OVERVIEW

Introduction

Background:

A bank manager is worried that more and more customers are leaving their credit card services. They want to figure out who is going to get churned so they can proactively to provide them better services and turn customers' decisions in the opposite direction.

Business questions:

- What are the three most critical characteristics that affect credit card customer churn?
- Which model performs best in predicting customer churn?
- What suggestions can we provide to help bank managers reduce customer churn?

Dataset – Bank Churners

This dataset from Kaggle has 10,127 rows and 20 useful variables (12 numerical and 8 categorical).

Target variable:

Attrition_Flag: 1: Churned, 0: Not Churned

Independent variables:

Numerical:

Customer_Age

Dependent_count

Months_on_book: months since loan origination date

Total_Relationship_Count: total number of products held

Credit_Limit

Total_Revolving_Bal: total unpaid portion

Avg_Open_To_Buy: average difference between the credit limit assigned and the present balance

Avg_Utilization_Ratio: average ratio of current borrowing to available limit

Total_Trans_Ct: total transaction count

Total_Trans_Amt: total transaction amount

Total_Ct_Chng_Q4_Q1: change in total transaction count from Q4 to Q1

Total_Amt_Chng_Q4_Q1: change in total transaction amount from Q4 to Q1

Categorical:

Gender

Education_Level

Marital_Status

Income_Category

Card_Category

Months_Inactive_12_mon

Contacts_Count_12_mon

Data Cleaning

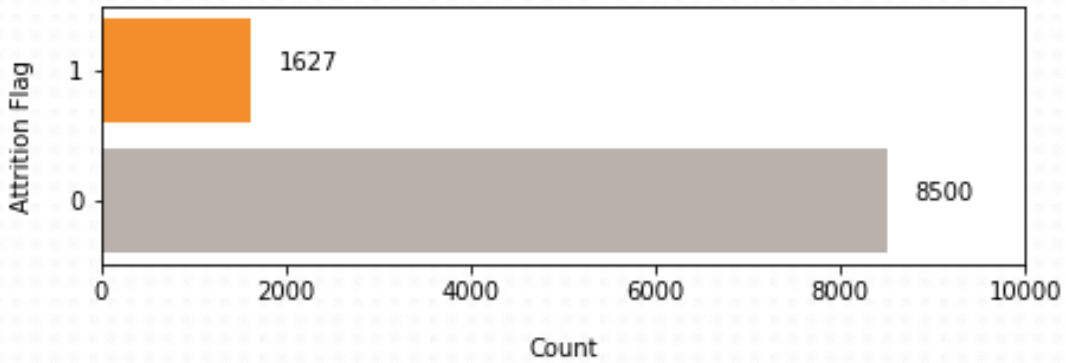
- Removed ClientNum and the last two useless variables (Naive_Bayes_Classifier 1 and 2).
- The ‘Blue’ category in the variable Card_Category accounts for the majority, so the other categories were merged into the ‘Others’ category.
- Avg_Open_To_Buy, and Total_Trans_Amt were right-skewed therefore need to be converted to their logarithms from.
- Used capping and flooring method to limit the outliers to within ±3 standard deviations from the mean for variables: Customer_Age, Total_Amt_Chng_Q4_Q1, Total_Ct_Chng_Q4_Q1 and Log_Avg_Open_To_Buy.

Exploratory Data Analysis

Imbalanced Target Variable: Attrition_Flag

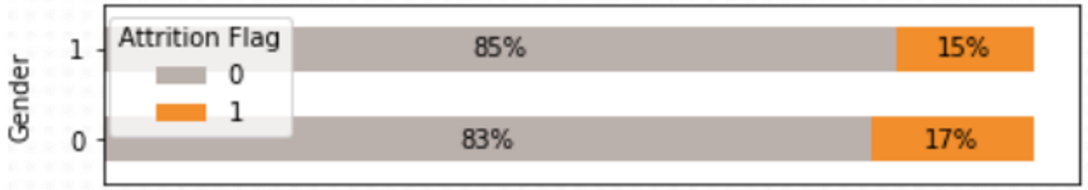
About 16% of the customers have churned, while the remaining 84% have not.

Count of Churned or Not

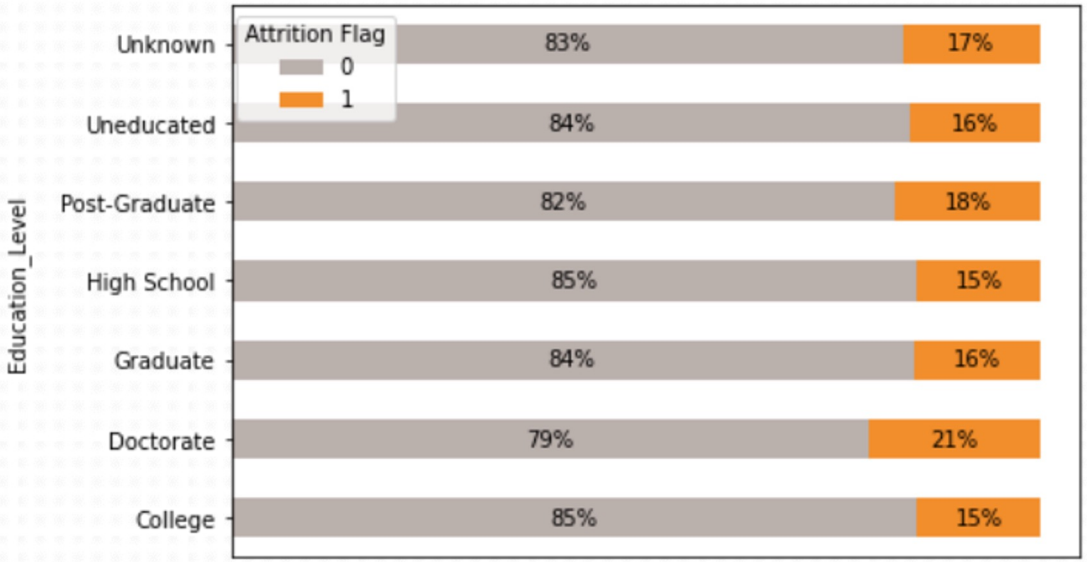


Gender, **Education Level**, **Marital Status**, **Income Category** and **Card Category** are not very discriminative for whether a customer has churned.

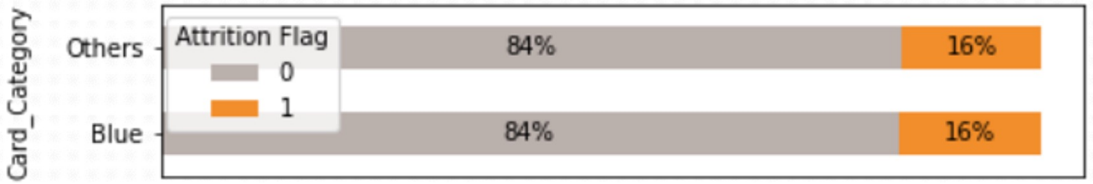
Percentage of Churned or Not by Gender



Percentage of Churned or Not by Education Level



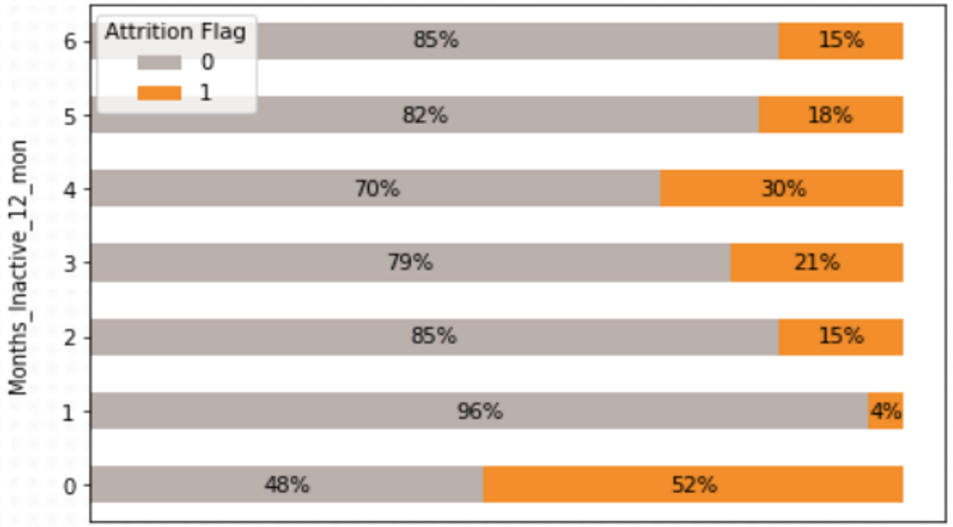
Percentage of Churned or Not by Card Category



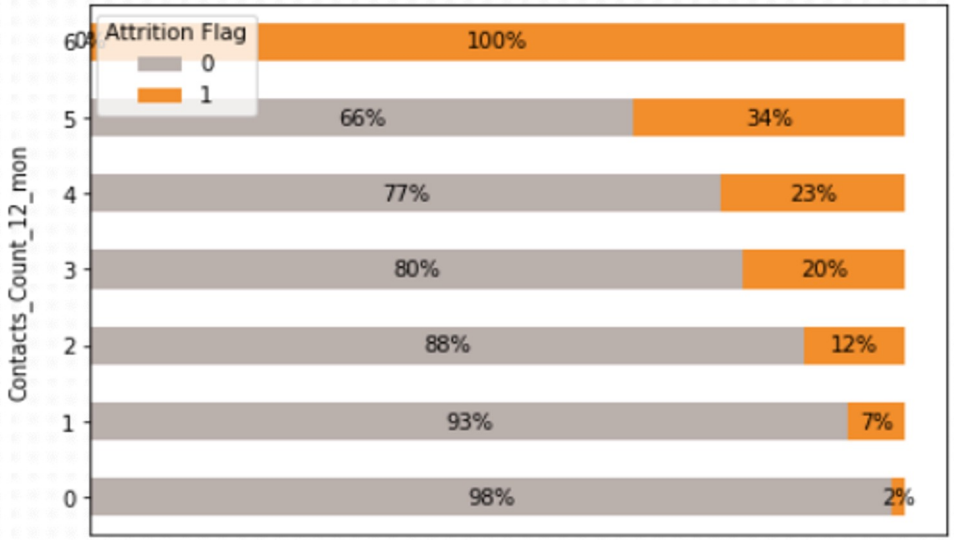
Exploratory Data Analysis

- The number of inactive months in a 12-month period shows a significant distinction between churned or not.
- The percentage of user churn is higher as the number of months of user inactivity increases.
- The churn rate for users who have not been active for 4 months is 30%.
- The higher the number of contacts in 12 months, the higher the percentage of churn.

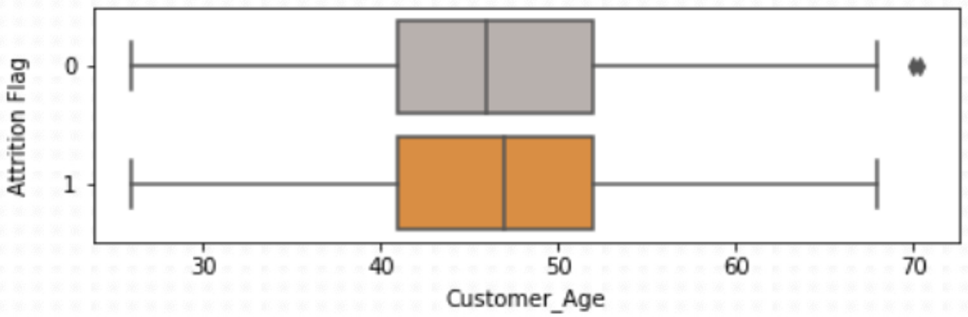
Percentage of Churned or Not by Months of Inactive in 12 Months

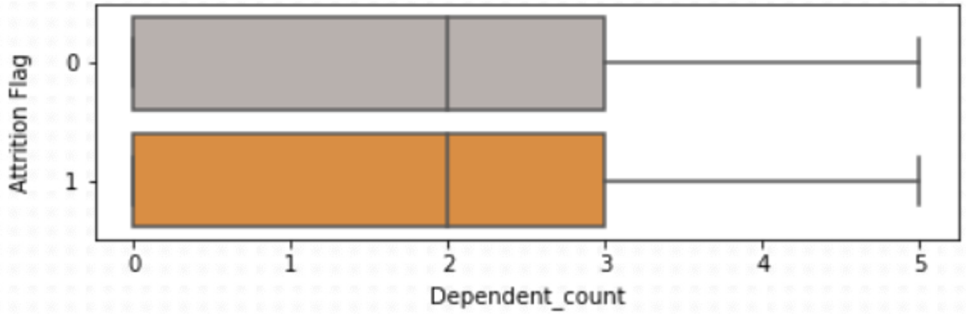


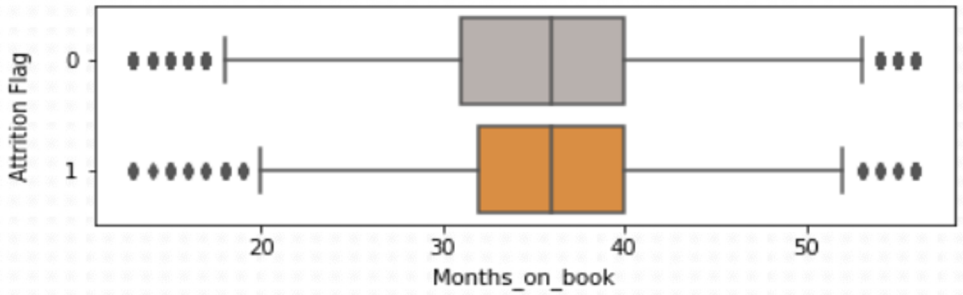
Percentage of Churned or Not by Contacts Count in 12 Months



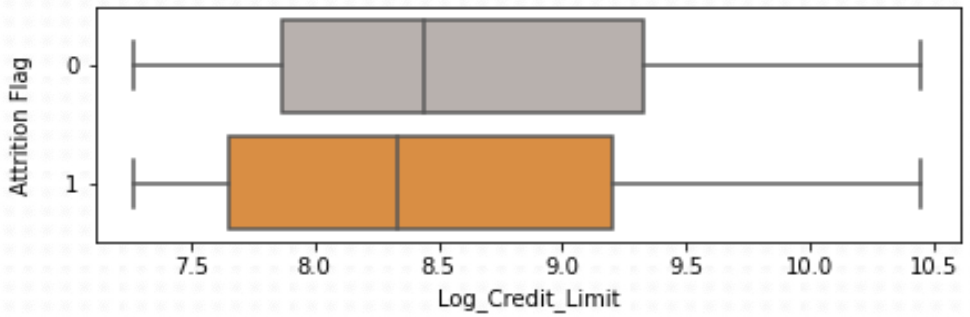
Boxplots of Age / Dependent Count / Months on Book Grouped by Churned or Not

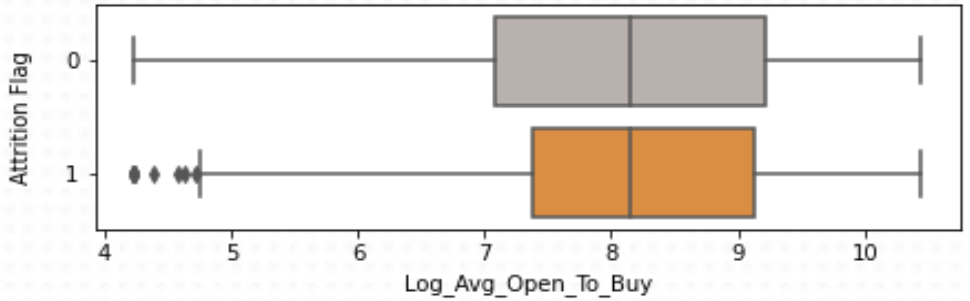


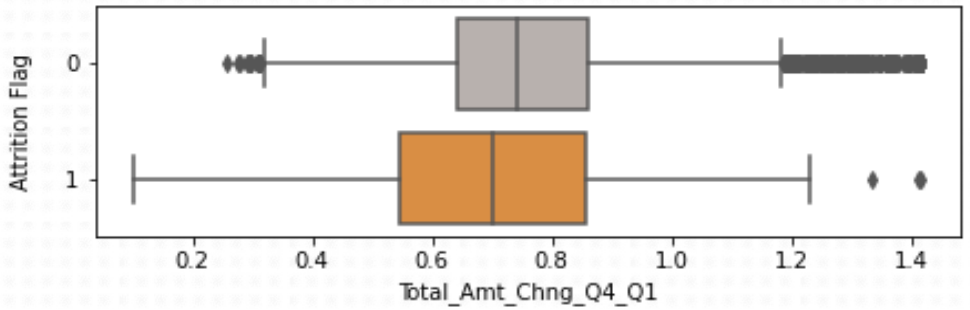




Boxplots of Log Credit Limit / Log Avg Open to Buy / Change in Total Amount Q4 - Q1 Grouped by Churned or Not





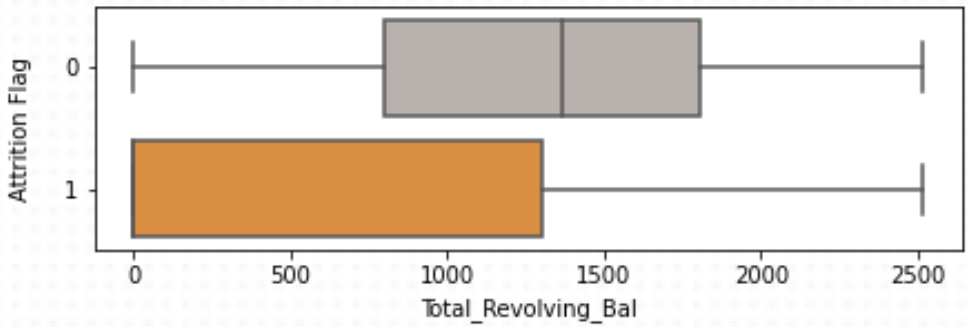


- Most numerical variables are not very discriminative for customers who churned or not.

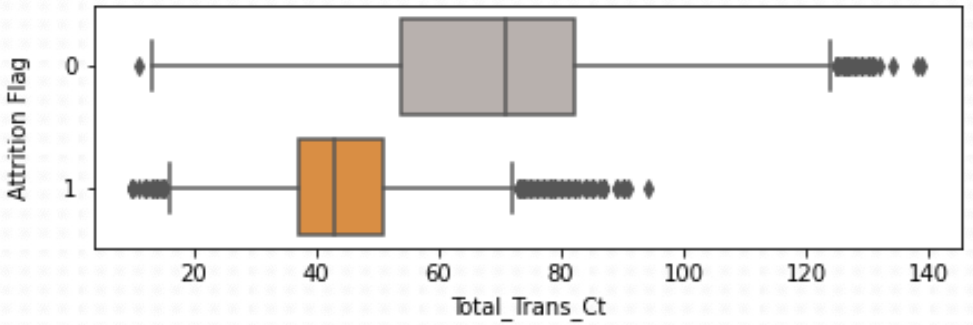
Exploratory Data Analysis

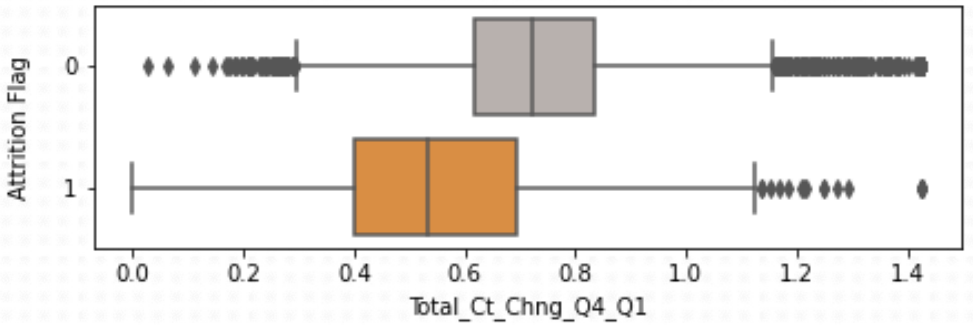
- The variables Total_Revolving_Bal, Total_Trans_Ct, Total_Ct_Chng_Q4-Q1, Log_Total_Trans_Amt can significantly distinguish whether the customers churned.
- The values for these variables are lower for churned customers than for non-churned customers.

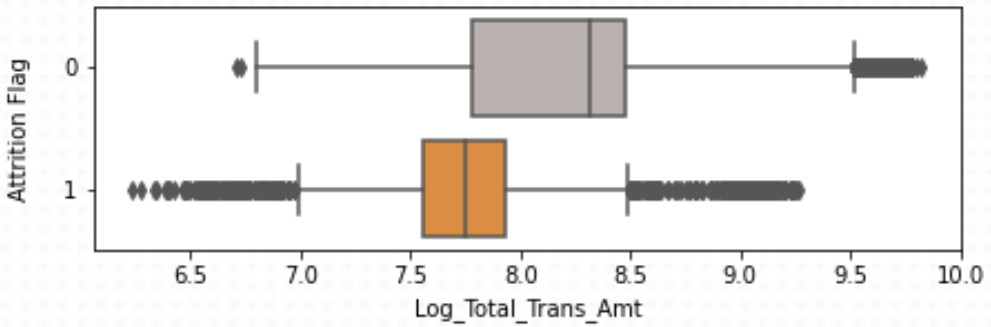
Boxplot of Total Revolving Balance Grouped by Churned or Not



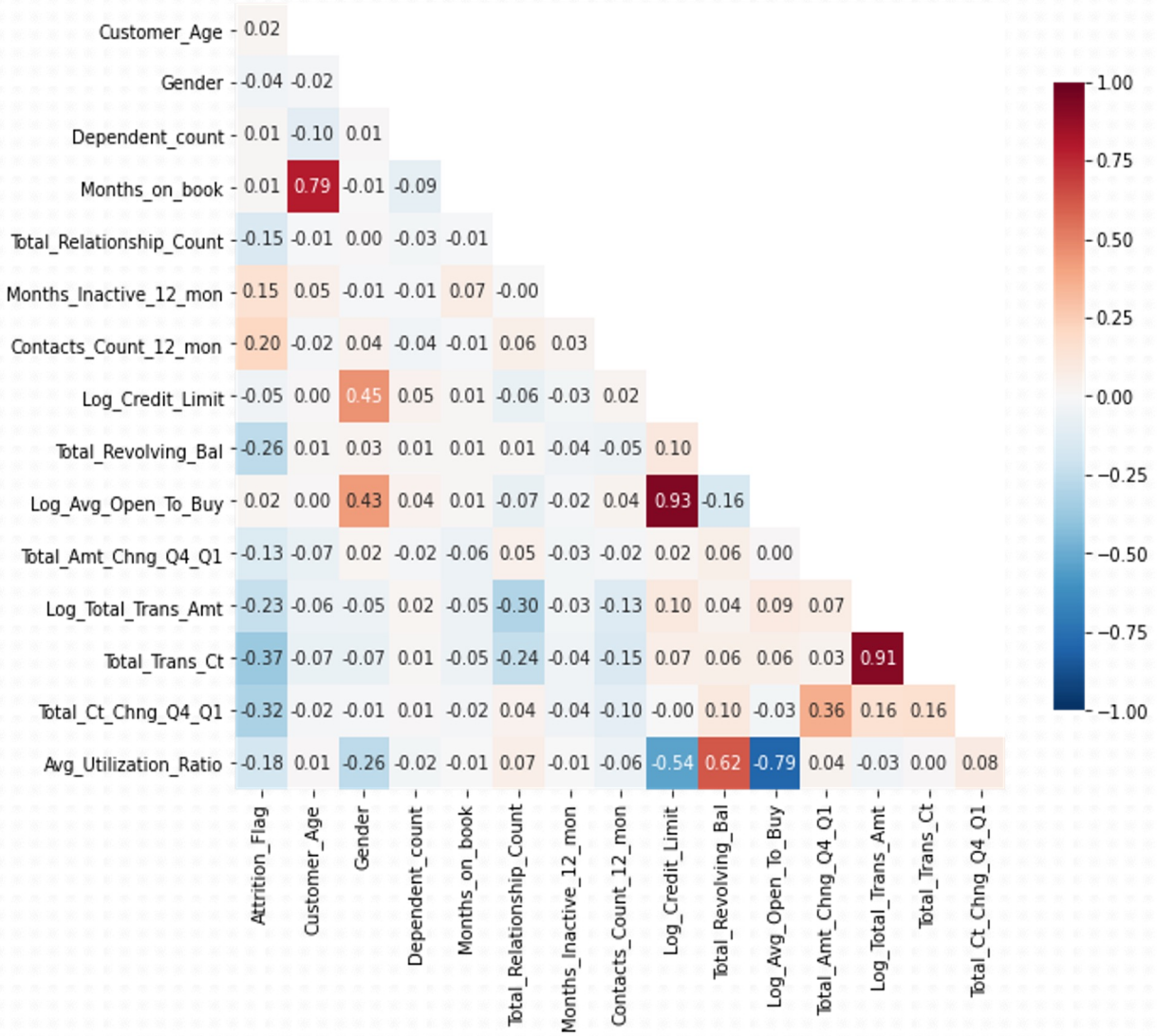
Boxplots of Total Transaction Count / Change in Total Transaction Count Q4 - Q1/ Log Total Transaction Amount Grouped by Churned or Not







Exploratory Data Analysis



- The target variable **Attrition_Flag** is slightly negatively correlated with **Total_Trans_Ct** and **Total_Ct_Chng_Q4_Q1** and slightly positively correlated with **Contacts_Count_12_mon**.
- The variable **Avg_Utilization_Ratio** exhibits strong positive or negative correlations with multiple variables.
- Potential multicollinearity risks need to be addressed.

Logistic Regression Model 1

Modeling:

- Used One Hot Encoding for all the categorical variables.
- Used 75% of the data as training set and 25% as test set.
- Used SMOTE to oversample the minority class so that the proportions of 0 and 1 are close.

Results:

- The customers with more total transactions have a lower probability of churn.
- The higher the transaction amount of the customer, the greater the likelihood of churn.

Logit Regression Results

Dep. Variable:	Attrition_Flag	No. Observations:	10837
Model:	Logit	Df Residuals:	10806
Method:	MLE	Df Model:	30
Date:	Mon, 21 Mar 2022	Pseudo R-squ.:	0.5935
Time:	21:48:34	Log-Likelihood:	-2984.5
converged:	True	LL-Null:	-7341.9
Covariance Type:	nonrobust	LLR p-value:	0.000

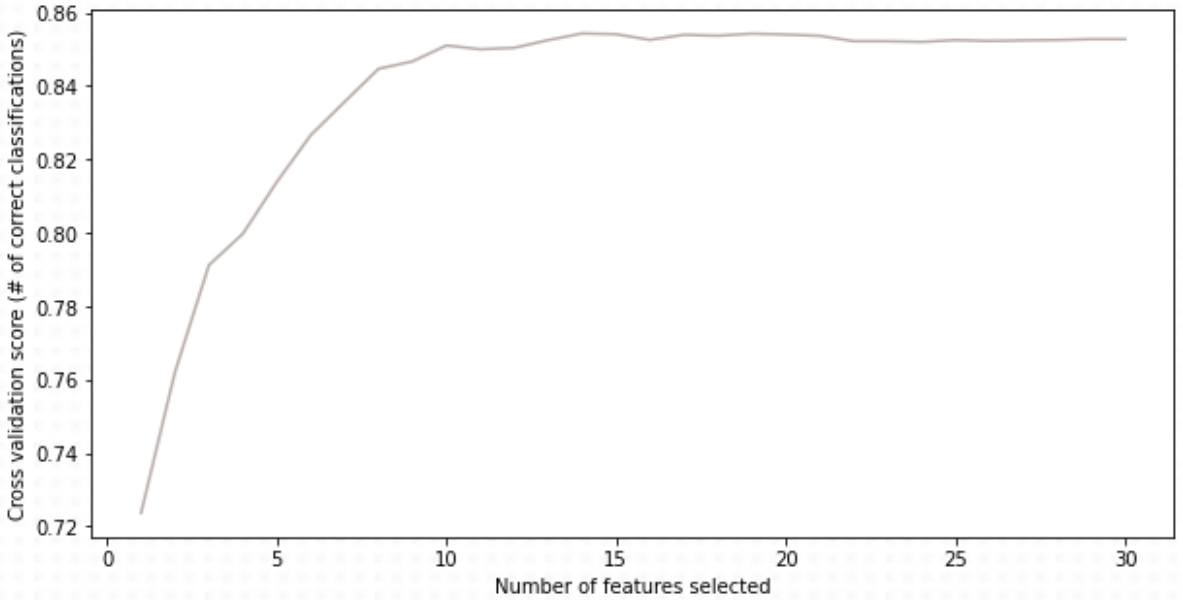
	coef	std err	z	P>|z|	[0.025	0.975]
const	11.4266	0.761	15.007	0.000	9.934	12.919
Customer_Age	0.4182	0.317	1.319	0.187	-0.203	1.040
Gender	-0.6553	0.131	-5.006	0.000	-0.912	-0.399
Dependent_count	1.3868	0.165	8.415	0.000	1.064	1.710
Months_on_book	-0.6117	0.310	-1.971	0.049	-1.220	-0.003
Total_Relationship_Count	-2.5384	0.122	-20.866	0.000	-2.777	-2.300
Months_Inactive_12_mon	3.2366	0.221	14.645	0.000	2.803	3.670
Contacts_Count_12_mon	3.3395	0.205	16.314	0.000	2.938	3.741
Log_Credit_Limit	7.2233	0.670	10.777	0.000	5.910	8.537
Total_Revolving_Bal	-1.4607	0.181	-8.052	0.000	-1.816	-1.105
Log_Avg_Open_To_Buy	-15.7103	1.325	-11.854	0.000	-18.308	-13.113
Total_Amt_Chng_Q4_Q1	-1.9641	0.250	-7.848	0.000	-2.455	-1.474
Log_Total_Trans_Amt	15.6781	0.514	30.514	0.000	14.671	16.685
Total_Trans_Ct	-25.5302	0.639	-39.925	0.000	-26.784	-24.277
Total_Ct_Chng_Q4_Q1	-4.6361	0.251	-18.447	0.000	-5.129	-4.144
Avg_Utilization_Ratio	-6.3636	0.547	-11.626	0.000	-7.436	-5.291
Education_Level_College	-0.3228	0.139	-2.329	0.020	-0.594	-0.051
Education_Level_Doctorate	-0.1446	0.184	-0.786	0.432	-0.505	0.216
Education_Level_Graduate	-0.2627	0.103	-2.548	0.011	-0.465	-0.061
Education_Level_High School	-0.1847	0.112	-1.647	0.100	-0.405	0.035
Education_Level_Post-Graduate	0.0854	0.178	0.479	0.632	-0.264	0.435
Education_Level_Uneducated	-0.2754	0.123	-2.243	0.025	-0.516	-0.035
Marital_Status_Divorced	-0.1605	0.188	-0.855	0.393	-0.528	0.208
Marital_Status_Married	-0.2536	0.130	-1.950	0.051	-0.508	0.001
Marital_Status_Single	0.8037	0.164	4.888	0.000	0.481	1.126
Income_Category_$120K +	0.7563	0.205	3.683	0.000	0.354	1.159
Income_Category_$40K - $60K	0.3129	0.147	2.132	0.033	0.025	0.601
Income_Category_$60K - $80K	0.2542	0.187	1.356	0.175	-0.113	0.622
Income_Category_$80K - $120K	0.5514	0.184	2.992	0.003	0.190	0.913
Income_Category_Less than $40K	0.3670	0.127	2.886	0.004	0.118	0.616
Card_Category_Blue	-0.2746	0.159	-1.725	0.084	-0.587	0.037

Logistic Regression Model 2

Improving:

- Performed feature selection using Recursive Feature Elimination with Cross Validation (RFECV).
- Checked the Variance inflation factor (VIF).
- 9 of the 30 variables were retained.

Results of Cross-Validation Recursive Feature Elimination



Models	Accuracy	F1-Score
Logistic Regression 1	88.59%	69.86%
Logistic Regression 2	87.01%	65.76%

Features	VIF
Total_Ct_Chng_Q4_Q1	9.96
Months_Inactive_12_mon	6.16
Total_Trans_Ct	6.16
Contacts_Count_12_mon	5.81
Dependent_count	5.22
Total_Relationship_Count	3.89
Marital_Status_Single	2.49
Total_Revolving_Bal	2.46
Gender	1.82

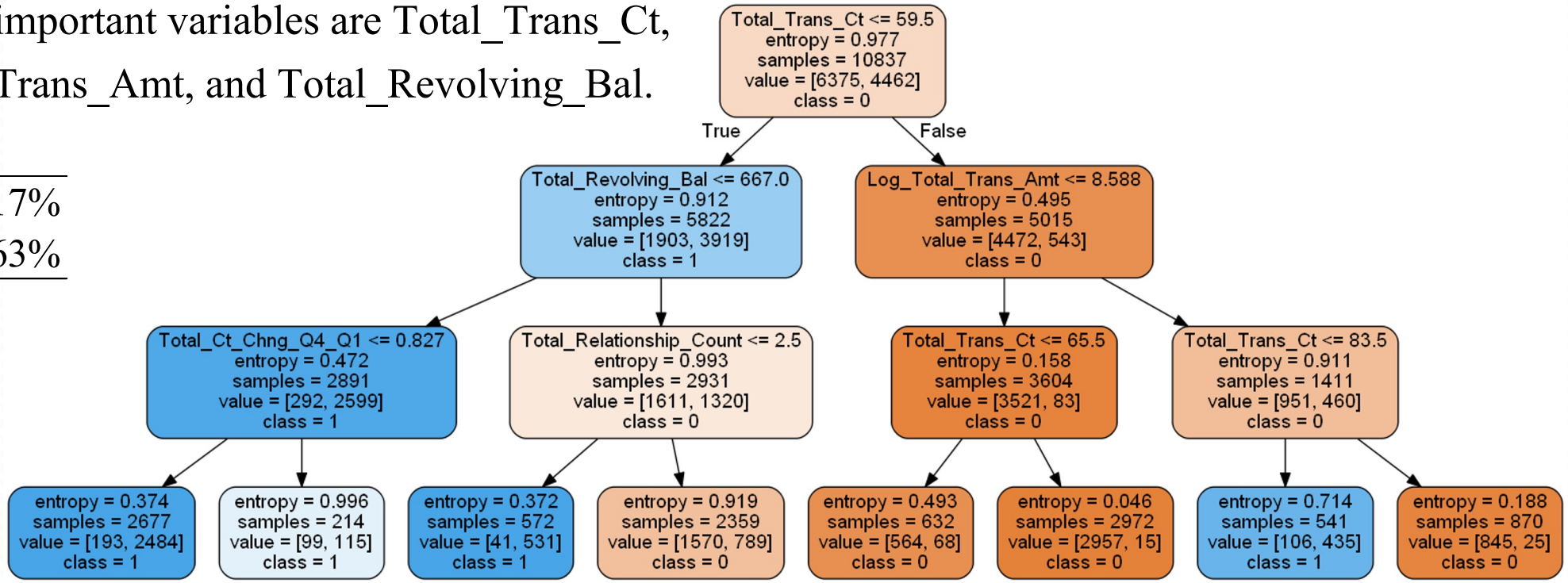
	coef	std err	z	P>|z|	[0.025	0.975]
const	4.0673	0.195	20.828	0.000	3.685	4.450
Gender	-0.4104	0.059	-6.927	0.000	-0.527	-0.294
Dependent_count	1.4359	0.143	10.007	0.000	1.155	1.717
Total_Relationship_Count	-2.7200	0.109	-25.016	0.000	-2.933	-2.507
Months_Inactive_12_mon	3.1631	0.193	16.383	0.000	2.785	3.541
Contacts_Count_12_mon	3.0122	0.179	16.853	0.000	2.662	3.362
Total_Revolving_Bal	-2.3957	0.089	-27.054	0.000	-2.569	-2.222
Total_Trans_Ct	-9.4799	0.238	-39.770	0.000	-9.947	-9.013
Total_Ct_Chng_Q4_Q1	-4.5538	0.211	-21.536	0.000	-4.968	-4.139
Marital_Status_Single	1.1281	0.107	10.583	0.000	0.919	1.337

Decision Tree Model

- The GridSearchCV method was used to obtain the optimal parameters of the decision tree model.
- Limited the depth of Decision Tree to 3 to see the most important features.
- The 3 most important variables are Total_Trans_Ct, Log_Total_Trans_Amt, and Total_Revolving_Bal.

Feature	Feature Importance by decision tree
Total_Trans_Ct	0.45483332
Log_Total_Trans_Amt	0.15290794
Total_Revolving_Bal	0.14767997
Total_Relationship_Count	0.08719741
Total_Ct_Chng_Q4_Q1	0.07200833

Accuracy	93.17%
F1-Score	80.63%



Random Forest Model

Modeling:

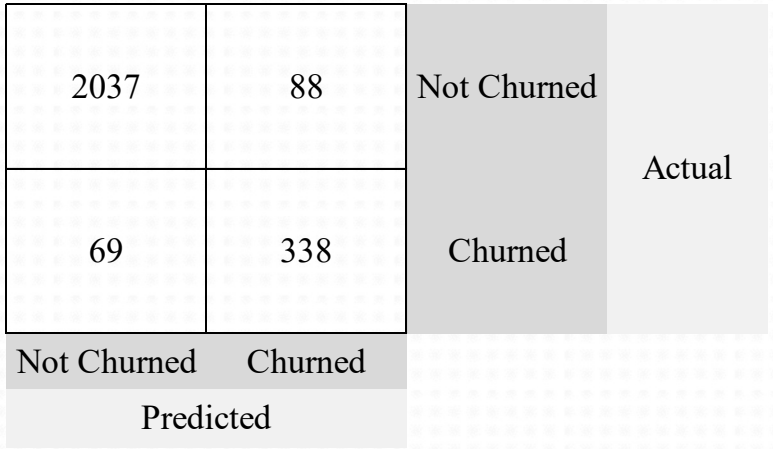
- The GridSearchCV method was also used to obtain the optimal parameters of the random forest model.

Result:

- The importance of the 3 variables Total_Trans_Ct, Log_Total_Trans_Amt, and Total_Revolving_Bal accounts for 54.11%.

Feature	Feature Importance by random foreset
Total_Trans_Ct	0.23653402
Log_Total_Trans_Amt	0.18307513
Total_Revolving_Bal	0.12146944
Total_Ct_Chng_Q4_Q1	0.09273619
Avg_Utilization_Ratio	0.06458751
Total_Relationship_Count	0.06185749
Total_Amt_Chng_Q4_Q1	0.04160009
Gender	0.02744243
Log_Avg_Open_To_Buy	0.02103119
Log_Credit_Limit	0.01979562
Months_Inactive_12_mon	0.01876189

Model	Accuracy	F1- Score
Logistic Regression 1	88.59%	69.86%
Logistic Regression 2	87.01%	65.76%
Decision Tree	93.17%	80.63%
Random Forest	93.80%	81.15%



Conclusion

- Three most important features:
  - ✓ The total transactions count
  - ✓ The total transaction amount
  - ✓ The total revolving balance
- The lower the total transactions count and the total revolving balance, the more likely customers will churn.
- The higher the total transaction amount, the more likely customers will churn.
- Using the random forest model to predict whether a user will churn can reduce the loss of customer churn and be more cost-effective (with the highest F1-Score).

Recommendation

- Using these variables as indicators for daily monitoring of user churn. Once a customer's three values reach a certain threshold, the credit card manager should contact the customer to resolve the problems.

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.

Thank you!