Capstone project 3

on

Credit card default prediction

By

Shubhangi Dharmik

(Individual)

Procedure

1. Introduction
2. Data summary
3. Factors
4. Exploratory data analysis
5. Modeling preparation
6. Algorithm used
7. Model performance
8. Hyperparameter tunning
9. Conclusions

Introduction

- In this analysis we used the dataset that consists of 30,000 credit card usage records and 3 machine learning models - Logistic Regression, Random Forest and ADA-Boost.
- Credit risk has traditionally the greatest risk among all the risks that the banking and credit card industries are facing.
- The main objective of this project is to conduct quantitative analysis on credit card default by using the machine learning models with accessible customer data.

Data summary

- The dataset contain following columns :
- X1: Amount of credit for both the individual and his/her family credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1-graduate school, 2-university, 3-high school, 4-others)
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6 = the repayment status in September
- X7 = the repayment status in August,
- X11 = the repayment status. 1-pay duly, 1-payment delay for 1 month, 2-payment delay for 2 months, likewise 9- payment delay for 9months & above.
- X12 to X17 = amount of bill statement in various months
- X18 to X23 = amount paid in various months

Factors

Following are the factors affecting to the default payment next month or our target variable:

- Gender
- Education
- Age
- Marital Status
- Credit Limit
- Inactive Customers

Exploratory data analysis

After loading and reading the dataset in notebook, we performed EDA. Comparing target variable which is default payment next month with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables and also we observed the distribution of variables. It gave us a better idea that how feature behaves with the target variable.

Modeling preparation

We define this as supervised machine learning for

better model performance.

- **Feature selection** : There are 25 columns in this dataset and the target variable is the column is default payment next month. We drop the column ‘ID’ and target variable
- **Imbalance data :** Imbalanced dataset will mislead machine learning algorithms and affect their performances so then we apply train test split to balance data.

Split Training and Test Data :

In the train test split we take two variables i.e. X and Y where X contain all the independent variables and Y contain dependent variable.

For the model, we use the ratio for training and test data split by 80% for training, 20% for test to ensure consistency.

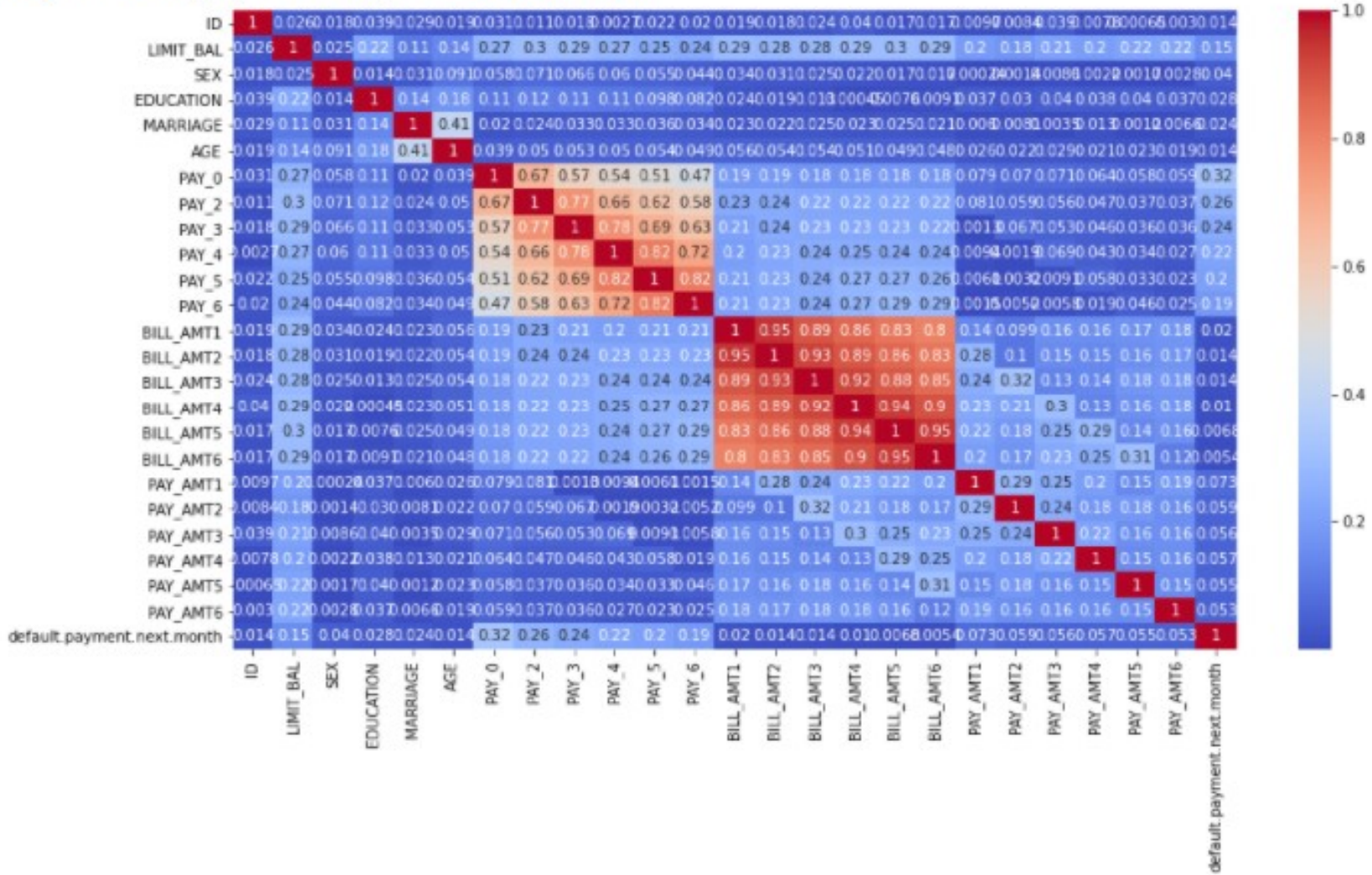
Correlation Analysis :

- We plot the heatmap to find the correlation between both dependent variable and independent variables.

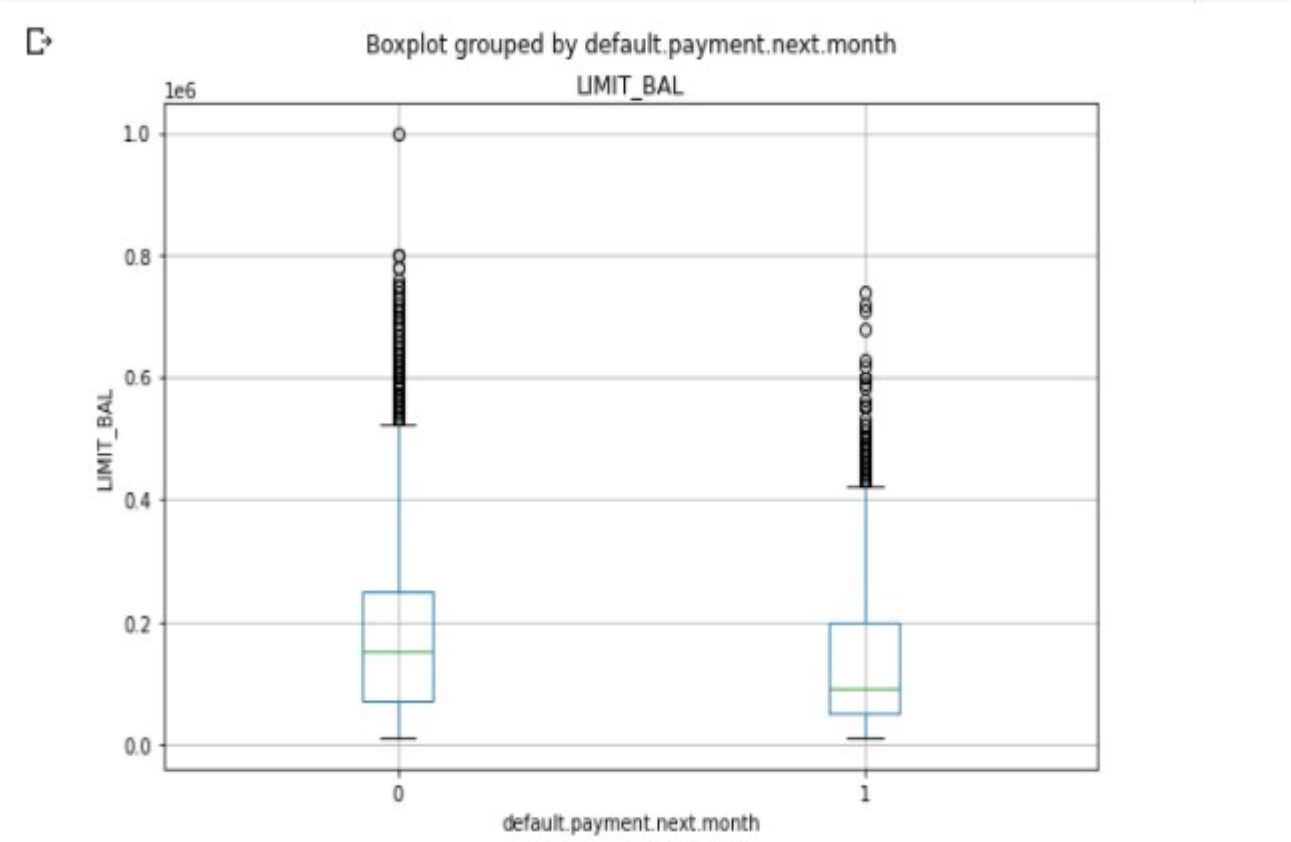
From the heatmap we observed that :

- Age, education, marital status, credit limit, gender is correlated to the default payment next month i.e. target variable.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa50035d7d0>
```



Correlation of default payment next month with limit credit



Algorithm used

We used the following algorithm to predict the value and for calculating the results. For finding out the result of predictive accuracy of the estimated probability of default.

- **Logistic Regression:**

  Logistic Regression is actually a classification algorithm that was given the name regression due to the fact that the mathematical formulation is very similar to linear regression.

  The function used in Logistic Regression is sigmoid function or the logistic function given by:

  $$f(x)= 1/1+e \text{ }^(-x)$$

- **Random Forest Classifier:**

  Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.

Model performance

- **Confusion Matrix**:

  The confusion matrix is a table that summarizes how successful the classification model is at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

- **Precision/Recall**:

  Precision is the ratio of correct positive predictions to the overall number of positive predictions i.e. TP/TP+FP

  Recall is the ratio of correct positive predictions to the overall number of positive examples in the set i.e. TP/FN+TP

- **Accuracy:**

  Accuracy is given by the number of correctly classified examples divided by the total number of classified examples. In terms of the confusion matrix, it is given by TP+TN/TP+TN+FP+FN

Hyperparameter tunning

Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used following hyperparameter tuning. This also results in cross validation and in our case we divided the dataset into different folds.

- **Grid Search CV**
- **Randomized Search CV**

Results

- **Logistic Regression :**

  Accuracy - 79%

  Precision/Recall – 1.00/0.79

  roc-auc score – 0.65

- **Random forest classifier :**

  Accuracy - 82% - 83%

  roc-auc score – 0.66

- **ADA-Boost:**

  Accuracy – 81.61%

Conclusions

- In Logistic Regression model if model has the highest recall but the lowest precision, if the firm expects high recall, then this model is the best candidate.
- If the balance of recall and precision is the most important metric, then Random Forest is the ideal model.
- Consider the applicants marital status. Married people seem to default more often.

- Consider the age of the applicant. Younger people are at higher risk of defaulting.
- We can say that random forest model gives us more accurate model instead of remaining ones.
- Our best prediction accuracy was around 82-83%, our lowest measured prediction accuracy was about 79%.

Thank You