Advanced Data Analytics - BUS 212A-2 Spring 2023

Title: Predicting Churn Rate of Credit Card Customers

Team Member Name: Hsiang-Tai Ling

Project Type: Kaggle Project

Project Page Link: shorturl.at/vWY29

Most Voted Notebook: shorturl.at/emrNQ

I. Introduction

1. Project Objective

The main objective of this final project is to deploy five major machine learning models, including Linear Regression Model, Logistic Regression Model with regularization, Naïve Bayes model, Nonlinear Support Vector Machine, and Random Forest algorithms to predict whether a customer is likely to stay with the bank for its credit card product, and identify some of the crucial features that best capture customers' decisions to provide insights for the company when making business decisions.

Although overall accuracy performance of the models is important, given the values of target y (0 if a customer leaves; 1 if a customer stays) in our dataset is imbalanced, we would put more emphasis on the performance of class 0 in the confusion matrix. From the company's perspective, the leaving of a customer is often costly, since it would directly affect revenues. Therefore, accuracy and recall rates are two major indicators in terms of model performance evaluation in our analysis.

2. Kaggle Approach

In the most voted notebook of this Kaggle case, the author uses Naïve Bayes and Logistic regression with L2 penalty to predict customer churn of the dataset. In data preprocessing part, the author utilizes LabelEncoder and StandardScaler functions turning categorical features into numerical values and rescaling variables to prevent extreme values affecting model accuracy. As for cross validation process, the author uses random split to obtain training and test sets. The author presents 100% accuracy in the two above-mentioned model.

However, the notebook suffers from several serious technical flaws, making the results not convincing. First, the author fails to identify, there are two features (hereinafter referred to as "two deterministic features") that had already been fine-tuned in predicting the target y. By conducting a correlation test, the result shows that these features are nearly perfectly correlated to the target y (0.99 and -0.99 respectively). By dropping these two deterministic features in the dataset, the whole case is another story.

Second, although the author plots a bar chart over the values of target y, the author does not take the imbalanced issue of target y into consideration when modeling. During the cross-validation process, the author not only conducts a single-time data split, but the data is split using shuffle approach, making results vulnerable. The codes of the above argument are attached in the Appendix section of the Jupyter Notebook for reference.

II. Approach

1. Data Preprocessing

After loading the raw data into Jupyter Notebook, we first check the shape of our dataset: there are originally 10,127 rows and 23 columns (see Appendix for complete list). Given the descriptive statistics for both categorical and numerical variables, we ensure that the dataset is complete and does not contain any missing value. Next, we investigate whether there is duplicate in the single variable "CLIENTNUM" and in the whole columns together. Again, the results are both zero, meaning the dataset is ready for analysis in terms of data completeness.

For features selection, as stated in the Introduction section, we find that the last two features "Naive_Bayes_Classifier…mon_1" and "Naive_Bayes_Classifier…mon_2" are fine-tuned deterministic features that would provide perfect predictions on our target y, making the task too easy. Therefore, we decide to drop these features in our dataset. Also, since "CLIENTNUM" is an irrelevant feature, we drop the column as well.

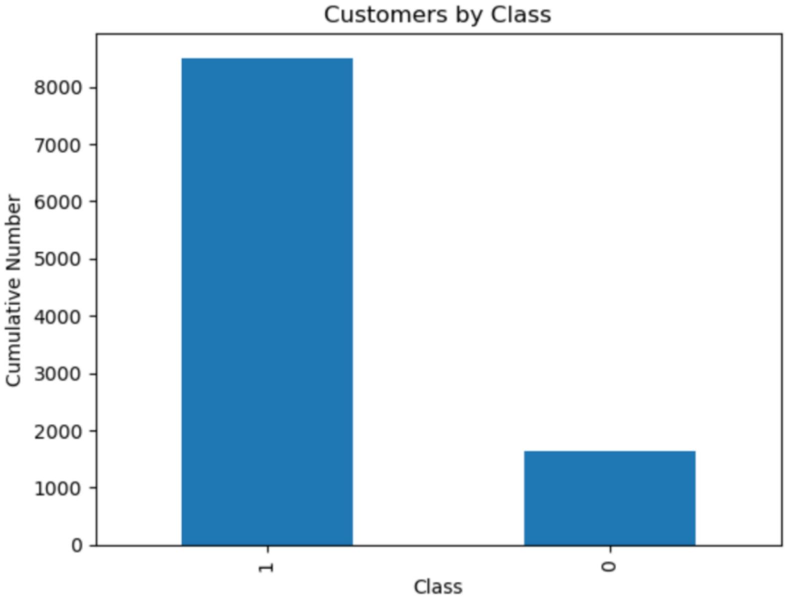
Target	"Attrition_Flag"
Features	"Customer_Age", "Gender", "Dependent_count", "Education_Level", "Marital_Status", "Income_Category", "Card_Category", "Months_on_book", "Total_Relationship_Count", "Months_Inactive_12_mon", "Contacts_Count_12_mon", "Credit_Limit", "Total_Revolving_Bal", "Avg_Open_To_Buy", "Total_Amt_Chng_Q4_Q1", "Total_Trans_Amt", "Total_Trans_Ct", "Total_Ct_Chng_Q4_Q1", "Avg_Utilization_Ratio"

Before feeding the data into our models, we transform six categorical features into numerical values using LabelEncoder, and rescale all the variables using StandardScaler to prevent the situation when spreadness of variables affect model performance. By far, we have selected 19 features and target y for analysis.

2. Modeling and Performance Evaluation

In this report, we use five different models to predict whether a customer stays. Linear Regression Model with Thresholds, Logistic Regression Model with Ridge Penalty, Naïve Bayes Model, Nonlinear Support Vector Machine, and Random Forest. The rationale behind model selection is that we choose models that have a linear classification boundary to nonlinear boundary, as we know in real world, most of the cases would not be linear separable. As for Linear Regression Model with Thresholds, it only serves the role of performance benchmark of the dataset.

For each model, we use stratified k-fold cross-validation approach to obtain training and test sets, since we find that the percentages of class 0 and 1 in the target y are imbalanced.



After fitting the data into different models, we use confusion matrix and classification reports, especially recall rate and weighted average accuracy, to compare the performance of each model. Overall, we would like to have a model that has the maximized recall rate, f1-score, and weighted average accuracy. In the end, we try to find some of the most crucial features that best capture the decisions made by customers.

III. Analysis

1. Data

The data describes the demographical (gender, education level) and financial attributes (credit limits, number of active months) of customer. However, plots of single and two variables is not very informative about whether they are related the attrition decision of the customer, since some variables are normally distributed, while some are heavily distributed toward certain class. In our case, we think presenting the bar chart graph of important features in Random Forest Model might be more interesting and insightful.

2. Algorithms

For each of the five methods, we conduct stratified k-fold cross validation, grid search on hyperparameter if available, and evaluate model performance based on confusion matrix and classification report. In particular, we focus more the predicting performance of class 0 (customers leaving), since companies tend to hurt directly when revenues decrease. Recall rate, F1-score and weighted average accuracy are three main indicators that we concerned most.

From our initial guess, we think nonlinear support vector machine and random forest would outperform other models, since they have flexible decision boundary, which is very suitable to capture nonlinear classification problem.

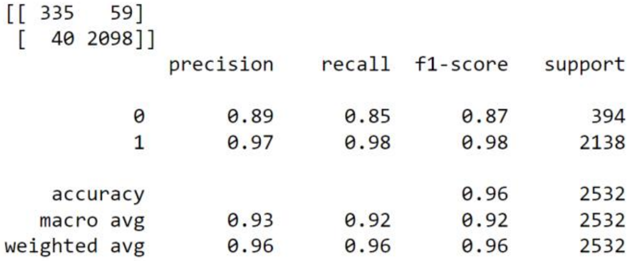
3. Experimental Details

For the cross-validation process, we use stratified_cv = StratifiedKFold(n_splits=nmc) in each model. We also test the run time of each model. In short, Random Forest has the fastest run time (2.36 seconds) and highest accuracy (0.96) and recall (0.85) performance. SVM takes longer time (2 minutes), having similar but lower accuracy (0.93) and recall (0.73). Surprisingly, Naïve Bayes do provide some quick (2.24 seconds) and decent performance (accuracy:0.89; recall: 0.68).

For Class 0	Accuracy	Recall	F1-score	Run Time
Random Forest	0.96	0.85	0.87	2.36 sec
Nonlinear SVM	0.93	0.73	0.78	111.64 sec
Naïve Bayes	0.89	0.68	0.66	2.24 sec

4. Results

Given the results of all five models, unsurprisingly, Random Forest has the best performance as graph below. The performance in Class 0 is better than I expected, since the classes are imbalanced in the raw data. Feature randomness of the tree model as well as the ensemble approach might be the reason behind, since it prevents the modeling from overfitting and has lower variance. The result is not comparable to the Kaggle notebook, since the notebook has wrong methodology.

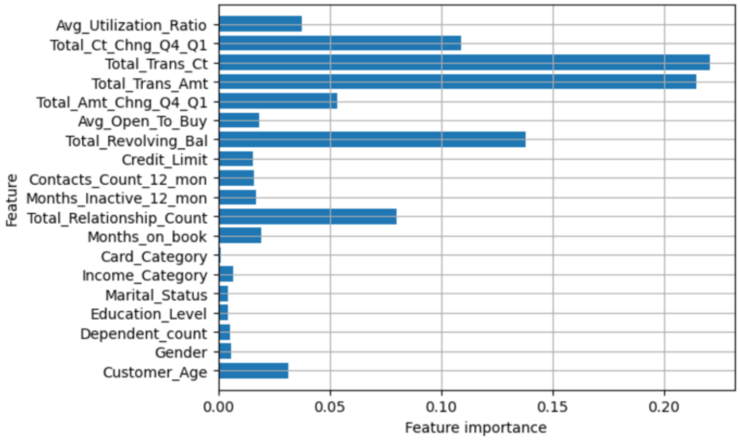


```
[[ 335   59]
 [  40 2098]]
              precision    recall  f1-score   support

           0       0.89      0.85      0.87       394
           1       0.97      0.98      0.98      2138

    accuracy                           0.96      2532
   macro avg       0.93      0.92      0.92      2532
weighted avg       0.96      0.96      0.96      2532
```

By utilizing Random Forest Model, we can observe by selecting certain features can help us better predict decisions made by customers. From the graph below, we can conclude it is more likely for company to identify attrition of customers based on transaction amount, count and total revolving balance of customer.



Appendix

Variable Table

Categorical variables are highlighted.

Features	Description
CLIENTNUM	Customer ID
Attrition_Flag	The decision whether a customer stays or leaves; Target y
Customer_Age	Age of customer
Gender	Gender of customer
Dependent_count	Number of dependents that customer has
Education_Level	Education level of customer
Marital_Status	Marital status of customer
Income_Category	Income category of customer
Card_Category	Type of card held by customer
Months_on_book	How long customer has been on the books
Total_Relationship_Count	Total number of relationships customer has with the credit card provider
Months_Inactive_12_mon	Number of months customer has been inactive in the last twelve months
Contacts_Count_12_mon	Number of contacts customer has had in the last twelve months
Credit_Limit	Credit limit of customer
Total_Revolving_Bal	Total revolving balance of customer
Avg_Open_To_Buy	Average open to buy ratio of customer
Total_Amt_Chng_Q4_Q1	Total amount changed from quarter 4 to quarter 1
Total_Trans_Amt	Total transaction amount
Total_Trans_Ct	Total transaction count
Total_Ct_Chng_Q4_Q1	Total count changed from quarter 4 to quarter 1
Avg_Utilization_Ratio	Average utilization ratio of customer
~~Naive_Bayes_Classifier…mon_1~~	~~Fine-tuned features~~
~~Naive_Bayes_Classifier…mon_2~~	~~Fine-tuned features~~