Final Project Proposal-BIA650
Jungho Park, Byeongseon Park

Credit Card Attrition Rate Analysis and Forecast

1. Introduction

Credit card company drives its revenue and profits from three factors: interest, annual fees charge to cardholders, and transaction fees paid by merchant businesses. (Walletbuddy, 2017) To maximize its revenue and profits, credit card company must focus on increasing number of its cardholders (customers) and merchant businesses. In the modern online payment market, fierce competition in the battle for market share has driven plastic issuers to keep churning out the zero-rate or low-rate balance-transfer offers they've come to rely on. (Weber, 2004) Gaining and retaining customers as cardholders is essential element to achieve higher profit. However, this project only focuses on customer retainment by exploring factors that affect churn rate because generating profits from incoming customers takes longer than existing customers. For example, first year credit cardholder is usually receiving promotions such as low interest rate, zero annual fee and extra mileage for travel. Our team found a credit card attrition data from Kaggle. Since such data is usually confidential, the credit card churn data is only available from open data source. We try to perform exploratory data analysis to derive insights and to find important factors for churn rate. Additionally, we compare various prediction algorithms to see which model fits best for our subject.

2. Problem Description

The goal is to build models with various algorithms to predict whether a customer will churn or not. Also, we will discover important factors of churn to introduce marketing strategy to retain existing cardholders.

3. Dataset Description

As mentioned in the introduction, our dataset is from open data source: Kaggle. The dataset is regarding to 10128 customers and 21 column factors including credit card churn, account information, demographic, income, financial information. The replacement of NaN values is necessary to solve the problem, since there are some unknown values in education level, marital status, income category. This data contains 15 continuous variables and 6 categorical variables and ‘Attrition_Flag’ column is a categorical variable whether customer is currently a cardholder. ‘Existing Customer’ attribute indicates the customer is still holding a credit card and ‘Attrited Customer’ indicates the customers who are no longer cardholders. Unique categorical values is found by using unique() function of pandas. The results are as stated on *Table 1.* The other 15 continuous variables are data types with either Int64 or Float64.

Table 1. Unique Categorical Values

Column Names	Unique Categorical Values
Attrition_Flag	['Existing Customer', 'Attrited Customer']
Gender	['M', 'F']
Education_Level	['High School', 'Graduate', 'Uneducated', 'Unknown', 'College', 'Post-Graduate', 'Doctorate']
Marital Status	['Married', 'Single', 'Unknown', 'Divorced']
Income_Category	['$60K - $80K', 'Less than $40K', '$80K - $120K', '$40K - $60K', '$120K +', 'Unknown']
Card_Category	['Blue', 'Gold', 'Silver', 'Platinum']

4. Data Preparation

In order to proceed with given dataset, categorical values must be preprocessed by assigning each unique category into integers. Attrition, Gender are assigned with integer values as in *Table 2*. Machine learning algorithms cannot work with categorical data directly. Thus, categorical data must be converted to numbers (dummy variables). Education Level, Income Category, Marital Status, and Card Category are using hot encoding to create dummy variables. However, to ensure degree of freedom, ‘Unknown’ column is dropped from first three categorical columns, and ‘Platinum’ from ‘Card Category’. Furthermore, the original columns ought to be dropped from the data frame because dummy variables are present.

Table 2. Conversion of categorical values into integers

Column Names	Unique Categorical Values
Attrition_Flag	{'Existing Customer' : 0, 'Attrited Customer' : 1}
Gender	{‘M’:0,’F’:0}

The dataset consists 8,500 existing cardholder data and 1,627 churned data. Up-sampling churned data is necessary to compare with existing customer data, because the churned customer’s data takes only 16.07% of the whole data. This project uses SMOTE() to perform an up-sampling in order to provide deeper insight with possibility of being neglected when analysis is done without it. Furthermore, it will add more flexibility to a model and solve imbalance problem. The training and testing of the data will be performed with 'train_test_split' function with ratio of 75% and 25%. The process will be executed after performing exploratory data analysis and visualization.

Fig 1.

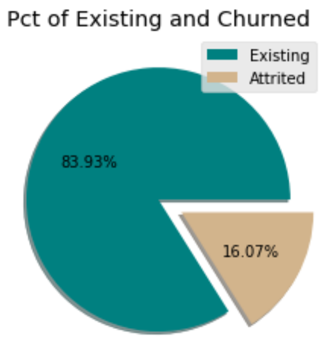
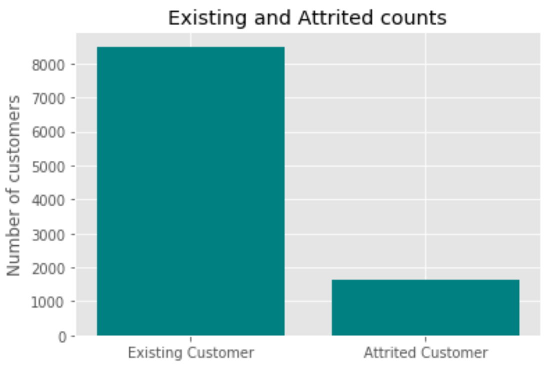


Fig 2.



5. Exploratory Data Analysis and Visualization

The data is divided in to two sub-parts: demographic data and credit card (financial) data. Demographic data are data that is related to customer's socioeconomic, demographic, and biological status, including age, gender, number of dependents, education levels, marital status, income status, and card category. Credit card data are months on book, total relationship count, months inactive 12 months, contacts count 12 months, credit limit, total revolving balance, average open to buy, total amount change q4 to q1, total transaction amount, total transaction ct, total ct change q4 to q1, average utilization ratio. Each and every visualizations and summary statistics will be introduced in the project report. However, we would like to introduce one of hidden insights from each data sub-parts that intrigued our interest.

- Demographic data: credit card category

Fig 3.

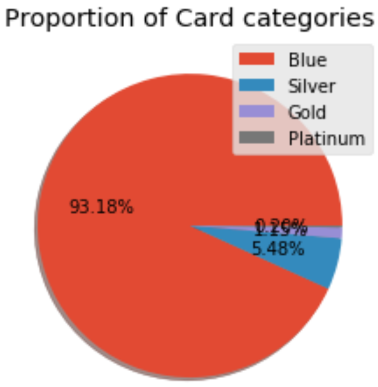
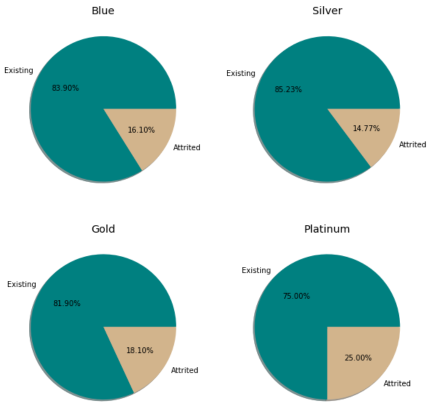


Fig 4. Attrition vs. existing for each card type



Unlike other demographic data attributes, the card category data shows that the churn proportion rate is significantly higher in platinum than blue, silver, and gold. This may suggest that the platinum credit cardholder have relatively less benefits than other cards. However, marketing department may ignore this factor because the absolute amount of platinum cardholder is less than 1%.

- Credit card data: average utilization ratio

Fig 5.

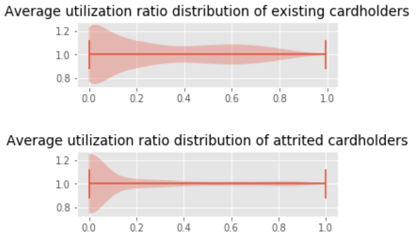


Fig 6.

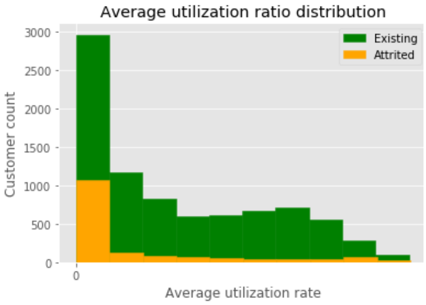
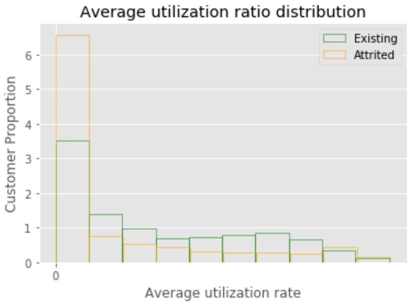


Fig 7.



Average utilization ratio (AUR) can be solved from dividing credit card debt by credit card limit. The lower the ratio is, lower the debt remain. In contrast to the common assumption: higher AUR will trigger churn, cardholder with lower AUR has higher chance of churn (*Fig 7.*).

6. Model Selection

After training and testing data split, we will fit data into various model such as random forest, logistic regression, k-NN, XGBoost, and gradient boosting. We will also measure each model's prediction performance by using confusion matrix and evaluate generated parameters. Accordingly, this project will carry conclusions and marketing remarks.

References

Weber, Joseph, and Mike McNamee. "A New Headache for the Credit-Card Biz." *BusinessWeek*, no. 3912, Dec. 2004, pp. 39–40. *EBSCOhost*, search.ebscohost.com/login.aspx?direct=true&db=buh&AN=15256295&site=ehost-live.

WalletBuddy. (2017, May 10). How do Credit Card companies make money - The Business Model. Retrieved December 01, 2020, from https://medium.com/walletbuddy-insights/how-do-credit-card-companies-make-money-the-business-model-d4892d301ac3

Goyal, S. (2020, November 19). Credit Card customers. Retrieved December 01, 2020, from https://www.kaggle.com/sakshigoyal7/credit-card-customers