MET CS-677 Final Project (Summer1 2020)

Neha Bais

<nehabais@bu.edu>

**Credit Card Defaulters Prediction**

**Purpose:**

The goal of this project is to compare several machine learning models to predict whether the customers are going to default their credit card payment or not for the next month.

The dataset is extracted from UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>), which contains the information of credit card clients of Taiwan. It includes 30,000 observations and 24 features.

**Technical Requirements:**

* Python 3.7.6
* Sklearn 0.23.2
* Pandas 1.0.1
* Matplotlib 3.1.3

**Project Deliverables:**

* Data pre-processing
* Data Exploration
* Feature selection
* Hyper Parameter Tuning using GridSearchCV
* Evaluating different ML models including Logistic Regression, kNN, Random Forest, SVC, Naïve Bayesian

To run the project, simply download the project folder, place it in your working directly and run the *“****credit\_card\_default.ipynb”*** file directly through Jupyter Notebook.

The data dictionary is also included in zipped project folder to get the definition of different features of data.

**Project Description:**

The dataset contains 24 features, describing the demographic values like age, marital status, education etc. along with bill amount and ther payments, and one dependent variable (default. payment.next.month) where value 1 indicates that the customer will default and 0 indicates No default.

1. **Data Preprocessing**

I started with checking Null or missing values and if any, decided to drop the columns with more than 70% of missing values. Also, for categorical variables will replace them with mode and for numerical columns, missing values will be replaced by median. Luckily, there were no missing values.

I dropped the ID column as it does not make sense for my analysis. Also, renamed target column so as to get better readability.

There was some Unknown category in Marital status and Education level column, so I combined them into Others category.

1. **Exploratory Data Analysis**

Analyzing the target variable, I found that the dataset is pretty much imbalanced with only 22% of customers as defaulters.

After analyzing the age and Credit Limit variable, I found that the customers with age group 25~30 tend to default more and Frequency of default is more for customers having a card with lower credit limit.

For Categorical variables, University students default more. Female tend to default more than men and the Marriage column does not show much distinction between defaulters and no defaulters.

Also, looking at the correlation between various variables, the only variable with a notable correlation with default is payment status. The last month’s (September) payment status has a Pearson correlation coefficient of 0.32 with default.

The next highest correlation with default occurs with balance limit at -0.15, which itself was somewhat negatively correlated with payment statuses. These measures indicate that customers with lower balance limits have more delays in payments and are more likely to default.

Bill amounts and payment amounts are also moderately positively correlated with balance limit, but that is to be expected, since customers with higher credit limit are more likely to spend more.

**3. Feature Selection**

I have used Recursive feature Elimination which is basically a backward elimination method to select 10 most relevant and non-redundant features.

I have used Random Forest as the base model for this due to its interesting feature of feature\_importances. The RFE recursively ranks each feature and remove the least important feature each time and continues this process until the specified number of features are achieved.

**4. Evaluating ML models**

I implemented a variety of individual classifiers; Logistic regression, random forest, SVM, kNN, Naive Bayesian to do the classification.

I split the dataset into 80% training and 20% testing set to evaluate the models and in order to find the optimal values for my models I have used GridSearchCV with 5-fold cross validation.

Also, for the evaluation metric, since my dataset was imbalanced, accuracy seemed not useful to make a meaningful prediction as the model can be biased towards majority class which is the Non-defaulter’s class.

I wanted a model that provides high recall rate but also a fair amount of precision. Since low precision indicates high False Positive rate which I do not want as it will be unfair towards the customers who paid their bills on time to be predicted as defaulters. I also evaluated the ROC-AUC curve, which tells us how good a model is in distinguishing between classes. Higher the area under curve, the better is the model in predicting defaulters as defaulters and non-defaulters as non-defaulters.

**4. Conclusion**

After analyzing the recall rate for these models, Logistic and Random forest models provided a high recall with fair precision. SVM and Naïve Bayesian performed pretty much the same.

Observing the ROC-AUC curve, Random forest stands out and has a better performance out of 5 models with an AUC score of 78% followed by Naive Bayesian and Logistic regression and then SVM.

However, through my results, I found that knn was not a good choice for my data as it gave a low recall rate also, AUC score for knn was low.

One thing I noticed that reducing the features did not show a major difference in the metrics results, but it helped to improve the training time, making the processing faster, so I decided to go with reduced feature set.