Predicting Churning Customers

Data 1030 Midterm Report —— Yingfei Hong

October 12, 2021

1. Introduction

A manager at the bank is disturbed by more and more customers leaving their credit card services. The problem I am trying to solve is to predict the "churned customers" for the bank managers based on the given data so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction. Predicting churn is very important especially when clear customer feedback is absent. Retaining existing customers and thereby increasing their lifetime value is something everyone acknowledges as being important, however, there is little the bank managers can do about customer churn if they don't see it coming in the first place. This is where predicting churn has its value. Early and accurate churn prediction empowers CRM and customer experience teams to be creative and proactive in their engagement with the customer. In fact, by simply reaching out to the customer early enough, 11% of the churn can be avoided [1].

This dataset comes from LEAPS [2] and contains 21 columns and 10127 data points with 'Attrition_Flag' as its target variable in which "Attrited customer" means that this customer is the churned customer we are looking for.

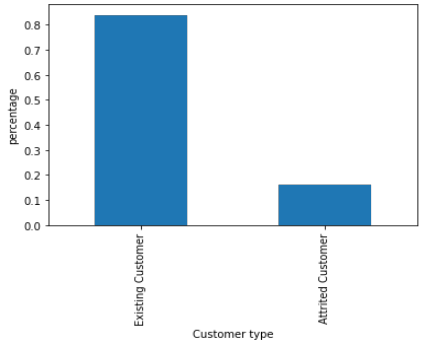

Figure 1 The distribution of 'Attrition_Flag'

It is an imbalanced dataset since the proportion of the 'Attrited customer' is about 16.07%. Besides, through the distribution of the target variable, this task is certainly a typical classification problem.

The rest 20 features are described in the following table which contains their feature name, type, and meaning.

Variable	Type	Description
Clientnum	Num	Client number. Unique identifier for the customer holding the account
Attrition_Flag	char	Internal event (customer activity) variable - flag to indicate if the customer is an existing customer or has attrited
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

Table 1 Description for each feature

For this data, there are two main tasks, one is to improve the performance of predicting churned customers while the other is to find the most influential factors that make the customers "churn". However, finding the factors that matter

most is a common strategy when enhancing the model performance so I rather treat them as a same task. Several projects have already been done to solve this problem and achieve good results.

Thomas[3] used SMOTE which is an approach to address imbalanced datasets by oversampling the minority class and found great improvement when processing the data generated by this strategy. The result showed that compared to the raw data, this new data could improve the F1 score from average 0.6 to average 0.9. Andi[4] looked into the details about the raw data and found some interesting relationships between the features and the target variable. The EDA showed that the likelihood of the customers' leaving is related to the money they spend annually, the months of inactivity in their bank account and their credit limit. Joseph[5] used Random Forest and LightGBM to predict with 97% recall and 95% accuracy and plotted the importance of these features which showed that the transaction feature ranked top in both models, so we need to look these features thoroughly when doing exploratory data analysis.

2. EDA

I've plotted the relationship between every feature and the target variable and found some relations that are worth notice.

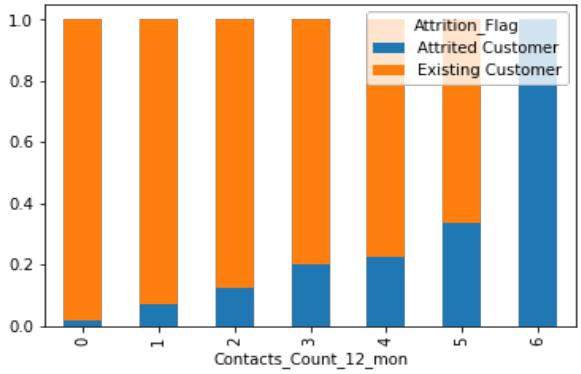

Figure 2 This graph displays how the number of contacts is distributed across two different customers. It seems that churning customers have had more contacts in the last 12 months with the bank managers.

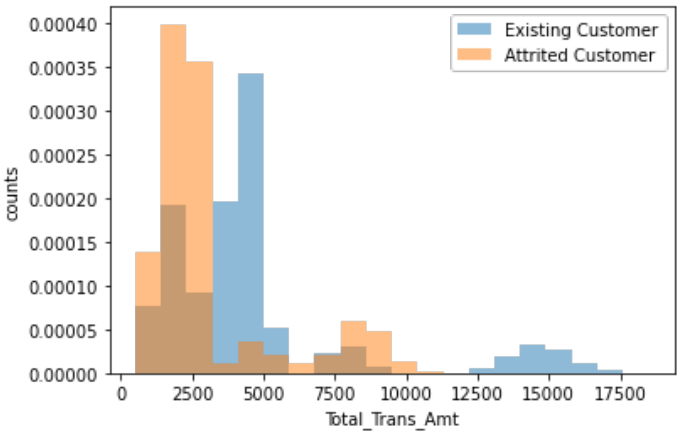

Figure 3 This graph shows that the total transaction amount of attrited customers is smaller than that of existing customers which explains why this feature ranks top in both models of Joseph's projects.

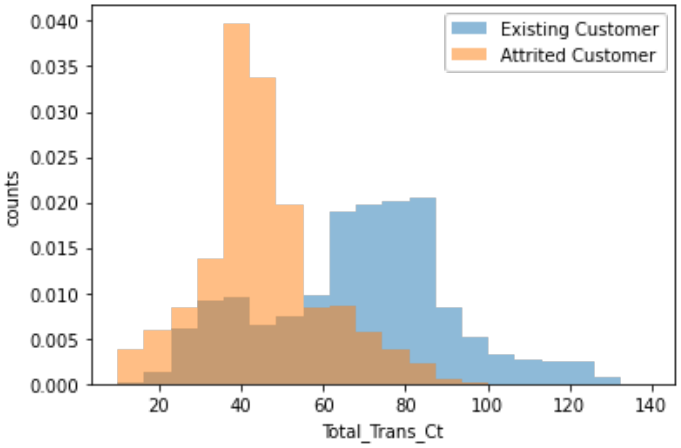

Figure 4 This graph has a similar pattern as figure 3 which shows that total transaction count is also a important feature when we try to classify attrited customer from the existing customer.

3. Data Proprocessing

This dataset is not IID since all features are not identically distributed. It does not have a group structure, nor does it have time-series data. But it is imbalanced, so I am going to use stratify method when splitting. The train size will be set as 0.6 and the validation size as well as the test size both as 0.2 because it is how people normally split their data when they have lots of data points.

I am going to treat discrete numerical features (dis_fea) as ordinal features and put them directly into the model since they are already in numerical form. Categorical features

(cat_fea) need to be treated with OneHotEncoder because it is not sensible if we put gender and marital status in order. For ordinal features(ord_fea), I am going to use OrdinalEncoder because it is obvious that there is ordinal information contained in educational level, income category, and card category. For continuous numerical feature(con_fea), I am going to use StandardScaler. Since in these columns, 'Customer_Age', 'Monts_on_book', and 'Total_Trans_Ct' are nearly normally distributed though some are skewed while other features have long tails which are not suitable to use MinMaxScaler.

4. Reference

[1] Why customers leave & what can banks do? Tiger Analytics. (2020, September 16). Retrieved October 12, 2021, from https://www.tigeranalytics.com/blog/addressing-customer-churn-in-banking/.

[2] Predict Customer Attrition Using Naïve Bayes Classification. ATH Leaps. Retrieved October 12, 2021, from https://leapsapp.analyttica.com/cases/11.

[3] Konstantin, T. (2021, May 1). Bank churn data exploration and churn prediction. Kaggle. Retrieved October 12, 2021, from https://www.kaggle.com/thomaskonstantin/bank-churn-data-exploration-and-churn-prediction.

[4] IDW, A. (2021, January 31). Customer churn - EDA, 95% ACC and 85% recall. Kaggle. Retrieved October 12, 2021, from https://www.kaggle.com/paotografi/customer-churn-eda-95-acc-and-85-recall.

[5] Chan, J. (2021, January 13). Bank Churners Classifier (Recall: 97% accuracy: 95%). Kaggle. Retrieved October 12, 2021, from https://www.kaggle.com/josephchan524/bankchurnersclassifier-recall-97-accuracy-95.

5. Github repository

https://github.com/YingfeiHong01/Data1030-FinalProject