CREDIT CARD CUSTOMER PREDICTION

Sritha Darbha - 11520315
Yamini Mangarai - 11535594

Prof. Dr. Onkar Malgonde

Programming Languages for Business Analytics

Table of Contents

Executive Summary	03
Project Motivation/ Background	04
Key Questions	05
Data Source	06
Data Description	06
Data Transformation	13
Data Analysis	15
Model Analysis	23
Findings	26
Conclusions	27
Appendix	28
References	33

Executive Summary

A manager at the bank is concerned with more and more customers leaving their credit card services. They would really appreciate if anyone could predict for them who is going to leave so they can proactively go to the customer to provide them better services and turn customer's decisions in the opposite direction.

In this dataset, we have almost 10,000 customers mentioning their age, salary, marital status, credit card limit, credit card category, etc.

Our goal is to analyze what features affect the attrition of the customers and how can the bank improve their customer service.

Project motivation/background

Everyone who handles their finances uses credit/debit cards on a day-to-day basis. Credit cards give you a certain amount of credit/limit with which you can borrow money from a bank, with the understanding that you will pay it back. Credit cards can help you establish credit and pay for unexpected expenses, but they can also put you in debt if you charge more than you can afford to pay back on time. There are so many benefits personally in using credit cards like receiving discounts, cash back, rewards, building credit score, etc. Credit cards are safer to carry than cash. With cards, consumers have recourse for fraudulent transactions and the card regulates dishonest business practices.

Even in a bigger picture, there are so many benefits in using cards. Cards enable more funds to flow through the economy. Change generated during a cash transaction potentially removes a certain amount of money from circulation. The change sitting unused hinders economic growth. Credit cards supported by global payment networks enable growth in key economic sectors including e-commerce and travel and tourism. Increasingly, business is transacted online and in a diverse global economy, the ability to buy and sell online has been essential for many businesses and consumers.

Predicting the Customer churn of the credit cards would help the authorities make better decisions regarding introducing new services to encourage new customers to the bank or even improve their current services. By predicting the churn banks can target on the right audience. Using credit card helps consumers to have a secure access to their funds, while reducing usage of cash and check handling for merchants. It reduces need to prospect for new customers, which allows banks to focus on building the relationship with existing customers. Usage of credit card encourages the manufacture of more goods and commitment of capital, creates tax revenues. The benefits are having lend money from banks with some interest rates when needed and can be paid off on monthly basis. Maintaining a credit score helps having more options to lend money with less interest rates, home and car buying.

We wanted to work on this dataset to utilize our python skills we learned in this class and apply it to this kind of dataset which has a mixture of categorical, nominal, binary, numerical etc.

Key Questions

Quantitative:

1. What is the effect of Credit Limit on Income Category?
2. How many of the Attrited Customers are inactive for more than 3 months in the last 12 months?
3. Is there any change in Maximum and Minimum Transaction Amount overall with Transaction Amount Change Rate greater than 1?
4. How is the Total Transaction Count is changing with Transaction Count Change Rate?
5. Which type of Credit card has the highest Attrition?

Descriptive:

6. Is there any relation between Dependent Count and Open to buy Credit Line Average?
7. What is the relation between Revolving Balance and Card Utilization Ratio?
8. How is the Age range for Attrited Customers?
9. How is the Customer Income affecting the Attrition Rate?
10. What gender is Attriting the most?

Data Source

The data is taken from the Kaggle Website below:

https://www.kaggle.com/datasets/whenamancodes/credit-card-customers-prediction

Data Description

The dataset has a total of 21 columns:

Column	Description
CLIENTNUM	Unique identifier for the customer holding the account
Attrition_Flag	If the customer is Existing or Attrited
Customer_Age	Customer's Age in Years
Gender	M=Male, F=Female
Dependent_count	Number of dependents
Education_Level	Educational Qualification of the account holder
Marital_Status	Married, Single, Divorced, Unknown
Income_Category	Annual Income Category of the account holder
Card_Category	Type of Card
Monthsonbook	Period of relationship with bank
TotalRelationshipcount	Total no. of products held by the customer
MonthsInactive12_mon	No. of months inactive in the last 12 months
ContactsCount12_mon	No. of Contacts in the last 12 months

Credit_Limit	Credit Limit on the Credit Card
TotalRevolvingBal	Total Revolving Balance on the Credit Card
AvgOpenTo_Buy	Open to Buy Credit Line (Average of last 12 months)
TotalAmtChngQ4Q1	Change in Transaction Amount (Q4 over Q1)
TotalTransAmt	Total Transaction Amount (Last 12 months)
TotalTransCt	Total Transaction Count (Last 12 months)
TotalCtChngQ4Q1	Change in Transaction Count (Q4 over Q1)
AvgUtilizationRatio	Average Card Utilization Ratio

Dataset Information:

```
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   CLIENTNUM                 10127 non-null  int64
 1   Attrition_Flag            10127 non-null  object
 2   Customer_Age              10127 non-null  int64
 3   Gender                    10127 non-null  object
 4   Dependent_count           10127 non-null  int64
 5   Education_Level           10127 non-null  object
 6   Marital_Status            10127 non-null  object
 7   Income_Category           10127 non-null  object
 8   Card_Category             10127 non-null  object
 9   Months_on_book            10127 non-null  int64
 10  Total_Relationship_Count  10127 non-null  int64
 11  Months_Inactive_12_mon    10127 non-null  int64
 12  Contacts_Count_12_mon     10127 non-null  int64
 13  Credit_Limit              10127 non-null  float64
 14  Total_Revolving_Bal       10127 non-null  int64
 15  Avg_Open_To_Buy           10127 non-null  float64
 16  Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 17  Total_Trans_Amt           10127 non-null  int64
 18  Total_Trans_Ct            10127 non-null  int64
 19  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 20  Avg_Utilization_Ratio     10127 non-null  float64
dtypes: float64(5), int64(10), object(6)
memory usage: 1.6+ MB
```

There are no null values in the dataset.

Dataset Description:

```
       Customer_Age  Dependent_count  Months_on_book  \
count  10127.000000     10127.000000    10127.000000
mean      46.325960         2.346203       35.928409
std        8.016814         1.298908        7.986416
min       26.000000         0.000000       13.000000
25%       41.000000         1.000000       31.000000
50%       46.000000         2.000000       36.000000
75%       52.000000         3.000000       40.000000
max       73.000000         5.000000       56.000000

       Total_Relationship_Count  Months_Inactive_12_mon  \
count              10127.000000            10127.000000
mean                   3.812580                2.341167
std                    1.554408                1.010622
min                    1.000000                0.000000
25%                    3.000000                2.000000
50%                    4.000000                2.000000
75%                    5.000000                3.000000
max                    6.000000                6.000000

       Contacts_Count_12_mon  Credit_Limit  Total_Revolving_Bal  \
count           10127.000000  10127.000000         10127.000000
mean                2.455317   8631.953698          1162.814061
std                 1.106225   9088.776650           814.987335
min                 0.000000   1438.300000             0.000000
25%                 2.000000   2555.000000           359.000000
50%                 2.000000   4549.000000          1276.000000
75%                 3.000000  11067.500000          1784.000000
max                 6.000000  34516.000000          2517.000000
```

	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	\
count	10127.000000	10127.000000	10127.000000	10127.000000	
mean	7469.139637	0.759941	4404.086304	64.858695	
std	9090.685324	0.219207	3397.129254	23.472570	
min	3.000000	0.000000	510.000000	10.000000	
25%	1324.500000	0.631000	2155.500000	45.000000	
50%	3474.000000	0.736000	3899.000000	67.000000	
75%	9859.000000	0.859000	4741.000000	81.000000	
max	34516.000000	3.397000	18484.000000	139.000000	

	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio
count	10127.000000	10127.000000
mean	0.712222	0.274894
std	0.238086	0.275691
min	0.000000	0.000000
25%	0.582000	0.023000
50%	0.702000	0.176000
75%	0.818000	0.503000
max	3.714000	0.999000

Customer_Age: Customer age has the mean value of 10127.00 and the standard deviation value equal to 46.32. It can be determined that the age of customers is mostly close to 52 years. Minimum age of customers is 26 and maximum 73.

Dependent_count: Dependent count has mean value of 2.34 and standard deviation value is 1.29. The highest dependent count is close to 3. The dependent count of minimum is 0 and maximum is 5.

Months_on_book: Customers month on book has mean value of 35.92 and the standard deviation value equals 7.98. The months on book high are around close to 40 months and the minimum and maximum ranges from 13 to 56.

Total_Relationship_Count: The mean value of total relationship count is 3.81 and the standard deviation value equals 1.55. Relationship count mostly in high are close to 5. The minimum and maximum total relationship count ranges from 1 to 6.

Months_Inactive_12_mon: The mean value of customers inactive for 12 months is a value of 2.34 and the standard deviation value equal to 1.01. Mostly the highest amount of months the customers were inactive is close to 3. The minimum and maximum months of inactive by customer in 12 months shows from 0 to 6.

Contacts_count_12_mon: The mean value of the number of contacts in the last 12 months is equal to 2.45 and the standard deviation value equal to 1.10. The high number of contacts is close to 3. The range of minimum and maximum contacts are from 0 to 6.

Credit_Limit: Mean value of credit limit equal to 8631.95 $ and the value of standard deviation equal to 9088.77 $. Mostly the highest credit limit amount is close to 11067.50. The minimum and maximum credit limit ranges from 1438.30 to 34516.

Total_Revolving_Bal: Mean value of total revolving balance on credit card is equal to 1162.81 and the standard deviation value is 814.98. Mostly high revolving balance is close to 1784. The minimum and maximum balance ranges from 0 to 2517.

Avg_Open_To_Buy: Mean value of open to buy credit line is equal to 7469.13 and the value of standard deviation is equal to 9090.68. Mostly the highest amount to buy credit line is close to 9859. The minimum and maximum amount of credit line open to buy ranges from 3 to 34516.

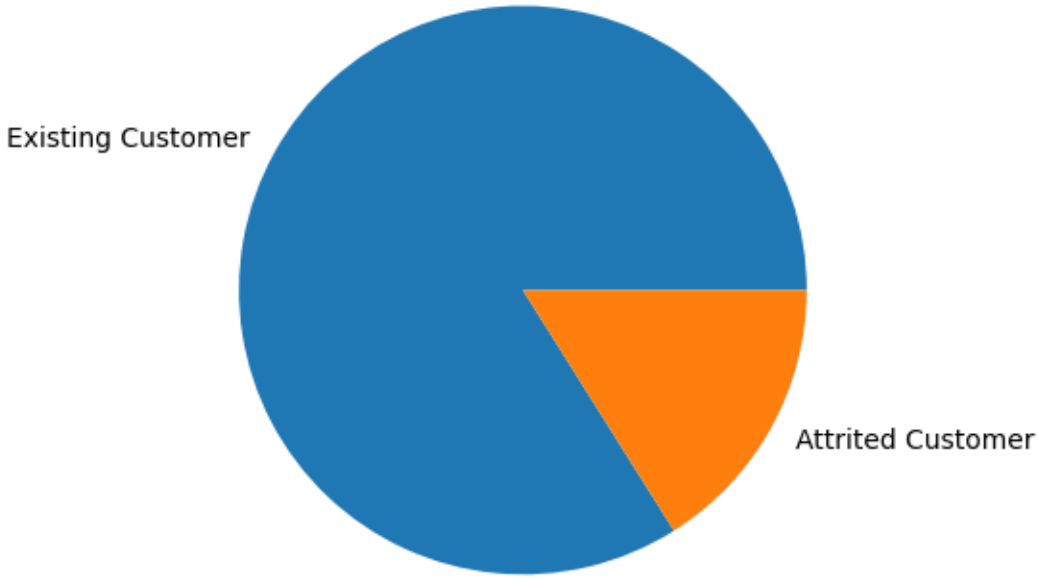
Total_Amt_Chng_Q4_Q1: Change in transaction amount Q4 over Q1, mean value is equal to 0.7599 and value of standard deviation is 0.2192. Mostly the high in change in transaction amount is close to 0.8590. The minimum and maximum transaction amount change ranges from 0 to 3.39.

Total_Trans_Amt: The mean value of total transaction amount in last 12 months is equal to 4404.08 $ and the value of standard deviation is equal to 3379.12 $. Mostly the high amount of total transactions are close to 4741 $. The minimum and maximum total amount of transactions ranges from 0 to 3.39.

Total_Trans_Ct: Total transaction count from past 12 months mean value is equal to 64.85 and the value of standard deviation is equal to 23.47. Transaction count is closely high to 81. The minimum and maximum transaction count ranges from 10 to 139.

Total_Ct_Chng_Q4_Q1: Change in transaction count of Q4 over Q1, mean value is equal to 0.71 and value of standard deviation is equal to 0.23. Mostly the high change in transaction count is close to 0.81. The minimum and maximum change in transaction ranges from 0 to 3.71.

Avg_Utilization_Ratio: The Average card utilization mean value is equal to 0.27 and standard deviation is equal to 0.27 where both mean and standard are similar. The highest rate of utilization mostly is close to 0.50. The minimum and maximum utilization ranges from 0 to 0.99.



This is a Pie chart showing the Attrition Rate to the Existing Customer Distribution.

Data Transformation

Data Transformation 1:

```
# cleaning 1 - Renaming the Last two columns and dropping them
ccdata.rename(columns={'Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1': 'PYes'}, inplace=True)
ccdata.rename(columns={'Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2': 'PNo'}, inplace=True)
ccdata = ccdata.drop(['PYes', 'PNo'], axis=1)
```

Data Transformation 2:

```
# cleaning 2 - Replacing the Attrition Flag Column to Binary values
ccdata['Attrition_Flag'].replace(['Attrited Customer', 'Existing Customer'], [0, 1], inplace=True)
```

Data Transformation 3:

```
# cleaning 3 - Replacing the Gender column to Binary values
ccdata['Gender'].replace(['M', 'F'], [0, 1], inplace=True)
```

Data Transformation 4:

```
# cleaning 4 - One hot encoding - for Education Level column
ccdata = pd.get_dummies(ccdata, columns=['Education_Level'], prefix=['Education-Level'])
ccdata = ccdata.reindex(columns=ccdata.columns, fill_value= 0)
```

Data Transformation 5:

```
# cleaning 5 - One hot encoding - for Marital Status Column
ccdata = pd.get_dummies(ccdata, columns=['Marital_Status'], prefix=['Marital_Status'])
ccdata = ccdata.reindex(columns=ccdata.columns, fill_value= 0)
```

Data Transformation 6:

```
# cleaning 6 - Ordinal Encoding - For Income category to see if the income level can have an affect in the credit score
# Dropping unknown income
print(ccdata.Income_Category.unique())
mapping = {'Unknown': 0, 'Less than $40K': 1, '$40K - $60K': 2, '$60K - $80K': 3, '$80K - $120K': 4, '$120K +': 5}
ccdata = ccdata.replace(mapping)
```

```
ccdataincome = ccdata[ccdata['Income_Category'] == 0].index
ccdata.drop(ccdataincome, inplace=True)
```

Data Transformation 7:

```
# cleaning 7 - One hot encoding - for Card category
ccdata = pd.get_dummies(ccdata, columns=['Card_Category'],
prefix=['Card_Category'])
ccdata = ccdata.reindex(columns=ccdata.columns, fill_value=0)
```

Data Transformation 8:

```
# drop columns as client number is unique to organization
ccdata = ccdata.drop(['CLIENTNUM'], axis=1)
```

Exploratory Data Analysis – Answering the Key Questions

Quantitative:

1. What is the effect of Credit Limit on Income Category?

```
Credit Limit Vs Income_Category
1    15987.0
2    23981.0
3    34516.0
4    34516.0
5    34516.0
Name: Credit_Limit, dtype: float64
```

```
With Unknown Income Range - Credit Limit Vs Income_Category
0    34516.0
1    15987.0
2    23981.0
3    34516.0
4    34516.0
5    34516.0
```

From the Analysis, we see the Income Categories 3, 4, 5, which means Income Category with Income above $60K have the highest credit limit of $34,516.0. We have checked the same without dropping the 'Unknown' Column of the Income category and see that the Income Category with Unknown Income also has the credit limit of $34,516.0. From this, we can conclude that the Income category, with income less than $60K, influences the Credit Limit.

2. How many of the Attrited Customers are inactive for more than 3 months in the last 12 months?

```
Number of Attrited customers Inactive for more than 3 months in the last 12 months: 155
```

Out of total Attrited Customers of 1440, only 155 of them are inactive for more than 3 months in the last 12 months. So, majority of the customers were active in the last 3 months before attrition. This shows it doesn't affect the active/inactive of the user with user attrition.

3. Is there any change in Maximum and Minimum Transaction Amount overall with Transaction Amount Change Rate greater than 1?

```
The minimum Total transaction amount overall is : 510
The minimum Total transaction amount for people with Transaction Amount Change Rate greater than 1 is : 602
The maximum Total transaction amount overall is : 18484
The maximum Total transaction amount for people with Transaction Amount Change Rate greater than 1 is : 16692
```

From the stats, we can see that minimum Transaction Amount with Transaction Amount Change rate greater than 1 is more than overall Transaction Amount, which shows that customers who tend to have a higher transaction amount change from q4 to q1 are the more likely ones to increase their transaction amount. While, on the other hand, the maximum transaction amount is higher than ones with transaction amount change greater than 1, which may indicate, customers with higher transaction amount change rate, probably had transactions which were low compared to overall amount.

4. How is the Total Transaction Count is changing with Transaction Count Change Rate?

```
The minimum Total transaction Count overall is : 10
The maximum Total transaction Count overall is : 139
The minimum Total transaction Count for people with Transaction Count Change Rate greater than or equal to 1 is : 10
The maximum Total transaction Count for people with Transaction Count Change Rate greater than or equal to 1 is : 118
The minimum Total transaction Count for people with Transaction Count Change Rate less than 1 is : 10
The maximum Total transaction Count for people with Transaction Count Change Rate less than 1 is : 139
```

We can see here that minimum transaction count remains the same with the transaction count change rate over the year, it probably means that customers transaction count is always at a minimum of 10. The maximum transaction count is higher for people with transaction count change rate of less than 1, means some customers probably had the same count of transactions over the year.

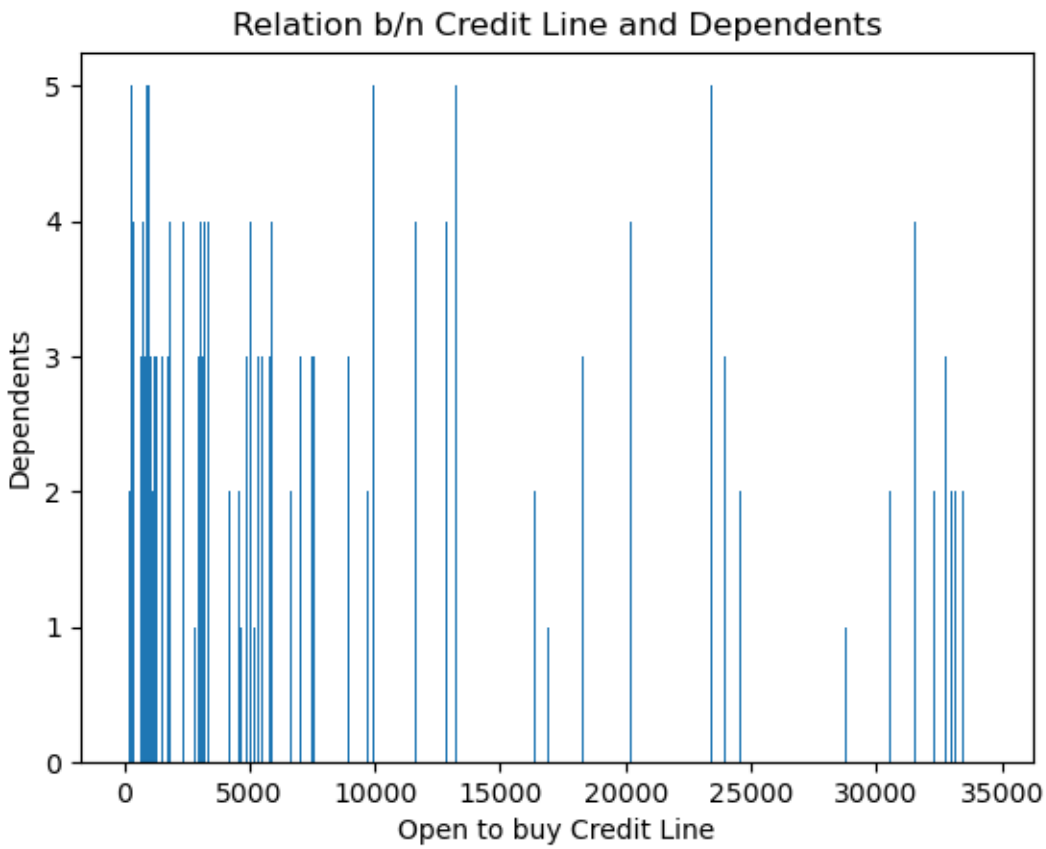
5. Which type of Credit card has the highest Attrition?

```
For Blue Card: Attrition_Flag
0    1343
1    7048
Name: Card_Category_Blue, dtype: uint64
For Gold Card: Attrition_Flag
0    19
1    88
Name: Card_Category_Gold, dtype: uint8
For Platinum Card: Attrition_Flag
0     3
1    12
Name: Card_Category_Platinum, dtype: uint8
For Silver Card: Attrition_Flag
0     75
1    427
Name: Card_Category_Silver, dtype: uint64
```

We see the Blue Card, which is more common among the consumers, have the highest Attrition compared to the rest of the cards. On the other hand, the Platinum card, although has the least customers, has the highest Attrition Rate of 25%, compared to the other cards.

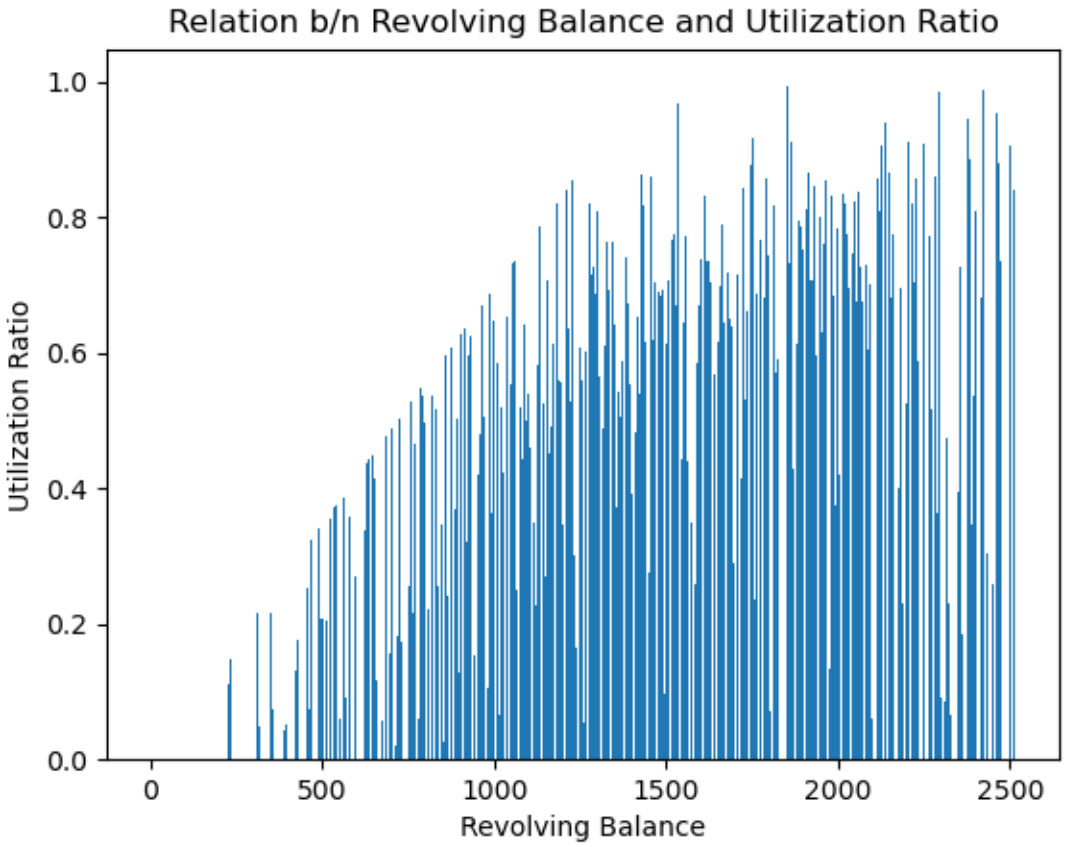
Descriptive:

6. Is there any relation between Dependent Count and Open to buy Credit Line Average?



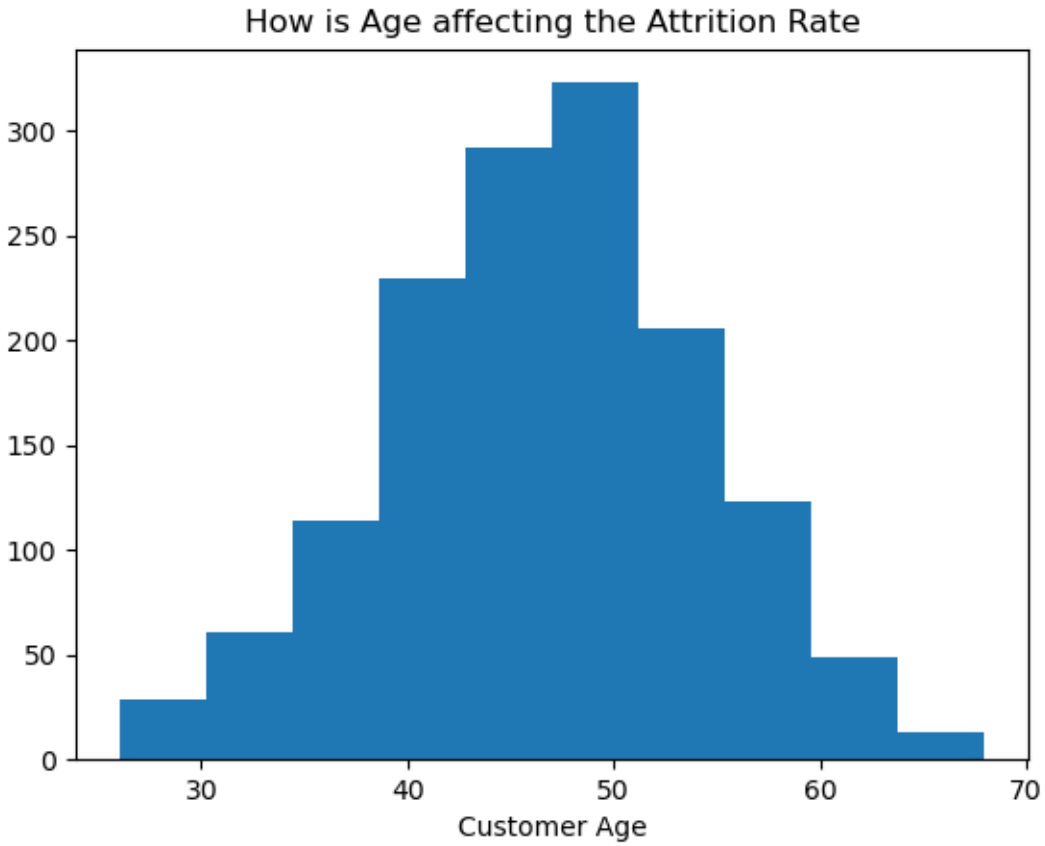
From the bar graph, we can see that the average Open to buy Credit Line is lower for of the people with a Dependent count of 3. Open to buy credit line is the difference between the credit limit and the present credit balance. This shows that, customers, with 3 or less dependents, have lower balance difference and are more likely to pay their bills on time.

7. What is the relation between Revolving Balance and Card Utilization Ratio?



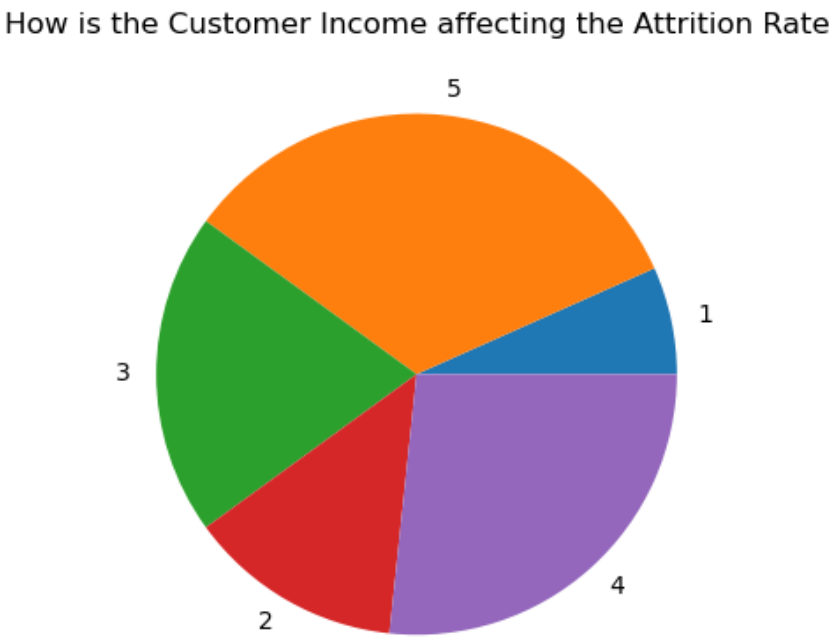
From this graph, we see the Revolving balance is increasing as there is an increase in the Utilization Ratio. The Revolving Balance is the outstanding amount on the credit card and the average utilization ratio is the utilization of card by the user. This shows that as more the customer uses the card, the higher is the Revolving Balance of the user.

8. How is the Age range for Attrited Customers?



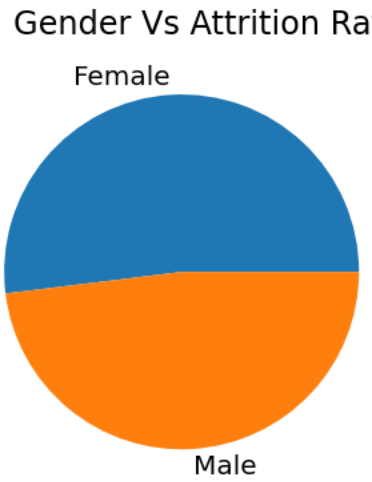
From the histogram, we can observe that the highest number of attired customers, over 300, are within the ages of 45 – 55. There are less than 50 customers below 30 years and above 60 years who are attired. This shows the Attrition Rate is higher among middle-aged people.

9. How is the Customer Income affecting the Attrition Rate?



From the Pie chart, we see the customers with Income Range above $120K have higher attired customer, followed by ranges $80K - $120K and $60K - $80K. This shows that people with higher income have higher attrition, maybe because they have higher earnings and might need less credit.

10. What gender is Attriting the most?



From the Pie chart, we can conclude that Females are attriting the most, with 748 being attrited, compared to Males, with 692 being attrited.

Models and Analysis

Our problem is a Classification problem and we have used the following four Models to check for the Accuracy and to choose the best Model for this dataset.

We have set the Attrition_Flag column as our predictor column and split the Test and Training data in 80-20 split.

The four models we used are:

1. Logistic Regression
2. Naïve Bayes
3. Random Forest
4. Decision Tree

```
LogisticRegression(max_iter=1000)
<class 'numpy.ndarray'> (1803,) <class 'pandas.core.series.Series'> (1803,)
0.8957293399889074
[[ 153  137]
 [  51 1462]]
                   precision    recall  f1-score   support

Attrited Customer       0.75      0.53      0.62       290
Existing Customer       0.91      0.97      0.94      1513

         accuracy                           0.90      1803
        macro avg       0.83      0.75      0.78      1803
     weighted avg       0.89      0.90      0.89      1803

GaussianNB()
<class 'numpy.ndarray'> (1803,) <class 'pandas.core.series.Series'> (1803,)
0.9018302828618968
[[ 183  107]
 [  70 1443]]
                   precision    recall  f1-score   support

Attrited Customer       0.72      0.63      0.67       290
Existing Customer       0.93      0.95      0.94      1513

         accuracy                           0.90      1803
        macro avg       0.83      0.79      0.81      1803
     weighted avg       0.90      0.90      0.90      1803
```

```
RandomForestClassifier()
<class 'numpy.ndarray'> (1803,) <class 'pandas.core.series.Series'> (1803,)
0.9561841375485303
[[ 230   60]
 [  19 1494]]
                  precision    recall  f1-score   support

Attrited Customer       0.92      0.79      0.85       290
Existing Customer       0.96      0.99      0.97      1513

         accuracy                           0.96      1803
        macro avg       0.94      0.89      0.91      1803
     weighted avg       0.96      0.96      0.95      1803

DecisionTreeClassifier(max_depth=2, random_state=0)
<class 'numpy.ndarray'> (1803,) <class 'pandas.core.series.Series'> (1803,)
0.8940665446478092
[[ 150  140]
 [  51 1462]]
                  precision    recall  f1-score   support

Attrited Customer       0.75      0.52      0.61       290
Existing Customer       0.91      0.97      0.94      1513

         accuracy                           0.89      1803
        macro avg       0.83      0.74      0.77      1803
     weighted avg       0.89      0.89      0.89      1803
```

Logistic Regression:

For Logistic Regression, we set the max iterations to 1000.
The Attrited customer precision is 0.75, recall is 0.53, f1-score is 0.62 and support 290.
The Existing customer precision is 0.91, recall is 0.97, f1-score is 0.94 and support 1513.
Accuracy of f1-score is 0.90 and support is 1803.
Macro average of precision is 0.83, recall is 0.75, f1-score 0.78 and support 1803.
Weighted average of precision is 0.89, recall is 0.90, f1- score is 0.89 and support is 1803.

Naive Bayes:

The Attrited customer precision is 0.72, recall is 0.63, f1-score is 0.67 and support 290.
The Existing customer precision is 0.93, recall is 0.95, f1-score is 0.94 and support 1513.
Accuracy of f1-score is 0.90 and support is 1803.
Macro average of precision is 0.83, recall is 0.79, f1-score 0.81 and support 1803.
Weighted average of precision is 0.90, recall is 0.90, f1- score is 0.90 and support is 1803.

Random Forest:

The Attrited customer precision is 0.92, recall is 0.79, f1-score is 0.85 and support 290.
The Existing customer precision is 0.96, recall is 0.99, f1-score is 0.97 and support 1513.
Accuracy of f1-score is 0.96 and support is 1803.
Macro average of precision is 0.94, recall is 0.89, f1-score 0.91 and support 1803.
Weighted average of precision is 0.96, recall is 0.96, f1- score is 0.95 and support is 1803.

Decision Tree:

For Decision Tree, we set the max depth of the tree to 2.
The Attrited customer precision is 0.75, recall is 0.52, f1-score is 0.61 and support 290.
The Existing customer precision is 0.91, recall is 0.97, f1-score is 0.94 and support 1513.
Accuracy of f1-score is 0.89 and support is 1803.
Macro average of precision is 0.83, recall is 0.9, f1-score 0. and support 1803.
Weighted average of precision is 0.89, recall is 0.89, f1- score is 0.89 and support is 1803.

Conclusion:

According to the model results , Random forest has the highest score of an accuracy which value equals to 0.96 meaning that 96% of the data are predicted accurate whereas Logistic regression (0.90), Naïve Bayes (0.90) and Decision Tree (0.89) have the lower value of accuracy. Additionally, comparing with the precision values, the result from Random Forest model has the higher value in both predicting of attributed customer (0.92) and existing customer (0.96) variable. As a result, the Random Forest is selected for the prediction model.

Findings and Manegerial Implications

Here are some of our Findings from our project. We conclude that the Income category, with income less than $60K, influences the Credit Limit. We also conclude that the active/inactive of the user doesn't affect with user attrition. The customers with higher transaction amount change rate, mostly had transactions which were low compared to overall transaction amount. Most of the Customers have the same transaction count over the year.

The Blue Card, has the highest Attrition and the Platinum card has the highest Attrition Rate, compared to the other cards. Customers, with 3 or less dependents, have lower balance difference and are more likely to pay their bills on time. The higher the utilization rate of the card for the user, the higher is the Revolving Balance of the user.

We also concluded that there is a higher Attrition Rate among middle-aged people and people with higher income have higher attrition, probably because they have higher earnings and might need less credit. Females are attriting the most, compared to Males being attrited.

We also found that Random Forest is best model for predicting our dataset of customer credit card churn.

Conclusions

This project aims at predicting the loss of bank credit card customers, we have almost 10,000 customers with data of their age, salary, marital status, credit card limit, credit card category, etc. and performed Predictions and Analysis based on it. We preprocess the data and then apply four classification models, Logistic Regression, Naive Bayes, Random Forest, Decision tree and found Random Forest to be the better model with higher accuracy rate.

From the Quantitative, Descriptive and Predictive Analysis from above, the manager of bank can use the data and improve credit card service to decrease the customer churn.

Appendix

```
# Importing Packages and all
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
from sklearn.naive_bayes import GaussianNB
import seaborn as sns
from matplotlib import pyplot as plt
from sklearn import tree
from sklearn import ensemble


# Pandas setting to show all columns when we print dataframe
pd.set_option('display.max_columns', None, 'display.max_rows', None)

#Reading the dataset and setting it as a DataFrame
ccdata = pd.read_csv("C:\\Users\\srith\\Downloads\\BankChurners.csv")
ccdata = pd.DataFrame(ccdata)

# cleaning 1 - Renaming the Last two columns and dropping them.
ccdata.rename(columns={'Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contact
s_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1': 'PYes'},
inplace=True)
ccdata.rename(columns={'Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contact
s_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2': 'PNo'},
inplace=True)
ccdata = ccdata.drop(['PYes', 'PNo'], axis=1)

# Checking Information about the Dataset

print(ccdata.shape)
print(ccdata.columns)
print(ccdata.describe())
print(ccdata.info())
print(ccdata.dtypes)


# Pie chart
ccdataval = ccdata['Attrition_Flag'].value_counts()
ccdatalabel = ccdata['Attrition_Flag'].unique().tolist()
plt.pie(ccdataval, labels= ccdatalabel, radius= 1)
plt.show()


# ------------------------------------------------------------------------------
-----------------------------------

# cleaning 2 - Replacing the Attrition Flag Column to Binary values
ccdata['Attrition_Flag'].replace(['Attrited Customer', 'Existing Customer'], [0,
1], inplace=True)

# cleaning 3 - Replacing the Gender column to Binary values
ccdata['Gender'].replace(['M', 'F'], [0, 1], inplace=True)
```

```
# cleaning 4 - One hot encoding - for Education Level column
ccdata = pd.get_dummies(ccdata, columns=['Education_Level'], prefix=['Education-
Level'])
ccdata = ccdata.reindex(columns=ccdata.columns, fill_value= 0)

# cleaning 5 - One hot encoding - for Marital Status Column
ccdata = pd.get_dummies(ccdata, columns=['Marital_Status'],
prefix=['Marital_Status'])
ccdata = ccdata.reindex(columns=ccdata.columns, fill_value= 0)

# cleaning 6 - Ordinal Encoding - For Income category to see if the income level
can have an affect in the credit score
# Dropping unknown income
print(ccdata.Income_Category.unique())
mapping = {'Unknown': 0, 'Less than $40K': 1, '$40K - $60K': 2, '$60K - $80K': 3,
'$80K - $120K': 4, '$120K +': 5}
ccdata = ccdata.replace(mapping)
ccdataincome = ccdata[ccdata['Income_Category'] == 0].index
ccdata.drop(ccdataincome, inplace=True)

# cleaning 7 - One hot encoding - for Card category
ccdata = pd.get_dummies(ccdata, columns=['Card_Category'],
prefix=['Card_Category'])
ccdata = ccdata.reindex(columns=ccdata.columns, fill_value=0)

# drop columns as client number is unique to organization
ccdata = ccdata.drop(['CLIENTNUM'], axis=1)

# Checking Information about the Dataset

# print(ccdata.tail())
# print(ccdata.shape)
# print(ccdata.columns)
# print(ccdata.head())
# print(ccdata.info())

# Data cleansing is done


# setting the prediction column if the customer is still existing or attrited
y = ccdata['Attrition_Flag']
print(type(y))

# selecting the dependent variables and dropping the prediction column
x = ccdata.drop(['Attrition_Flag'], axis=1)
print(type(x), x.shape)

# creating test and training datasets from x and y
# test_size -->  represent the proportion of the dataset to include in the test
split
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.20)
print("X_train",type(X_train), X_train.shape)
print("y_train",type(y_train))
print("X_test",type(X_test), X_test.shape)
print("y_test",type(y_test))
```

```
# Fitting into Models
# Logistic Regression
model1 = LogisticRegression(solver='lbfgs', max_iter=1000)
print(model1)
model1.fit(X_train, y_train)
predictions = model1.predict(X_test)
print(type(predictions),predictions.shape,type(y_test),y_test.shape)
print(accuracy_score(y_test,predictions))
print(confusion_matrix(y_test,predictions))
print(classification_report(y_test, predictions, target_names=['Attrited Customer',
'Existing Customer'], zero_division=1))

# Naive Bayes
model2 = GaussianNB()
print(model2)
model2.fit(X_train,y_train)
predictions = model2.predict(X_test)
print(type(predictions),predictions.shape,type(y_test),y_test.shape)
print(accuracy_score(y_test,predictions))
print(confusion_matrix(y_test,predictions))
print(classification_report(y_test,predictions,target_names=['Attrited Customer',
'Existing Customer'], zero_division=1))

# Random Forest
model4 = ensemble.RandomForestClassifier()
print(model4)
model4.fit(X_train,y_train)
predictions = model4.predict(X_test)
print(type(predictions),predictions.shape,type(y_test),y_test.shape)
print(accuracy_score(y_test,predictions))
print(confusion_matrix(y_test,predictions))
print(classification_report(y_test,predictions,target_names=['Attrited Customer',
'Existing Customer'], zero_division=1))

# Decision Tree Classifier
model3 = tree.DecisionTreeClassifier
model3 = tree.DecisionTreeClassifier(random_state= 0, max_depth= 2)
print(model3)
model3.fit(X_train,y_train)
predictions = model3.predict(X_test)
print(type(predictions),predictions.shape,type(y_test),y_test.shape)
print(accuracy_score(y_test,predictions))
print(confusion_matrix(y_test,predictions))
print(classification_report(y_test,predictions,target_names=['Attrited Customer',
'Existing Customer'], zero_division=1))

# ------------------------------------------------------------------------------
----------------------------------

# questions
# 1. What is the effect of Credit Limit on Income Category?

print('Credit Limit Vs', ccdata.groupby('Income_Category')['Credit_Limit'].max())

# 2. How many of the Attrited Customers are inactive for more than 3 months in the
last 12 months?
```

```
# print(ccdata.head())
attrdcust = []
for row in ccdata.itertuples():
    if row[8] > 3 and row[1] == 0:
        attrdcust.append(row)
print('Number of Attrited customers Inactive for more than 3 months in the last 12
months:',len(attrdcust))


# 3. Is there any change in Maximum and Minimum Transaction Amount overall with
Transaction Amount Change Rate greater than 1?

ccdataamt = []
for row in ccdata.itertuples():
    if row[13] > 1.0:
        ccdataamt.append(row[14])
print('The minimum Total transaction amount overall is :',
min(ccdata.Total_Trans_Amt),
      '\nThe minimum Total transaction amount for people with Transaction Amount
Change Rate greater than 1 is :',min(ccdataamt),
      '\nThe maximum Total transaction amount overall is :',
max(ccdata.Total_Trans_Amt),
      '\nThe maximum Total transaction amount for people with Transaction Amount
Change Rate greater than 1 is :', max(ccdataamt))


# 4. How is the Total Transaction Count is changing with Transaction Count Change
Rate?
ccdatacntg1 = []
ccdatacntg2 = []
for row in ccdata.itertuples():
    if row[16] >= 1.0:
        ccdatacntg1.append(row[15])
    else:
        ccdatacntg2.append(row[15])
print('The minimum Total transaction Count overall is :',
min(ccdata.Total_Trans_Ct),
      '\nThe maximum Total transaction Count overall is :',
max(ccdata.Total_Trans_Ct),
      '\nThe minimum Total transaction Count for people with Transaction Count
Change Rate greater than or equal to 1 is :',min(ccdatacntg1),
      '\nThe maximum Total transaction Count for people with Transaction Count
Change Rate greater than or equal to 1 is :',max(ccdatacntg1),
      '\nThe minimum Total transaction Count for people with Transaction Count
Change Rate less than 1 is :', min(ccdatacntg2),
      '\nThe maximum Total transaction Count for people with Transaction Count
Change Rate less than 1 is :', max(ccdatacntg2))


# 5. Which type of Credit card has the highest Attrition?

print('For Blue Card:',
ccdata.groupby('Attrition_Flag')['Card_Category_Blue'].sum())
print('For Gold Card:',
ccdata.groupby('Attrition_Flag')['Card_Category_Gold'].sum())
print('For Platinum Card:',
```

```
ccdata.groupby('Attrition_Flag')['Card_Category_Platinum'].sum())
print('For Silver Card:',
ccdata.groupby('Attrition_Flag')['Card_Category_Silver'].sum())

# 6. Is there any relation between Dependent Count and Open to buy Credit Line
Average?

plt.bar(ccdata.Avg_Open_To_Buy, ccdata['Dependent_count'])
plt.title('Relation b/n Credit Line and Dependents')
plt.xlabel('Open to buy Credit Line')
plt.ylabel('Dependents')
plt.show()

# 7. What is the relation between Revolving Balance and Card Utilization Ratio?

plt.bar(ccdata['Total_Revolving_Bal'], ccdata['Avg_Utilization_Ratio'])
plt.title('Relation b/n Revolving Balance and Utilization Ratio')
plt.xlabel('Revolving Balance')
plt.ylabel('Utilization Ratio')
plt.show()

# 8. How is the Age range for Attrited Customers?

ccdataattr = ccdata.loc[ccdata.Attrition_Flag == 0]
plt.hist(ccdataattr.Customer_Age)
plt.title('How is Age affecting the Attrition Rate')
plt.xlabel('Customer Age')
plt.show()

# 9.What is the relation between Income Range and Attrition?

ccdataincval = ccdataattr['Income_Category'].value_counts()
ccdataincunqlst = ccdataattr['Income_Category'].unique().tolist()
plt.pie(ccdataincunqlst, labels= ccdataincunqlst, radius= 1)
plt.title('How is the Customer Income affecting the Attrition Rate')
plt.show()


# 10. What gender is Attriting the most?

fig = plt.figure(figsize=(4,3), dpi= 144)
mylabels = ['Female','Male']
ccdatagenatr = ccdataattr['Gender'].value_counts().to_frame()
plt.pie(ccdatagenatr.Gender, labels=mylabels)
plt.title('Gender Vs Attrition Rate', loc= 'right')
plt.show()
```

References

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8898559/
- https://prosperitythinkers.com/money/mpact-of-credit-in-the-economy/
- https://mygreatlakes.org/educate/knowledge-center/credit.html
- https://www.kaggle.com/datasets/whenamancodes/credit-card-customers-prediction
- https://usa.visa.com/dam/VCOM/download/corporate/media/moodys-economy-white-paper-feb-2013.pdf
- https://merchantservicesnewjersey.com/definitions/open-to-buy/
- https://www.rocketloans.com/learn/financial-smarts/revolving-debt#:~:text=If%20you%20have%20an%20outstanding%20balance%20on%20your,your%20credit%20score%20%E2%80%93%20if%20it%E2%80%99s%20handled%20correctly.
- https://python-course.eu/machine-learning/confusion-matrix-in-machine-learning.php