Customer Attrition Prediction Based on Characteristics

Zaniar Javanroodi

Northeastern University

Toronto, ON

z.javanroodi@northeastern.com

Abstract

In this model we are trying to predict if a customer is likely to churn and take steps to prevent that from happening before a decision is made by the customer. The goal is to analyze the data we have on specific characteristics and see if there is a connection.

Dataset

A part of the data is shown below with columns.

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_bo
0	768805383	Existing Customer	45	M	3	High School	Married	$60K-80K$	Blue	
1	818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	
2	713982108	Existing Customer	51	M	3	Graduate	Married	$80K-120K$	Blue	
3	769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	
4	709106358	Existing Customer	40	M	3	Uneducated	Married	$60K-80K$	Blue	
...	...	...	...	...	...	...	...	...	...	
10122	772366833	Existing Customer	50	M	2	Graduate	Single	$40K-60K$	Blue	
10123	710638233	Attrited Customer	41	M	2	Unknown	Divorced	$40K-60K$	Blue	
10124	716506083	Attrited Customer	44	F	1	High School	Married	Less than $40K	Blue	
10125	717406983	Attrited Customer	30	M	2	Graduate	Unknown	$40K-60K$	Blue	
10126	714337233	Attrited Customer	43	F	2	Graduate	Married	Less than $40K	Silver	

10127 rows × 21 columns

Customer Attrition Prediction Based on Characteristics

Data Description

For sake of understandability and ease of use the columns are renamed. Here is the names and a short description of them:

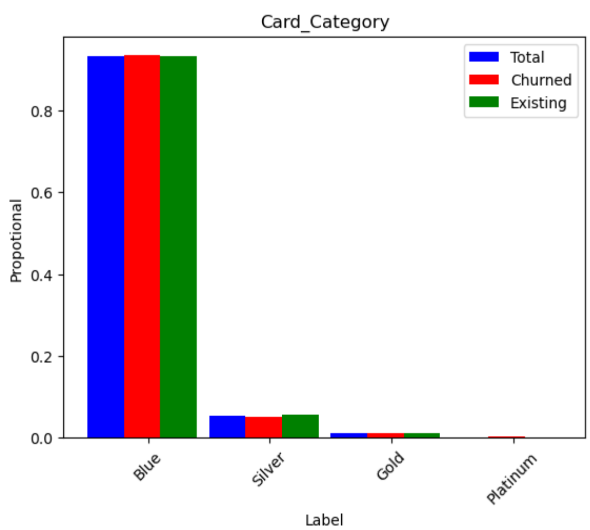
CLIENTNUM	Unique identifier for each customer. (Integer)
Attrition_Flag	Flag indicating whether the customer has churned out. (Boolean)
Customer_Age	Age of customer. (Integer)
Gender	Gender of customer. (String)
Dependent_count	Number of dependents that customer has. (Integer)
Education_Level	Education level of customer. (String)
Marital_Status	Marital status of customer. (String)
Income_Category	Income category of customer. (String)
Card_Category	Type of card held by customer. (String)
Months_on_book	How long customer has been on the books. (Integer)
Total_Relationship_Count	Total number of relationships customer has with the credit card provider. (Integer)
Credit_Limit	Credit limit of customer. (Integer)
Total_Revolving_Bal	Total revolving balance of customer. (Integer)
Avg_Open_To_Buy	Average open to buy ratio of customer. (Integer)
Total_Amt_Chng_Q4_Q1	Total amount changed from quarter 4 to quarter 1. (Integer)
Total_Trans_Amt	Total transaction amount. (Integer)
Total_Trans_Ct	Total transaction count. (Integer)
Total_Ct_Chng_Q4_Q1	Total count changed from quarter 4 to quarter 1. (Integer)
Avg_Utilization_Ratio	Average utilization ratio of customer. (Integer)

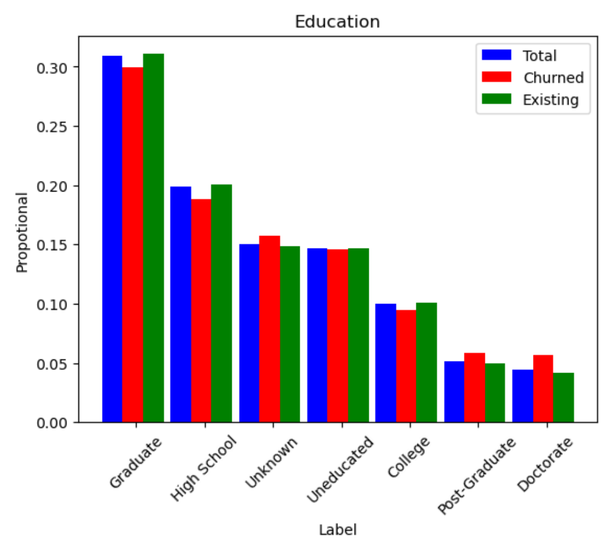
Customer Attrition Prediction Based on Characteristics

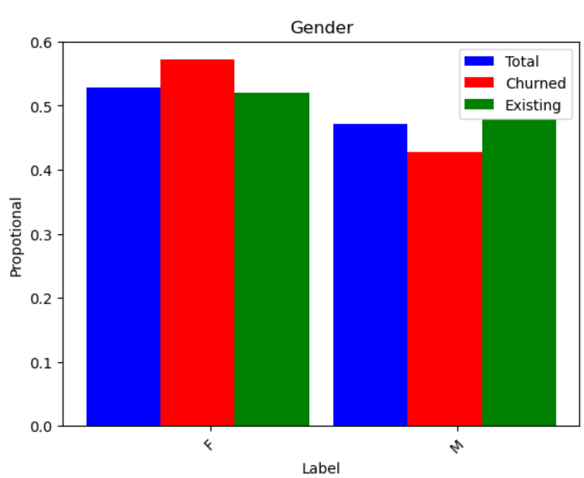
As we can see many of the features are categorical and we need to transform them into numbers to feed to our algorithm and due to the relation and umber of categories I chose labeling the data with numbers rather than One-Hot encoding.

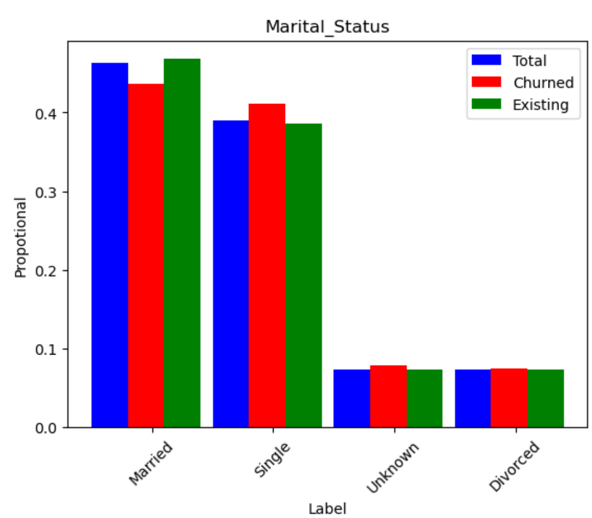
Data Analysis

I started with comparing categorical data from the two sides of the data (existing ,churned) to see if there is a obvious connection between the them. For example, if female customer are more likely to churn or a specific category of income. Here are some figures on those data types:

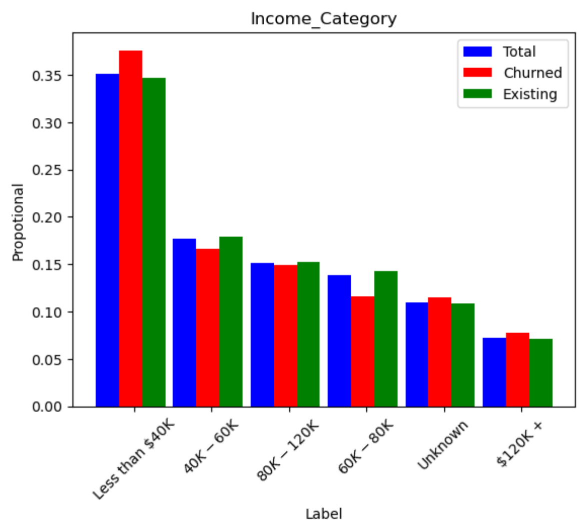








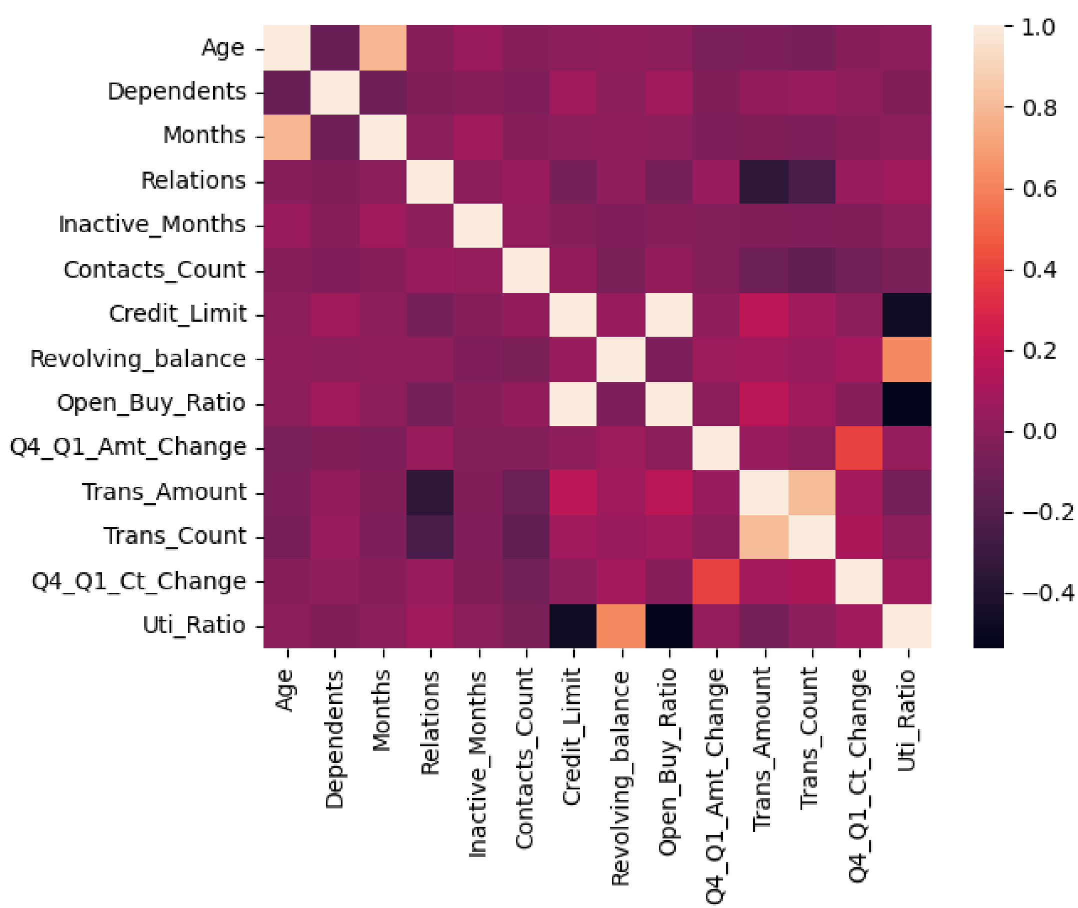
Customer Attrition Prediction Based on Characteristics



This graph shows the percentage of the different parts of the data compared (obviously the number of total is always greater that the other ones that's why we have to use percentage).

In the graphs above we can see the card category might not be a good feature since the percentage of both outcomes are almost the same.

We can also explore the relations of numerical data with a histogram:



Customer Attrition Prediction Based on Characteristics

There is a relationship between Uti_Ratio and Credit_Limit and Open_buy_Ratio because Utilization ratio is by definition related to those features. Also transaction amount and relations have a relation which also make sense because married customers tend to spend more. Other than that, there seems to be no apparent relation between data. There is also n missing values in our data.

Model Process

After the data analysis we can now start to train our models and compare the results. I am going to use K Neighbors Classifier, Decision Tree Classifier, Gradient Boosting Classifier, Ada Boost Classifier, Random Forest Classifier and Logistic Regression algorithms and compare their accuracy and speed. In figure.1 you can see the first results of the training of all algorithms.

- The data has been split for train and test by factor of 0.8 –

Module name in Sklearn	Training Accuracy	Testing Accuracy	Roc Auc Score	Time for Prediction
KNearestNeighbors	91.27268	89.23988	0.758664	0.263808
DecisionTreeClassifier	98.17307	94.22507	0.891052	0.00267
RandomForestClassifier	97.86446	94.91609	0.875057	0.056115
LogisticRegression	88.92729	89.38796	0.723416	0.002
AdaBoostClassifier	96.198	95.95262	0.91461	0.035068
GradientBoostClassifier	97.6793	96.79171	0.922231	0.012012

Figure .1

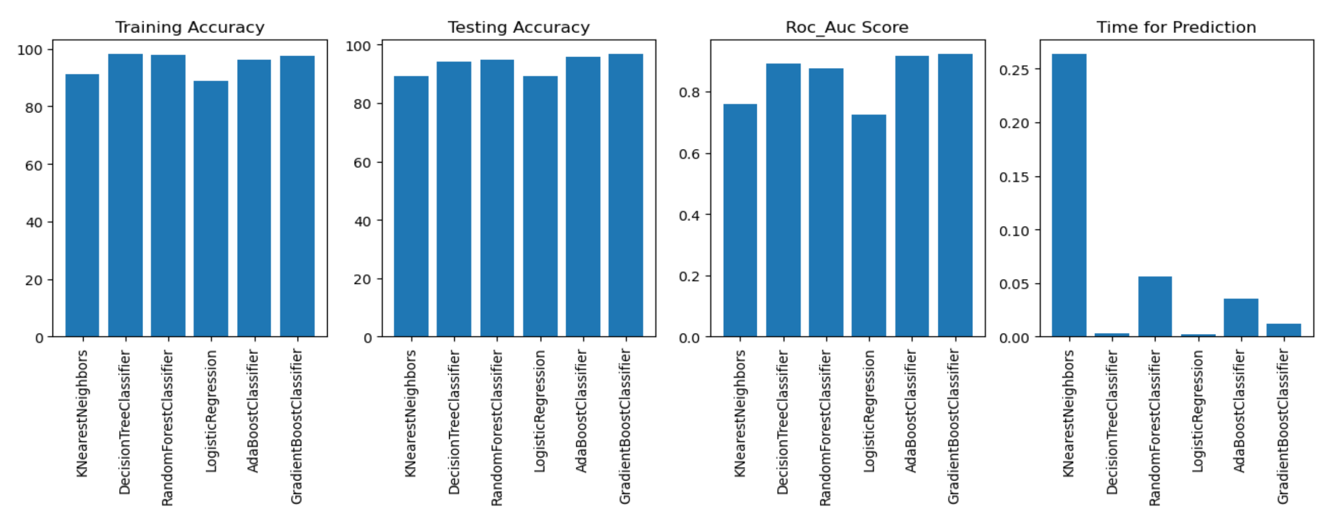


Figure .2

Customer Attrition Prediction Based on Characteristics

As we can see Logistic Regression and K Nearest Neighbor has low testing accuracy so we can use the other algorithms. The best overall algorithms are Gradient Boost Classifier if you prefer accuracy and Decision Tree Classifier if you prefer speed.

Now to test if we can improve the models by scaling the numerical numbers. Here are the mean and std. of the data:

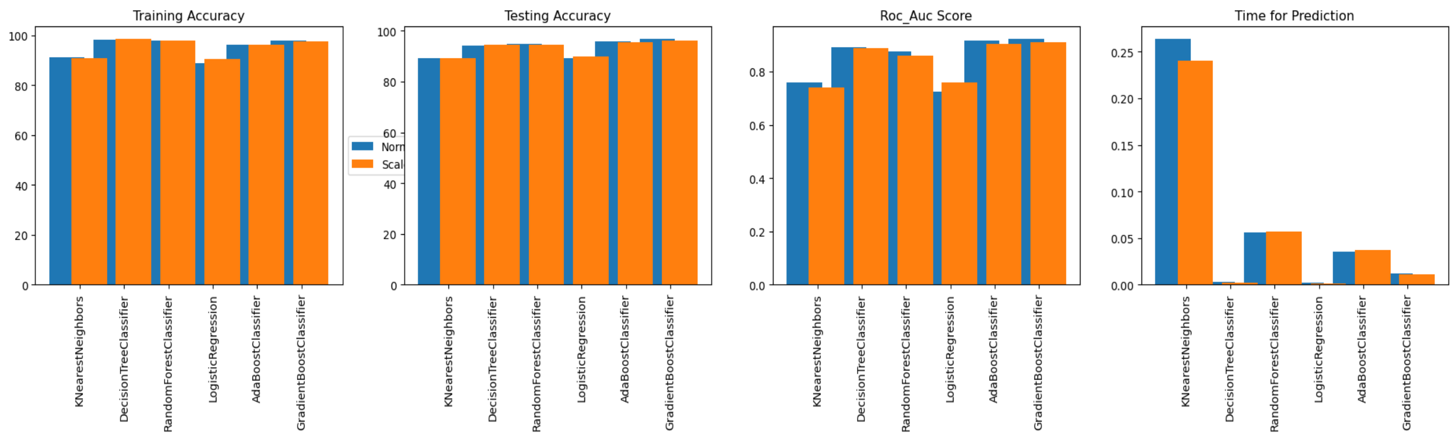
features	mean	std
Attrition_Flag	0.83934	0.367235
Age	46.32596	8.016814
Gender	0.470919	0.499178
Dependents	2.346203	1.298908
Education	3.096574	1.834812
Marital_Status	1.463415	0.737808
Income_Category	2.863928	1.5047
Card_Category	0.179816	0.693039
Months	35.92841	7.986416
Relations	3.81258	1.554408
Inactive_Months	2.341167	1.010622
Contacts_Count	2.455317	1.106225
Credit_Limit	8631.954	9088.777
Revolving_balance	1162.814	814.9873
Open_Buy_Ratio	7469.14	9090.685
Q4_Q1_Amt_Change	0.759941	0.219207
Trans_Amount	4404.086	3397.129
Trans_Count	64.85869	23.47257
Q4_Q1_Ct_Change	0.712222	0.238086
Uti_Ratio	0.274894	0.275691

There are some features with very high mean and std so I can try to scale these data to see if we can improve our models. – the scaled features are shown by color in figure 3 –

	Training Accuracy	Testing Accuracy	Roc_Auc Score	Time for Prediction
KNearestNeighbors	90.9147	89.3386	0.739376	0.24092
DecisionTreeClassifier	98.49401	94.47187	0.8869[illegible]	[illegible]02009
RandomForestClassifier	97.69164	94.47187	0.8595[illegible]	[illegible]57286
LogisticRegression	90.39625	89.83218	0.760218	0.001003
AdaBoostClassifier	96.29675	95.36032	0.904232	0.037673
GradientBoostClassifier	97.58055	96.29812	0.911045	0.011019

Figure .3

Now to compare the two results:



Conclusion

In this report I compared the results of different models and algorithms to see their differences in accuracy and speed. Also after seeing mean and std of the data it might seem like a good idea to scale the data but as we saw it did not make a lot of difference. I also strongly suggest that the code for this report to be examined to see the details of the work done. Another aspect of this models is the probability of the customer churn.

We can implement the code as such if the probability exceeds for example 40 percent we take measures to ensure the customers stays with bank like offer discount or any other form of advertising.

Resources

1 -Dataset: [Dataset](https://zenodo.org/record/4322342#.Y_ezgXbMJD9)(https://zenodo.org/record/4322342#.Y_ezgXbMJD9)

2 - https://pandas.pydata.org/docs/

3 - https://scikit-learn.org/stable/user_guide.html

4 - https://matplotlib.org/stable/index.html