ROUND 2 Epsilon – Case Study

Piyush Raj Gupta **Team – Parag** ***Bhushan Chaudhari***

College – Indian Institute Of Management, Calcutta (PGDBA)

To approach the problem of defaulter prediction we had used various sophisticated machine learning algorithm and statistical methods. The methods which we had used is successfully giving the results of around 96% of accuracy. Our methodology and suggestion are discussed in subsequent report and code base is attached in folder.

We had approach this problem in following sub-steps.

Step 1 – Data Exploration

Step 2 – Data Pre-Processing

- Step 2.1 – Outlier Detection
- Step 2.2 – Outlier Ranking
- Step 2.3 – Outlier Removal
- Step 2.4 – Imputations Removal
- Step 2.5 – Splitting Training & Test Datasets
- Step 2.6 – Balancing Training Dataset

Step 3 – Exploratory Data Analysis

Step 4 – Features Selection and model fitting

- Step 4.1 – Correlation Analysis of Features
- Step 4.2 – Ranking Features
- Step 4.3 – Feature Selection

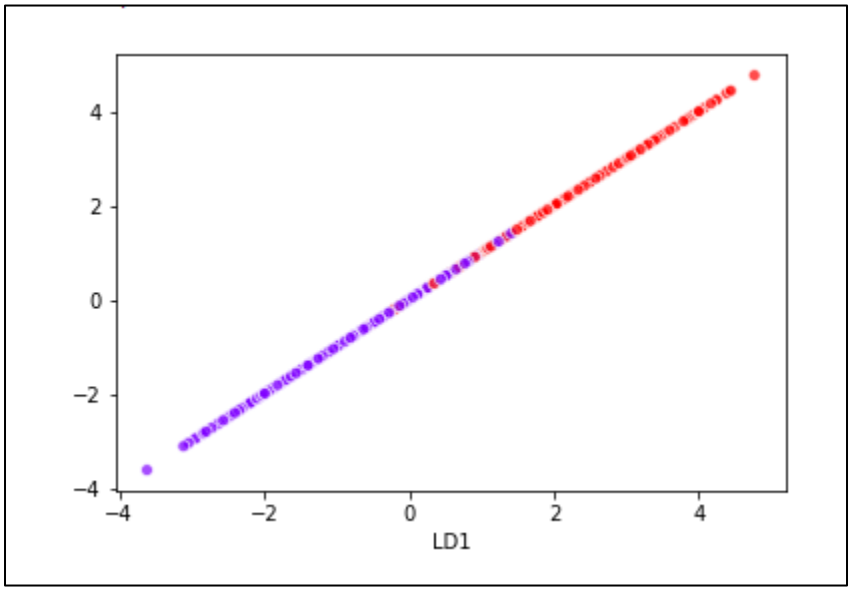
Step 5 – Building Classification Model

Step 6 – Predicting Class Labels of Test Dataset

Step 7 – Evaluating Predictions

Data Preprocessing – After exploring data it had been observed that there are few missing values in the data. We had treated the missing values by filling them with median in case of features having continuous values like Amount, Credit score, Term etc. and with mode in case of categorical features with mode of corresponding feature. Label Encoding is done using sklearn library to convert categorical variables into numerical variables. We had split data into split of 85:15 training is to validation set . As the available data is not too much we haven't removed any outliers.

Exploratory Data Analysis - We have used linear discriminant analysis technique to check the seperability of class **.** However after doing LDA it has been observed that all the variance of data had been explained by only one variable and hence we came to know that there is large correlation exists between given variables. We had used this fact in feature selection method and hence reduced the features in order to achieve high accuracy.



LDA Classifier Red- Defaulter and blue non defaulter

In above figure we can clearly see that there is linear boundary between two classes and hence we can use LDA classifier effectively.

Feature Selection and Model Fitting – We had used various probabilistic machine learning classifiers in order to predict the defaulters as accurately as possible. We had tried different models and finally it has been observed that LDA classifier had worked best for this classification problem.We had used following Models :

1) XGBoost Classifier
2) Suppory Vector Classifier
3) Naïve Bayes Probabilistic Model
4) Random Forest Model
5) LDA Classifier

Classification of reports for Each of Model is attached below :

```
              precision    recall  f1-score   support

         0.0       0.97      0.94      0.95        88
         1.0       0.86      0.91      0.88        33

   micro avg       0.93      0.93      0.93       121
   macro avg       0.91      0.93      0.92       121
weighted avg       0.94      0.93      0.93       121

[[83  5]
 [ 3 30]]
```

Support Vector Classifier

```
0.9338842975206612
              precision    recall  f1-score   support

         0.0       0.97      0.94      0.95        88
         1.0       0.86      0.91      0.88        33

   micro avg       0.93      0.93      0.93       121
   macro avg       0.91      0.93      0.92       121
weighted avg       0.94      0.93      0.93       121

[[83  5]
 [ 3 30]]
```

XGBoost

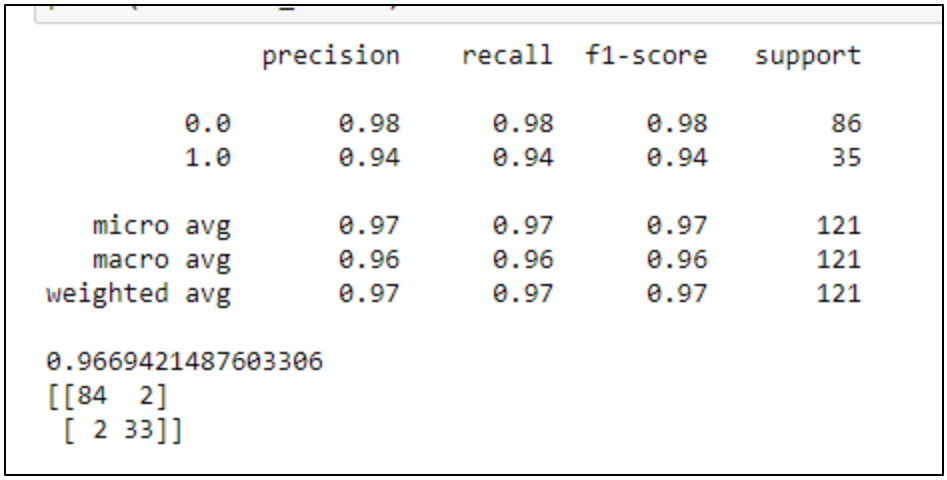
```
              precision    recall  f1-score   support

         0.0       0.98      0.97      0.97        87
         1.0       0.91      0.94      0.93        34

   micro avg       0.96      0.96      0.96       121
   macro avg       0.95      0.95      0.95       121
weighted avg       0.96      0.96      0.96       121

0.9586776859504132
[[84  3]
 [ 2 32]]
```

RandomForest



	precision	recall	f1-score	support
0.0	0.98	0.98	0.98	86
1.0	0.94	0.94	0.94	35
micro avg	0.97	0.97	0.97	121
macro avg	0.96	0.96	0.96	121
weighted avg	0.97	0.97	0.97	121

0.9669421487603306

```
[[84  2]
 [ 2 33]]
```

LDA Classifier

It is evident from the above reports that LDA classifier is working best on validation data and hence we had proceed with this algorithm to predict the defaulters.

Hence finally we had train our model on full train dataset by using above mentioned features and Random forest classifier.

Feature Engineering - After applying various feature selection techniques such as forward selection, Backward selection and proper subset selection we came to conclusion that there are only 4 features which should be selected to extract maximum accuracy which are listed as :

- Age
- Checking Amount
- Credit Score
- Saving Amount

As we have very less number of important feature it is not appropriate to fit complex models to data as it can result into overfitting of data. Hence we had selected LDA classifier to apply on this dataset.

Suggestions – As mentioned above there are only 4 important features which influence the decision variable. In today's generation there is abundance of data available hence we can incorporate more features like applicant's income, education level, property details etc so that we can apply more complex machine learning algorithms to improve accuracy further.

Supporting python code and jupyter notebook is attached in the folder.

Team - Parag

Piyush Raj Gupta
PGDBA (2019-21)
IIM Calcutta | ISI Kolkata | IIT Kharagpur
+91-8381872655 | piyushrba2021@email.iimcal.ac.in

Bhushan Chaudhari
PGDBA (2019-21)
IIM Calcutta | ISI Kolkata | IIT Kharagpur
+91-7030407502 | chaudharibba2021@email.iimcal.ac.in