance by random foreset
Total_Trans_Ct	0.21666637
Log_Total_Trans_Amt	0.16092909
Total_Revolving_Bal	0.09417426
Total_Ct_Chng_Q4_Q1	0.07862575
Total_Relationship_Count	0.05903932
Avg_Utilization_Ratio	0.04791708
Total_Amt_Chng_Q4_Q1	0.04780474
Log_Avg_Open_To_Buy	0.03073833
Customer_Age	0.02956163
Log_Credit_Limit	0.02861484
Gender	0.02621855
Months_on_book	0.01985765
Months_Inactive_12_mon	0.01819525