A

Capstone Project

in

Data Science

BANK CUSTOMER CHURN

By:
Ganesh Bhandari

@

January 2021

1. Problem, background and data:

A manager at a ABC bank is disturbed because of more and more customers are leaving their credit card services. The manager would really appreciate it if someone could predict for them who is going to get churned so they can proactively go to the customer to provide them better services and turn customers decisions in the opposite direction.

I would like to help the manager by solving the problem. I talked with the manager and he sent me a data set which consists the record of more than 10,000 customers at the bank mentioning their age, salary, marital status, credit card limit, credit card category, etc. There are nearly 21 features. The description of the features of the data is as follows:

Variable	Type	Description
Clientnum	Num	Client number. Unique identifier for the customer holding the account
Attrition_Flag	char	Internal event (customer activity) variable - if the account is closed then 1 else 0
Customer_Age	Num	Demographic variable - Customer's Age in Years
Gender	Char	Demographic variable - M=Male, F=Female
Dependent_count	Num	Demographic variable - Number of dependents
Education_Level	Char	Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.)
Marital_Status	Char	Demographic variable - Married, Single, Unknown
Income_Category	Char	Demographic variable - Annual Income Category of the account holder (< $40K$, 40K - 60K, $60K-$ 80K, $80K-$ 120K, > $120K, Unknown)
Card_Category	Char	Product Variable - Type of Card (Blue, Silver, Gold, Platinum)
Months_on_book	Num	Months on book (Time of Relationship)
Total_Relationship_Count	Num	Total no of products held by the customer
Months_Inactive_12_mon	Num	No of months inactive in the last 12 months
Contacts_Count_12_mon	Num	No of Contacts in the last 12 months
Credit_Limit	Num	Credit Limit on the Credit Card
Total_Revolving_Bal	Num	Total Revolving Balance on the Credit Card
Avg_Open_To_Buy	Num	Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1	Num	Change in Transaction Amount (Q4 over Q1)
Total_Trans_Amt	Num	Total Transaction Amount (Last 12 months)
Total_Trans_Ct	Num	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Num	Change in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio	Num	Average Card Utilization Ratio

The data, provided by the Manager contains data related to the record of 10127 customers at the bank. At this time our target variable is 'Attrition Flag'.

2. Methods used:

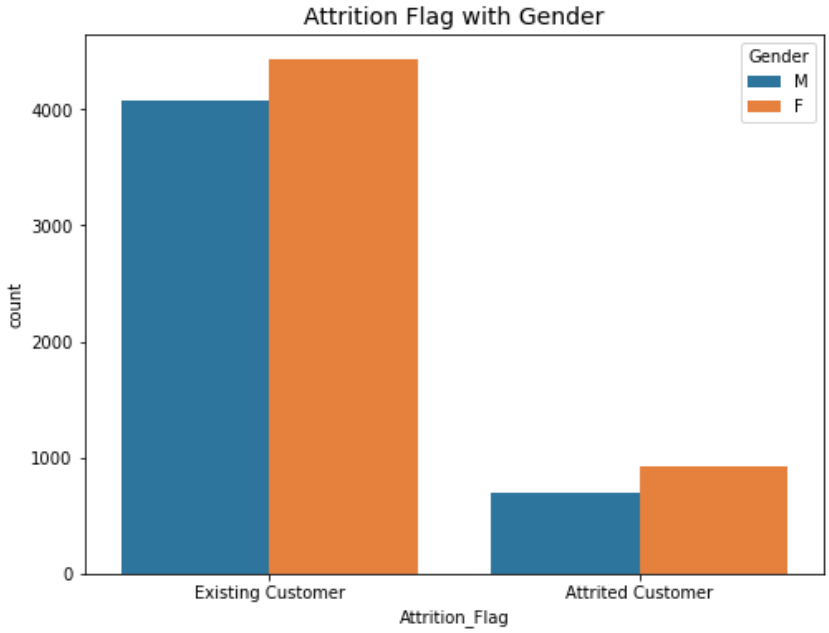
This problem is a classification problem. So to solve this problem I have created a model based on classification algorithms . In this project, I used classification algorithms like Logistic Regression, Naive Bayes, KNN, SVM, Decision Tree or Random Forest and chose the best model based on their accuracy performance. I used the best selected model to predict which customer is going to get churned. The steps that I used to get the result

are as follows:

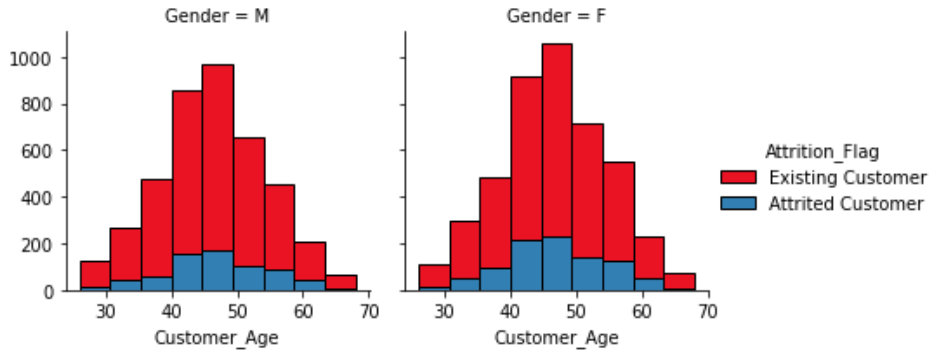
- Loading data
- Cleaning and organizing data
- Performing Exploratory Data Analysis (EDA) to understand the data and important features.
- Creating model and training the model
- Using suitable metric to check the performance of the model
- Deploying the model to get the results

3. Current Situation:

The data shows that the current situation of the bank is not good because of the number customers churning out. I found 16% of the customers are gone to churn. Also there is an interesting scenario gender wise.17 % of the female and 15 % of the male customers are churned.



Another interesting finding is that most of the churned customers are of age 40 - 50.



4. Model:

In this project, I used classification algorithms like Logistic Regression, Naive Bayes, KNN, SVM, Decision Tree or Random Forest to create a suitable model. The performance report of these model on the data is as follows:

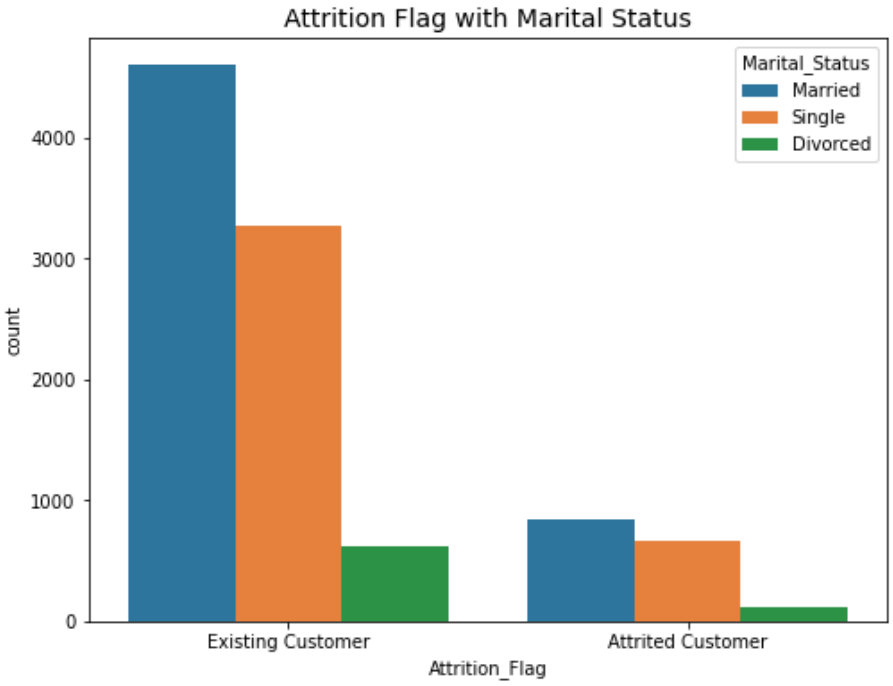
Algorithm	Accuracy Score	F1-score	LogLoss
Logistic Regression	0.91	0.91	0.21
Naive Bayes	0.88	0.88	0.48
KNN	0.88	0.87	1.43
SVM	0.92	0.91	NA
Decision Tree	0.92	0.92	0.22
Random Forest	0.95	0.95	0.14

The report table shows that, among all these algorithm, the Random Forest model 'RF' is more consistent and works better than the others. So, I chose the Random Forest model as the final model for my project.

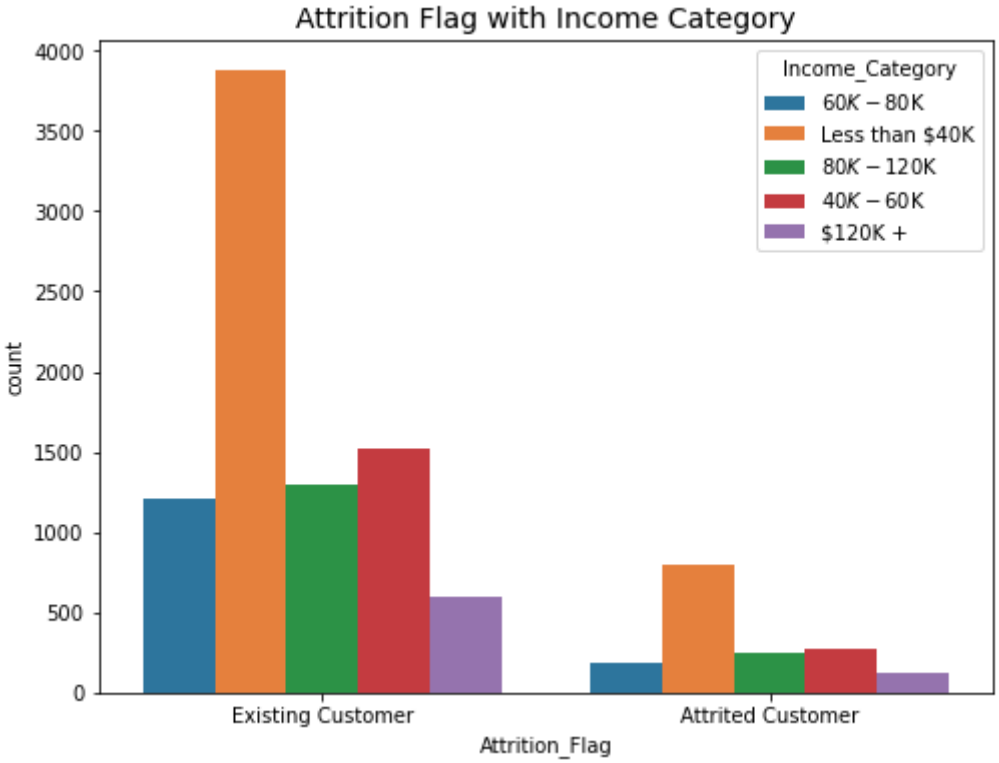
5. Findings:

I tried to do my best research on the data. After performing exploratory data analysis and deploying the model, I have some interesting findings on the data :

- More than 16% of the customers are gone to churn
- Divorced and Single customers have higher churned rate.



- 17 % of the female and 15 % of the male customers are churned.
- Most of the churned customers are of age 40 - 50.
- Customers with income level less than 40K have higher churning rate.



6. Recommendations:

Based on the findings obtained during data analysis process and deploy of the model, I have some recommendations for the bank:

- Emphasize on providing cards for male customers than female.
- Reduce the number of cases of providing cards for divorced and Single customers
- Reduce the number of cases of providing cards for the people of age around 40-50.
- Reduce the number of providing cards for the customers with income level less than 40K.

7. Conclusion:

Apply all the credit check methods strictly, collect correct data and do good research about the costumers before providing any credit card service. Also apply the model for each customer's information to predict whether the customer will be churned or not.

8. Acknowledgments:

I am very grateful to my mentor for his valuable suggestions while completing this project. Also I am thankful to the bank authority for providing valuable information.