DATA 1030 Midterm Report

Manlin Li
https://github.com/Marlenahaslee/DATA1030_Project

1 INTRODUCTION

Customer churn is when a customer stops using the service of a company's product, and in this case, stop being a paying client for a particular business. This problem is very important, especially for the audience of banks and credit card companies. For credit card companies to be profitable, they need to prevent their clients from gaming the system by churning. This project aims to build a classification tool for banks and credit card companies to help them predict whether a costumer will churn or not based on his/her demographical information as well as the record of how they have used the credit card. As an early warning of whether the existing customers are going to churn, this tool is useful in helping the bank and credit card companies in taking some action to maintain their customer retention.

The dataset obtained consists of a total of 10127 data with 21 valid columns. (There were originally 23 columns, but the author of the dataset suggested the last two columns containing data from Naïve Bayes Classifier Level to be removed.) The target variable `Attrition_Flag` is marked as either "Existing Customer" or "Attrited Customer".

Based on the literature research, all publications have the same goal of classification given the emphasis of this dataset on whether customers churned or not. Luc Chetboun and Thomas Konstantin were both interested in building a classifier to predict the `Attrition_Flag` for users (whether they churned or not). Moreover, they both chose SVM as one of their machine learning models, given that the mechanism of SVM is to find the best hyperplane that separates all data points into two classes. Luc Chetboun found in his project that AdaBoost achieved the best performance with precision of 0.88 and recall of 0.82 [1], and Thomas Konstantin in his project had random forest as the most robust model among all. Thomas Konstantin solves the class imbalance problem by SMOTE (Synthetic Minority Over-sampling Technique), and he also applied a PCA after the one-hot encoding to compress the dimensionality of the dataset [2].

2 Exploratory Data Analysis

During exploratory data analysis, the distribution, existence of null values, and unique values for each variable is explored by printing out the summary statistics and visualization. Bivariate visualizations between the target variable and the predictor variables are also conducted.

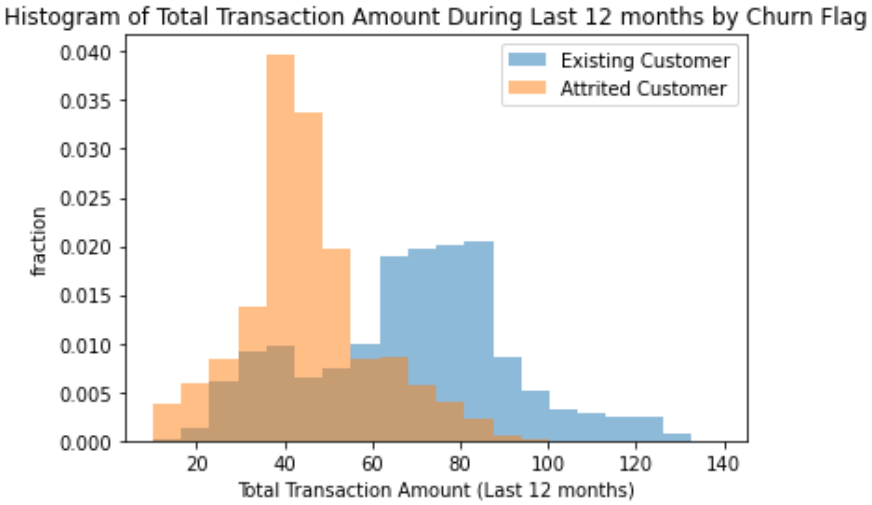

Figure 1 This category-specific histogram demonstrates the distribution of total transaction amount during the last 12 months by the churn flag of the customer: "Attrited Customer" or "Existing Customer". The bell-shaped pattern for each category indicates that they both approximate a Gaussian distribution. However, it can be easily seen that the distribution of total transaction amount for attrited customers is centered around 40 while that for existing customer is centered around 70. At the meantime, the former category has a smaller variance. This obvious difference in the distribution of transaction amount indicates that this variable is very likely to be an influencing factor in the classification model.

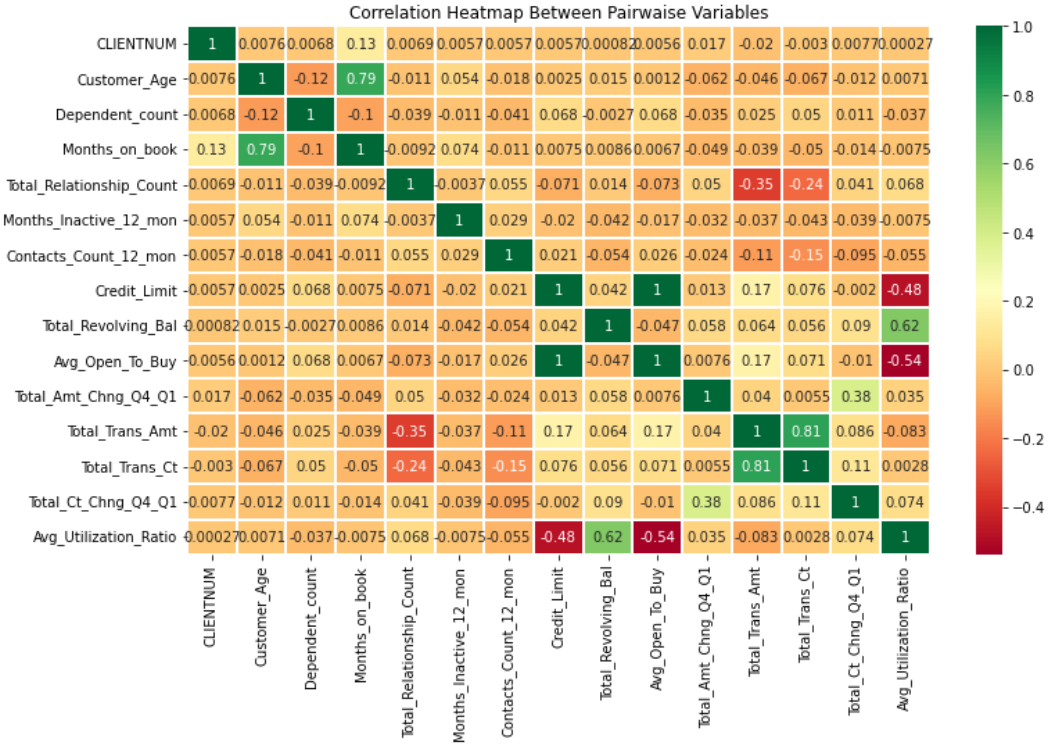

Figure 2 This correlation matrix heat map gives an overview of the linear relationship between predictor variables that are of continuous type. It can be seen from the graph that a cell with dark

green/red indicates a strong positive/negative correlation. For example, the correlation coefficient between `Months_on_book` and `Customer_Age` is 0.79, indicating that the period of relationship with bank has a positive linear relationship with the age of the customer. This visualization is important to understand because it may indicate the multicollinearity issue between predictor variables, which requests future attention when constructing the models.

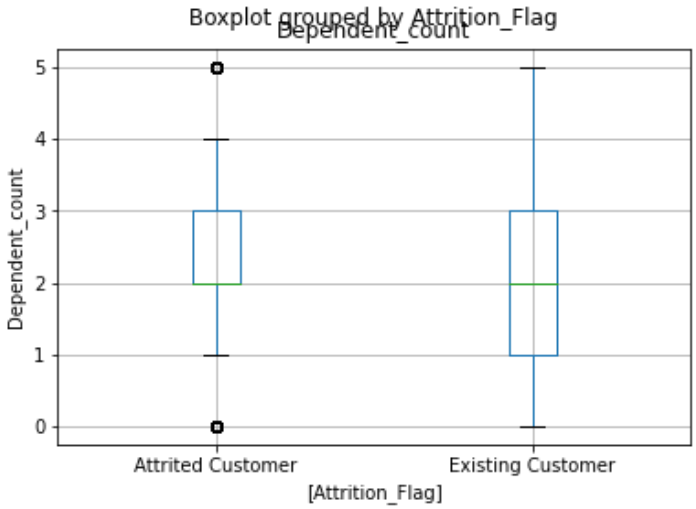

Figure 3 The above box plot shows the distribution of customers with different number of dependents on their credit cards. From the plot, the range of number of dependents for "Attrited Customer" is smaller than that for "Existing Customer". Also, the overlapping between the mean and the 25% quantile for "Attributed Customer" indicates the skewness for that category comparing to "Existing Customer" which follows a more bell-shaped distribution.

3 Data Preprocessing

After EDA, the `CLIENTNUM` column is dropped since it is a unique identifier that is meaningless in our classification model. Given that the target variable is a binary variable, it is also encoded by 0(Existing Customer) or 1(Attrited Customer) before splitting the data.

The data is first split into a training set with 70% data, and the rest 30% data is then split into a validation set with 20% data and a test set with 10% data. During splitting, stratified splitting is applied given the fact that this is an imbalanced dataset with 84% data in the category of "Existing Customer" in the target variable and only 16% data in the category of "Attrited Customer". This step is important to make sure that the splitting preserves the same proportions of data in both class as in the original dataset. Therefore, the final train-validation-testing split is 70-20-10.

The dataset is assumed to be Independent and Identically Distributed (iid). Each data entry represents the credit card usage pattern and personal information of an individual customer. Also, it was shown during the previous EDA section that each customer only has one unique data entry in the dataset. Therefore, it is neither a group-structured data nor a time-series data.

In the encoding step, a pipeline consisting of StandardScaler, OneHotEncoder, OrdinalEncoder, and MinMaxScaler is implemented to avoid the leaking statistics. StandardScaler is applied to the continuous variables that have no clear pattern of being bounded including `Dependent_count`, `Months_on_book`, etc. MinMaxScaler is applied to variables such as `Age` and `No. of Contacts in the last 12 months` which have clear bounds of 0 to 100 and 0 to 12, respectively. OrdinalEncoder is applied to orderd categorical variables including `Income_Category`, `Education_Level`, `Card_Category`. OneHotEncoder is applied to categorical variables `Gender` and `Marital_Status`. As a result, there are 23 features included in the preprocessed data.

Reference

[1]: Chetboun, Luc. Data Exploration, Model Evaluation on BankChurners. https://www.kaggle.com/chetbounl/data-exploration-model-evaluation-on-bankchurners

[2]: Konstantin, Thomas. Bank Churn Data Exploration And Churn Prediction. https://www.kaggle.com/thomaskonstantin/bank-churn-data-exploration-and-churn-prediction/notebook