CREDIT CARD CHURN ANALYSIS

Nitish Baradwaj, Nuo Chen, Yueru Zhang, Jianing Liu

ISE 535 Fall 2022 Group-3

University of Southern California

Link to our recording:
https://drive.google.com/file/d/1YoYTnwaqbMqzLezh5a_muOUa1M00OFFd/view?usp=share_link

Overview

Credit Card Customer Churn Analysis

Overview

Business Objectives

- Credit cards are a good source of income for banks, so the bank wants to analyze the data and id who wilentify the customers **leave or stay on their credit card services and reason for leaving**
- Thus the bank could improve upon those areas

Credit Card Customer Churn Analysis

Data Dictionary

Category Variable (6)
Attrition_Flag
Gender
Education_level
Marital_Status
Income_Category
Card_Category

OUTCOME

Numeric Variable (15)		
CLIENTNUM	Avg_Open_To_Buy	Credit_Limit
Customer_Age	Total_Amt_Chng_Q4_Q1	Total_Revolving_Bal
Months_on_book	Total_Trans_Amt	Dependent_count
Total_Relationship_Count	Total_Trans_Ct	
Months_Inactive_12_mon	Total_Ct_Chng_Q4_Q1	
Contacts_Count_12_mon	Avg_Utilization_Ratio	

10,127 observations divided into 21 attributes

Initial Data Review

Numeric

Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
CLIENTNUM	0	10127	7.391776e+08	708082083.0	8.283431e+08	3.690378e+07
Customer_Age	0	45	4.632596e+01	26.0	7.300000e+01	8.016814e+00
Dependent_count	0	6	2.346203e+00	0.0	5.000000e+00	1.298908e+00
Months_on_book	0	44	3.592841e+01	13.0	5.600000e+01	7.986416e+00
Total_Relationship_Count	0	6	3.812580e+00	1.0	6.000000e+00	1.554408e+00
Months_Inactive_12_mon	0	7	2.341167e+00	0.0	6.000000e+00	1.010622e+00
Contacts_Count_12_mon	0	7	2.455317e+00	0.0	6.000000e+00	1.106225e+00
Credit_Limit	0	6205	8.631954e+03	1438.3	3.451600e+04	9.088777e+03
Total_Revolving_Bal	0	1974	1.162814e+03	0.0	2.517000e+03	8.149873e+02
Avg_Open_To_Buy	0	6813	7.469140e+03	3.0	3.451600e+04	9.090685e+03

Attribute <chr>	Missing Values <int>	Unique Values <int>	Mean <dbl>	Min <dbl>	Max <dbl>	SD <dbl>
Total_Amt_Chng_Q4_Q1	0	1158	7.599407e-01	0.0	3.397000e+00	2.192068e-01
Total_Trans_Amt	0	5033	4.404086e+03	510.0	1.848400e+04	3.397129e+03
Total_Trans_Ct	0	126	6.485869e+01	10.0	1.390000e+02	2.347257e+01
Total_Ct_Chng_Q4_Q1	0	830	7.122224e-01	0.0	3.714000e+00	2.380861e-01
Avg_Utilization_Ratio	0	964	2.748936e-01	0.0	9.990000e-01	2.756915e-01

Factor (Category)

Attribute <chr>	Missing Values <int>	Unique Values <int>
Attrition_Flag	0	2
Gender	0	2
Education_Level	0	7
Marital_Status	0	4
Income_Category	0	6
Card_Category	0	4

Logical Groupings of Variables

Category (6)

- **Outcome**
  - Attrition_Flag
- **Customer information**
  - Gender, Education_Level, Marital_Status, Income_Category
- **Card information**
  - Card_Category

Measures (14)

- **Customer information**
  - Customer_Age, Dependent_count
- **The relationship between bank and customer**
  - Months_on_book, Total_Relationship_Count, Months_Inactive_12_mon, Contacts_Count_12_mon
- **The credit card information**
  - Credit_Limit, Total_Revolving_Bal, Avg_Open_To_Buy, Total_Amt_Chng_Q4_Q1,Total_Ct_Chng_Q4_Q1, Total_Trans_Amt, Total_Trans_Ct, Avg_Utilization_Ratio

Univariate & Bivariate Analysis

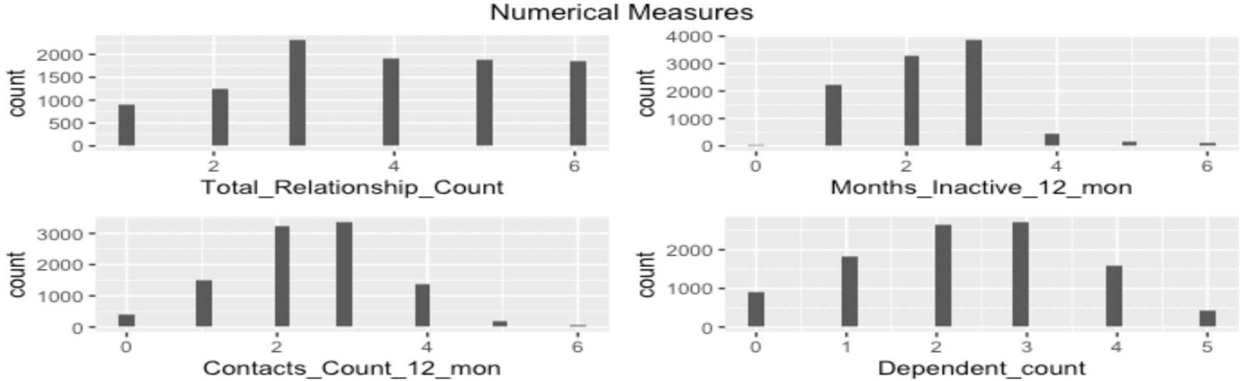
Univariate Analysis - Numerical Summary

High cardinality variable

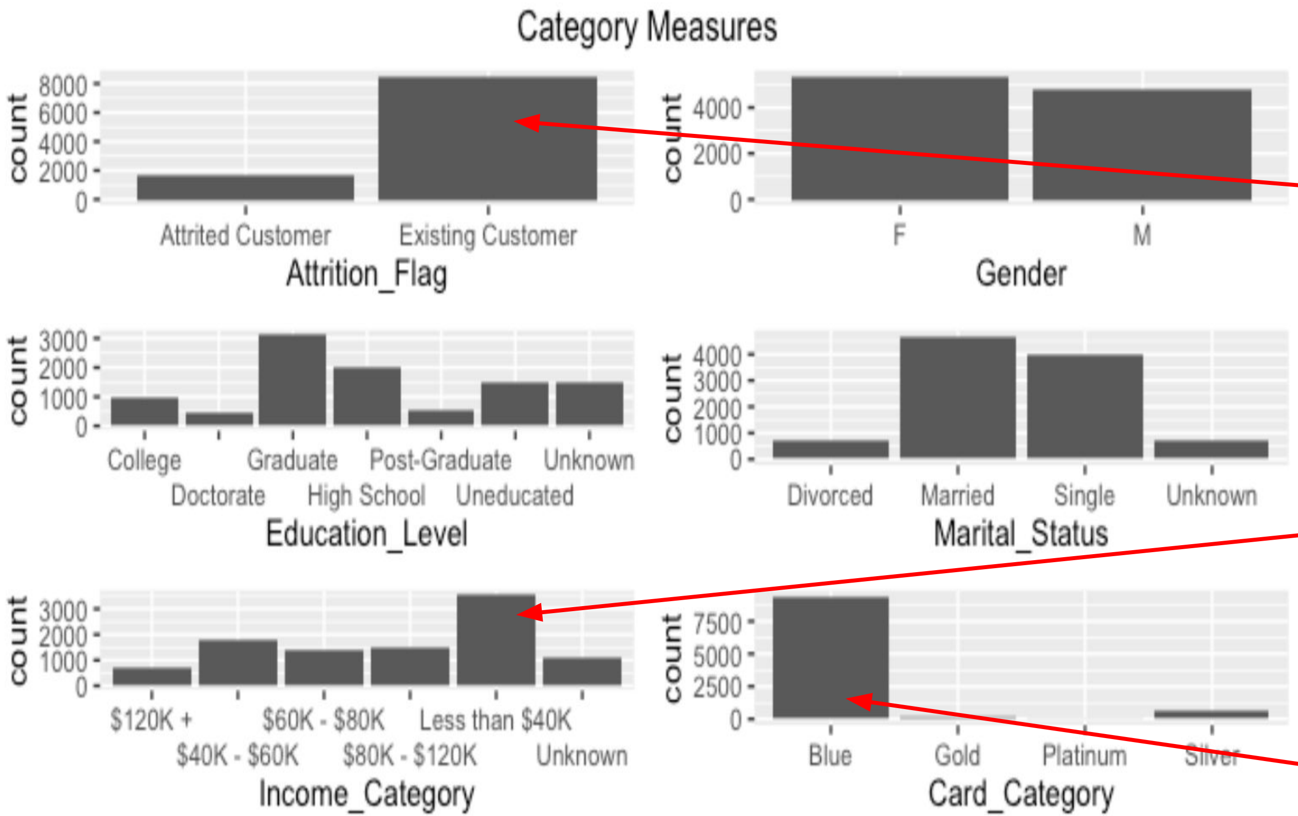
Credit limit is **right skewed**. Most customers have less than 5,000 credit limit. Same as credit limit.

Three clusters of Total_trans_Amt
Two clusters of total_Trans_Ct

Low cardinality variable



Univariate Analysis - Category Summary

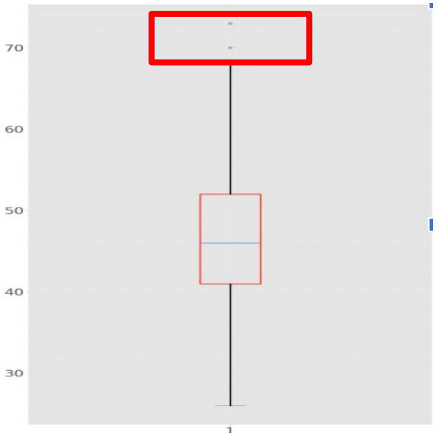


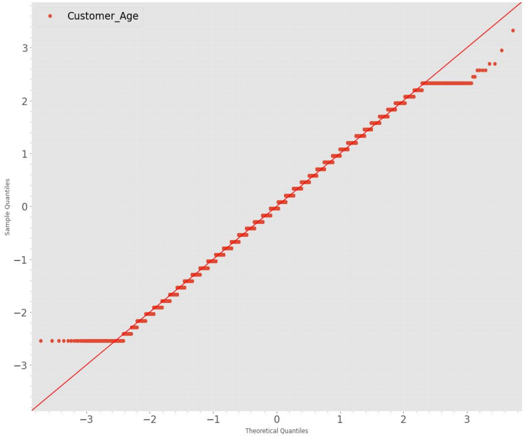
Attrited Customer: About 84% of the customers are existing (active) customers.

Income: 35% customers earn less than $40K;

Card Category: 93% customers own Blue Card.

Univariate Outlier Analysis - Customer Age





Attrition_Flag <chr>	Credit_Limit <dbl>	Customer_Age <dbl>	Gender <chr>	Education_Level <chr>	Income_Category <chr>	Card_Category <chr>
Existing Customer	4469.0	73	M	High School	$40K - $60K	Blue
Existing Customer	3252.0	70	M	High School	Less than $40K	Blue
Existing Customer	13860.0	68	M	Graduate	Unknown	Blue
Attrited Customer	1438.3	68	M	High School	Less than $40K	Blue
Existing Customer	3006.0	67	F	Graduate	Less than $40K	Blue
Existing Customer	5876.0	67	M	Graduate	$40K - $60K	Blue
Existing Customer	3106.0	67	M	Uneducated	$40K - $60K	Blue
Existing Customer	10509.0	67	F	Unknown	Unknown	Blue
Attrited Customer	7882.0	66	F	Doctorate	Unknown	Blue
Existing Customer	3171.0	66	F	High School	Less than $40K	Blue

1-10 of 111 rows | 1-7 of 9 columns Previous 1 2 3 4 5 6 ... 12 Next

Customer Age:

- Data is normally distributed.
- The average age of customers is around 46, most customers are younger than 50.
- Two outliers at the upper bound side. **Remove those.**

Univariate Outlier Analysis - Total_Ct_Q4_Q1 and others

Total_Ct_Chng_Q4_Q1	`n()`
<dbl>	<int>
3.71	1
3.57	1
3.5	1
3.25	1
3	2
2.88	1
2.75	1
2.57	1
2.5	3
2.43	1

… with 819 more rows

Total_Amt_Chng_Q4_Q1	`n()`
<dbl>	<int>
3.40	1
3.36	1
2.68	1
2.59	1
2.37	1
2.36	1
2.32	1
2.28	1
2.28	1
2.27	1

… with 1,145 more rows

Attrition_Flag <fctr>	Total_Ct_Chng_Q4_Q1 <dbl>	Total_Amt_Chng_Q4_Q1 <dbl>
Attrited Customer	0.000	0.153
Attrited Customer	0.000	0.000
Attrited Customer	0.000	0.000
Attrited Customer	0.000	0.000
Attrited Customer	0.000	0.010
Attrited Customer	0.000	0.000
Attrited Customer	0.000	0.000
Existing Customer	0.028	0.459
Attrited Customer	0.029	0.046
Attrited Customer	0.038	1.214

the rows that Total_Ct_Chng_Q4_Q1 is 0.00, and Attrition_Flag is **Attrited_Flag**. **Keep it!**

Months_Inactive_12_mon	`n()`
<dbl>	<int>
6	124
5	178
4	434
3	3841
2	3281
1	2233
0	29

Contacts_Count_12_mon	`n()`
<dbl>	<int>
6	54
5	175
4	1389
3	3379
2	3226
1	1498
0	399

Avg_Open_To_Buy	`n()`
<dbl>	<int>
34516	98
34362	1
34300	1
34297	1
34286	1
34238	1
34227	1
34140	1
34119	1
34117	1

… with 6,798 more rows

Total_Trans_Amt	`n()`
<dbl>	<int>
18484	1
17995	1
17744	1
17634	1
17628	1
17498	1
17437	1
17390	1
17350	1
17258	1

… with 5,022 more rows

Customer_Age	`n()`
<dbl>	<int>
73	1
70	1
68	2
67	4
66	2
65	101
64	43
63	65
62	93
61	93

… with 35 more rows

Credit_Limit	`n()`
<dbl>	<int>
34516	507
34496	1
34458	1
34427	1
34198	1
34173	1
34162	1
34140	1
34058	1
34010	1

… with 6,192 more rows

Total_Trans_Ct	`n()`
<dbl>	<int>
139	1
138	1
134	1
132	1
131	6
130	5
129	6
128	10
127	12
126	10

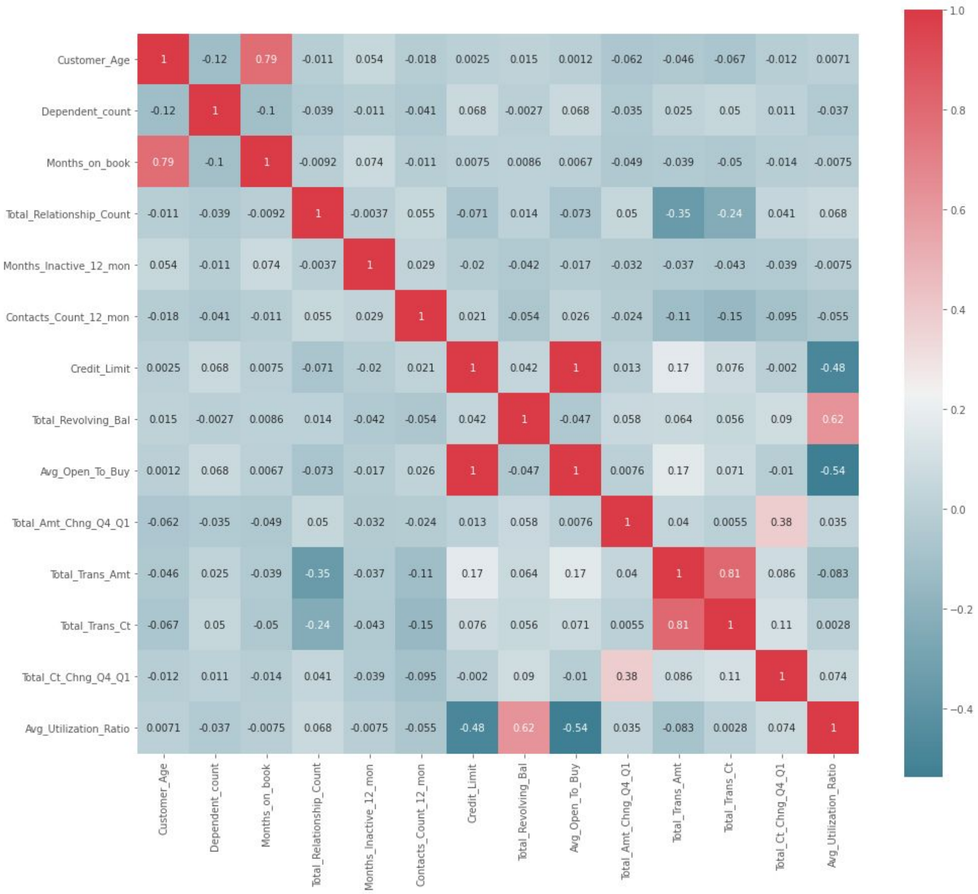
… with 116 more rows

- **None** of these outliers appears out of the range of being possible, so we will leave them in the dataset.

Univariate Analysis Summary

- The blue card owner is 93%
- Most customers inactive for 2 or 3 months in the last 12 months
- Credit Limit and Average Open to Buy are all right-skewed
- Total revolving balance contains lots of 0 and the distribution of both credit limit and average open to buy are right-skewed.
- Total transaction amount has 3 parts, total transaction count has 2 parts, we need to do further research.
- Most customers did less transactions in Q4 compared to Q1, also spent less total amount of money. The removing outlier are all attributed customer

Bivariate Analysis

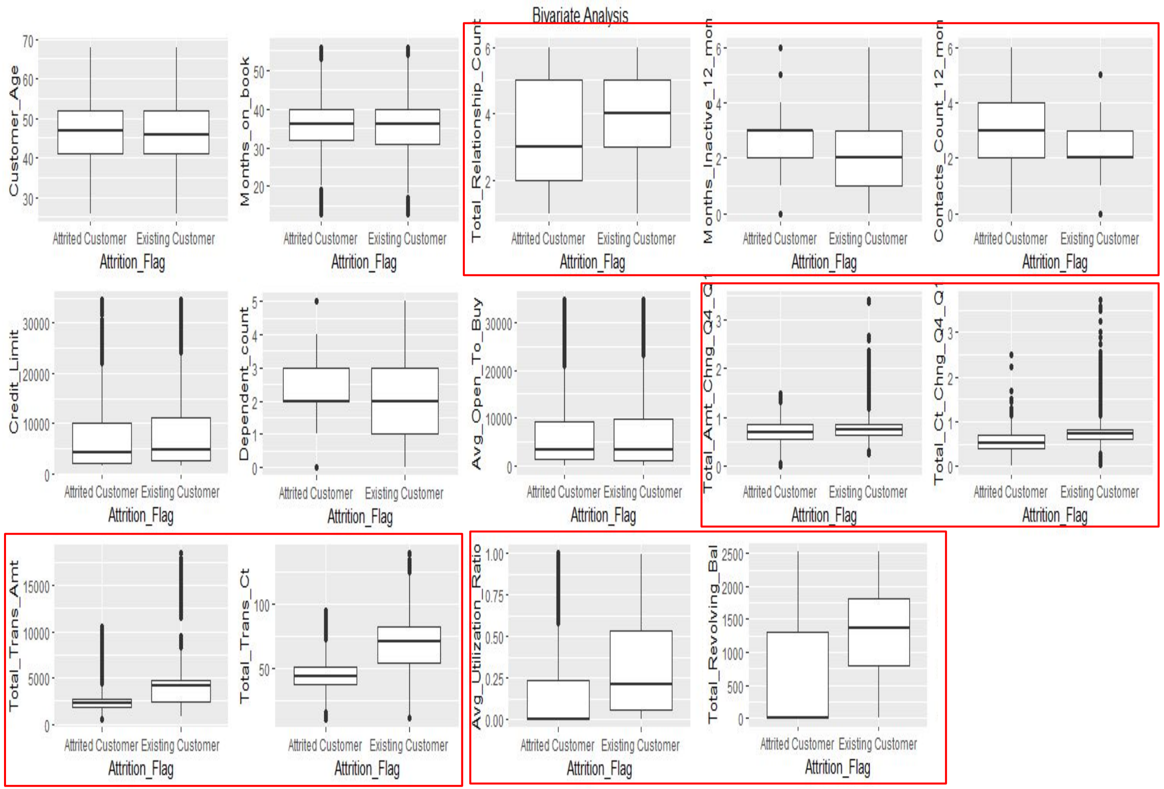


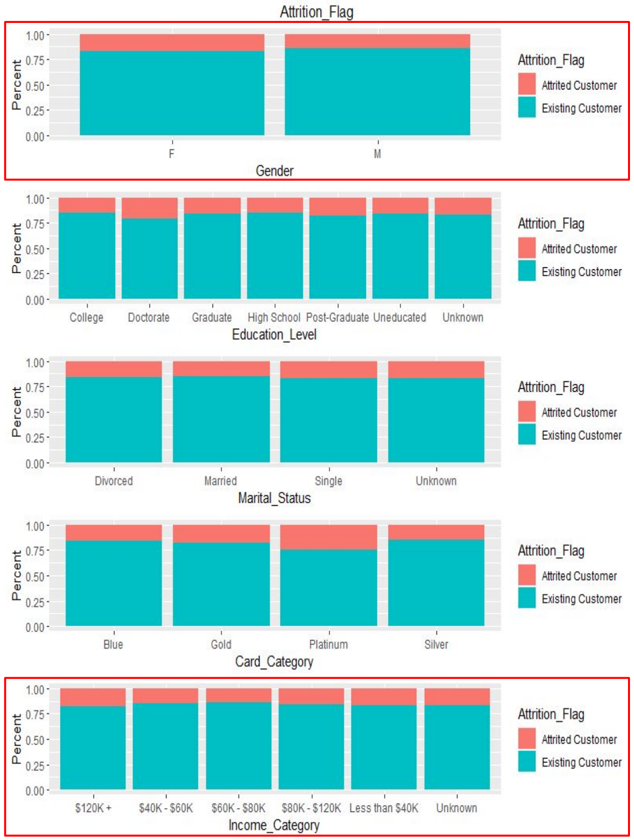
- Avg_Open_To_Buy and Credit_Limit have 100% collinearity
- Months_on_book and Customer_Age, Total_Trans_Ct and Total_Trans_Amt have quite strong correlation
- Total_Revolving_Bal and Avg_Utilization_Ratio also have positive correlation

$$Avg\_Utilization\_Ratio = \frac{Total\_Revolving\_Bal}{Credit\_Limit}$$

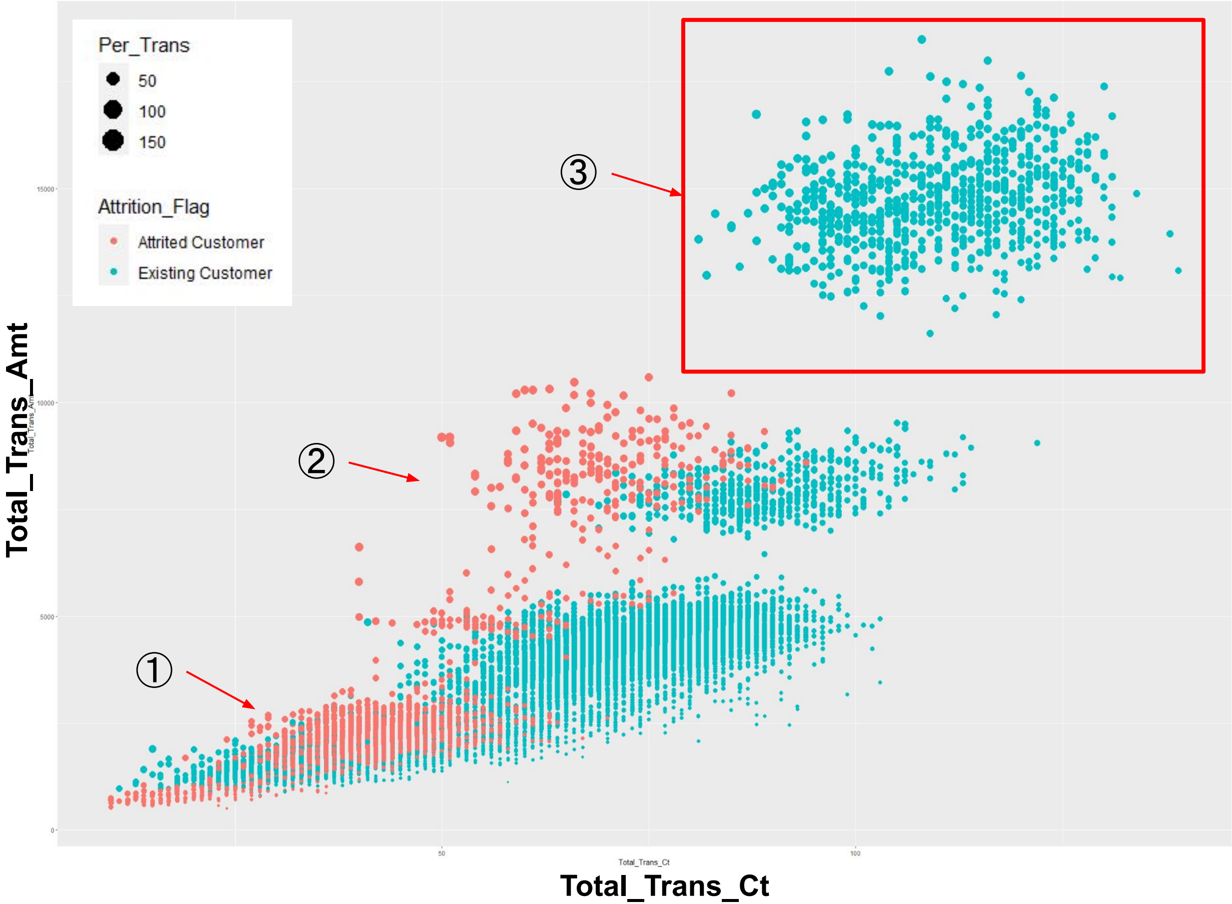
$$Total\_Revolving\_Bal = Credit\_Limit - Avg\_Open\_To\_Buy$$

Bivariate Analysis - Attrition Flag

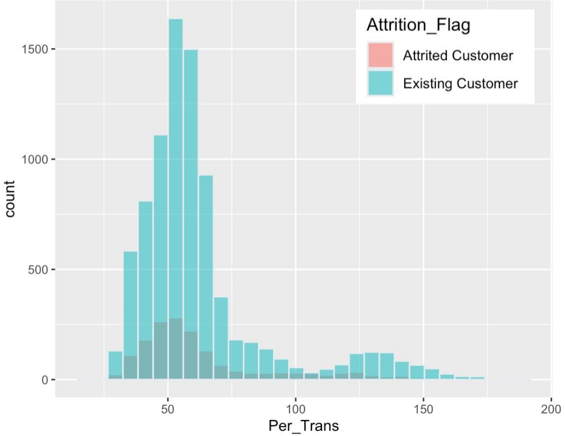




Bivariate Analysis



Average Per Trans	Existing Customer	Attrited Customer
①	53.31	54.61
②	88.85	**121.97**
③	134.32	**NA**



Bivariate Analysis Summary

- Avg_Open_To_Buy and Credit_Limit, Total_Trans_Amount and Total_Trans_Ct have strong correlation.
- **Existing customers** have **higher transaction count**, **transaction amount**, and **amount per count** (more total revolving balance and average utilization). They are more likely to connect with the bank.
- Male customers have higher credit limit than female customers. But female customers have higher average utilization ratio.
- Attrition rates are significantly different for customers of different genders and income categories.

Models

Outline

The Various Models we will be building are:

1. Logistic Regression
2. Decision Trees

- **The reason why we choose the following two models are just because of their probabilistic nature.**
- **The Question that we want to answer is, Given the details of the customer (Data) can we predict with some probability if the customer stays or leaves.**

Generation of Training and Test Dataset

- We will be partitioning the dataset into two. We will use one to train the model and measure in-sample accuracy and the other as a test dataset to measure the out of sample accuracy.
- It is very important to make sure the training data and the test data come from the same distribution. We will randomly sample the data into training and testing data. We have used the 80-20 partition. 80% as training data and the 20 % as test data.

~~~~
```{r}
m = nrow(churn_df)

set.seed(693)
train_ids <- sample(m, 0.8 * m)
```

```{r}
train_ids
```

```{r}
train_df <- churn_df[train_ids,]
test_df <- churn_df[-train_ids,]
```
~~~~

Here we have used the last three digits of our USC id as a random number seed. This makes it unique

Logistic Regression

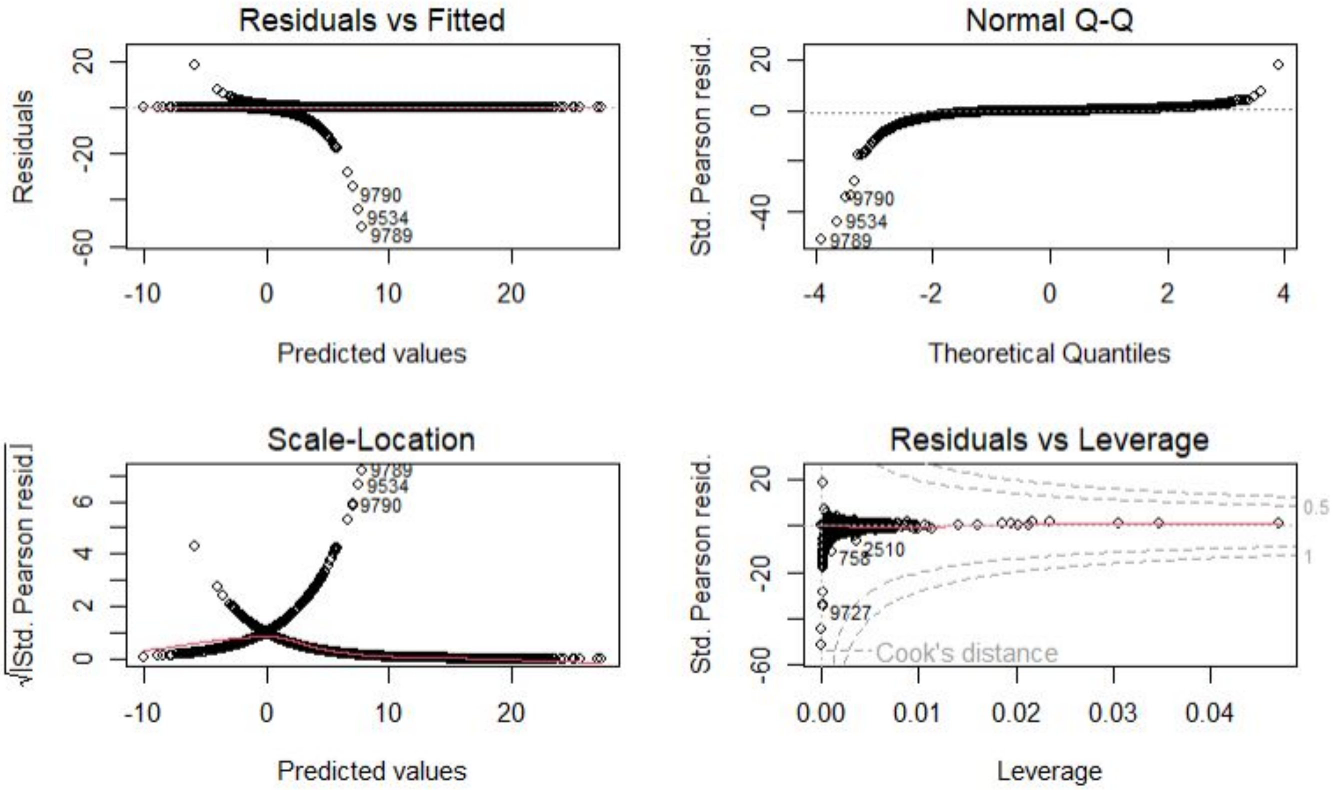
```
Call:
glm(formula = Attrition_Flag ~ Total_Trans_Ct + Total_Trans_Amt +
    Contacts_Count_12_mon + Total_Relationship_Count + Months_Inactive_12_mon +
    Total_Ct_Chng_Q4_Q1 + Total_Trans_Ct:Total_Trans_Amt, family = binomial,
    data = train_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7693   0.0313   0.1394   0.3649   2.5919

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -9.173e-01  2.954e-01  -3.105   0.0019 **
Total_Trans_Ct                  8.700e-02  4.736e-03  18.369   <2e-16 ***
Total_Trans_Amt                -2.927e-03  1.354e-04 -21.616   <2e-16 ***
Contacts_Count_12_mon          -4.370e-01  3.961e-02 -11.033   <2e-16 ***
Total_Relationship_Count        5.046e-01  3.018e-02  16.718   <2e-16 ***
Months_Inactive_12_mon         -4.726e-01  4.297e-02 -10.998   <2e-16 ***
Total_Ct_Chng_Q4_Q1             3.573e+00  2.080e-01  17.179   <2e-16 ***
Total_Trans_Ct:Total_Trans_Amt  2.759e-05  1.574e-06  17.529   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7083.1  on 8100  degrees of freedom
Residual deviance: 3738.7  on 8093  degrees of freedom
AIC: 3754.7

Number of Fisher Scoring iterations: 8
```

Why choose Logistic Regression?

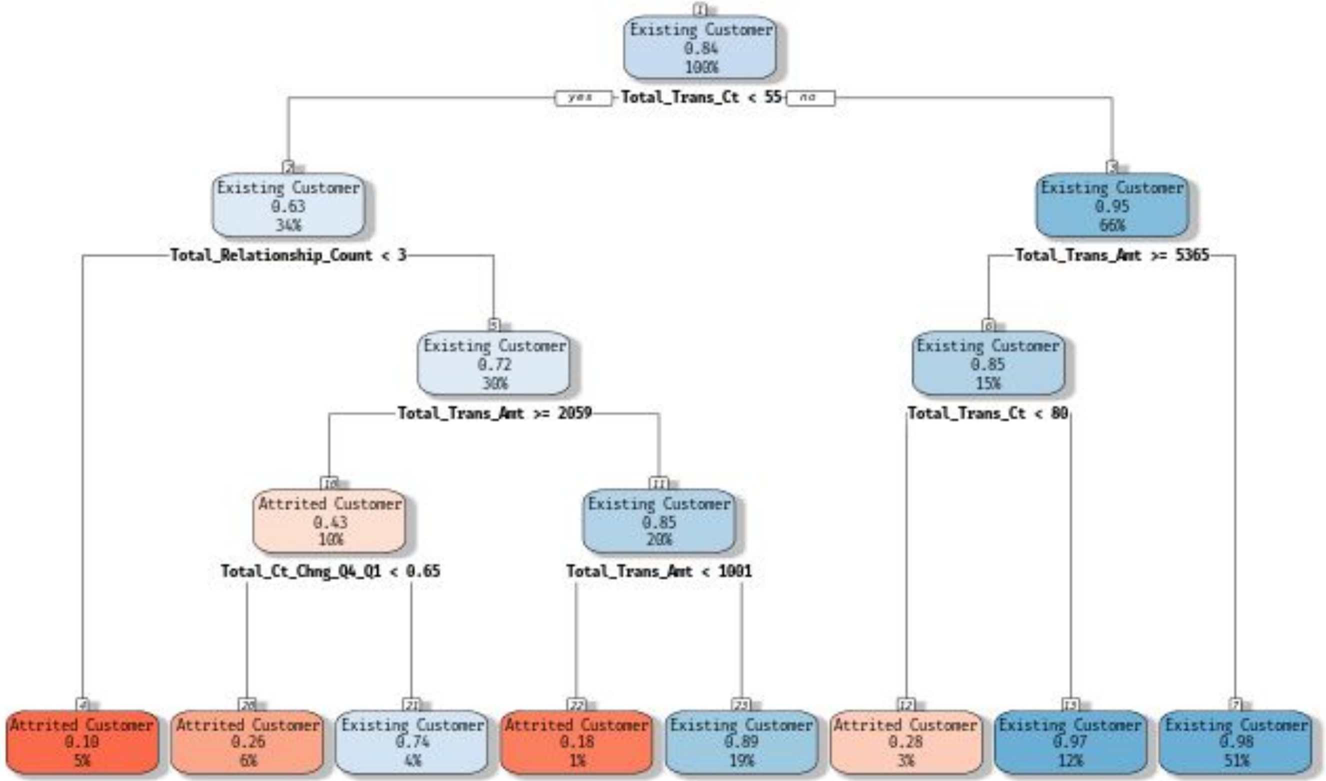
Our Ultimate aim is to provide the bank with a prediction of which customer will stay and which one will leave based on the data given to us. Since the logistic regression predicts a probability distribution. This model is ideal for predicting with a certain confidence of which customers will stay.

Total Transaction count and Total Transaction Amount have the lowest p value and are therefore the most important feature to keep in mind while building the model

Logistic Regression

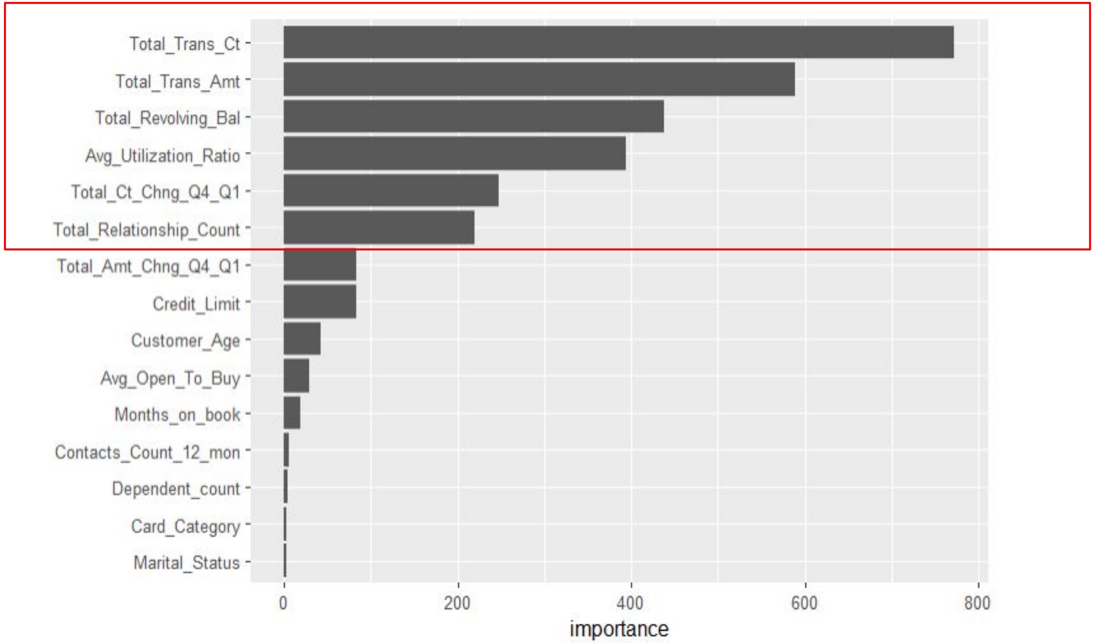


Decision Tree



Decision Tree

```
      Total_Trans_Ct           Total_Trans_Amt       Total_Revolving_Bal     Avg_Utilization_Ratio       Total_Ct_Chng_Q4_Q1 Total_Relationship_Count
          772.291032                588.636118                438.724452                393.403610                247.906216               220.243796
 Total_Amt_Chng_Q4_Q1              Credit_Limit              Customer_Age           Avg_Open_To_Buy            Months_on_book      Contacts_Count_12_mon
           82.943242                 82.893151                 41.954620                 29.502282                 19.029682                 5.277538
     Dependent_count             Card_Category            Marital_Status
            4.712738                  3.079317                  2.465411
```



Comparing the Models

Something important to understand about probabilistic models:

- Output is a probability
- The Training data does not explicitly contain any probability Data(x,y) is binary in y which is generated by a hidden target function that we don't have access to.
- p(y|x) = {f(x) if y=1 or
  1-f(x) if y = 0} here f(x) is the hidden target function.

- A good error measure is what we call the 'cross-entropy' error or the log loss error. Since we are dealing with probabilities it is good to choose a likelihood type error.

Comparing the Models

Logistic Regression

Log-Loss Error:

```{r}
prob <- predict(fit.logit,test_df,type = "response")
LogLoss(y_pred = prob, y_true = test_df$flag)
```

```
[1] 0.2296575
```

Confusion Matrix:

```
                   Predicted
Actual              Attrited Customer Existing Customer
  Attrited Customer               204               139
  Existing Customer                28              1655
```

Accuracy:

```
[1] 0.9175716
```

Decision Tree

Log-Loss Error:

```{r}
LogLoss(y_pred = tree_prob, y_true = test_df$flag)
```

```
[1] 2.506055
```

Confusion Matrix:

```
                   Predited
Actual              Attrited Customer Existing Customer
  Attrited Customer               259                84
  Existing Customer                63              1620
```

Accuracy:

```
[1] 0.9274432
```

Conclusions

- The most important feature to keep in mind when building a model is the Total Transaction count and the Total Transaction Amount. This is verified by both the models that we built.
- Although the Decision Tree does well in the Accuracy metric calculated from the confusion matrix, It does poorly in the log-loss error metric.
- For probabilistic models such as logistic regression and decision trees that deal with an inherent hidden probability function the log-loss error metric is a better reflection of the model accuracy.
- We recommend the Bank to use a logistic regression model with emphasis on the features of Total Transaction Amount and Total Transaction Count to determine with a confidence interval for a given customer, the probability of staying or leaving

THANK YOU FOR YOUR TIME