Credit Risk Model

Anubhab Bose
Soumya Ghatak

July 31, 2022

1 Introduction:

Banks play a crucial role in market economies. They decide who can get finance and on what terms and can make or break investment decisions. For markets and society to function, individuals and companies need access to credit.

Credit scoring algorithms, which make a guess at the probability of default, are the method banks use to determine whether or not a loan should be granted. This competition requires participants to improve on the state of the art in credit scoring, by predicting the probability that somebody will experience financial distress in the next two years.

The goal of this project is to build a model that borrowers can use to help make the best financial decisions.

Historical data are provided on 150,000 borrowers .

2 Dataset Variables:

1) SeriousDlqin2yr:Person experienced 90 days past due delinquency or worse

2) RevolvingUtilizationOfUnsecuredLines :Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits

3) NumberOfTime30-59DaysPastDueNotWorse :Number of times borrower has been 30-59 days past due but no worse in the last 2 years.

4) DebtRatio :Monthly debt payments, alimony,living costs divided by monthy gross income

5) MonthlyIncome :Monthly income

6) NumberOfOpenCreditLinesAndLoans :Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)

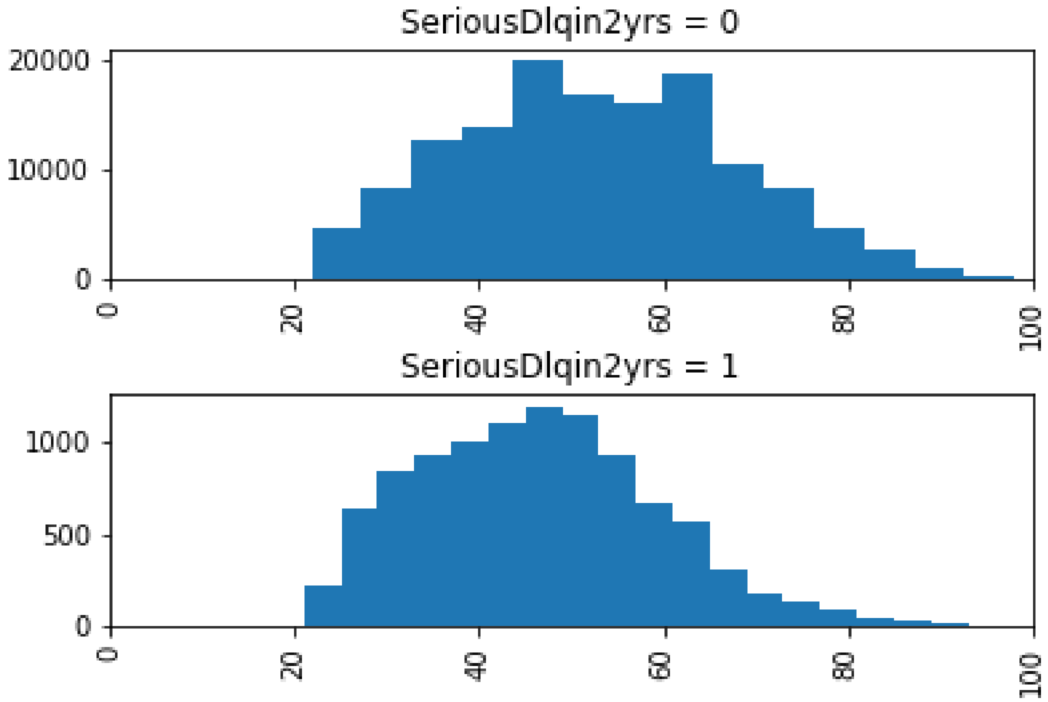
7) NumberOfTimes90DaysLate :Number of times borrower has been 90 days or more past due.

8)NumberRealEstateLoansOrLines :Number of mortgage and real estate loans including home equity lines of credit

9) NumberOfTime60-89DaysPastDueNotWorse :Number of times borrower has been 60-89 days past due but no worse in the last 2 years.

10)NumberOfDependents :Number of dependents in faIn our original dataset, we have 6.684

3 Exploratory Data Analysis And Data Pre-Processing:



- Generally, younger people were more responsible for defaulting than older people as evident from the 2nd histogram.

- 2.5 percent of the persons that is roughly 4 lakh people have debt ratio over 3489. From the figures concerning Debt Ratio, it is evident that there are many outliers. So, we replace values (nearly 20.85 percent of the data set) outside the 3rd quartile+1.5* IQR(1.908) with that particular value.

- The 'NA' values in the 'MonthlyIncome' are repalced by 0 as 'MonthlyIncome' will be 0 in the worst case scenario. This will make our model more robust.

- 'NA' values in 'NumberofDependents' are replaced by the median of the observations rounded off to the nearest integer as the observations are positively skewed.

- In our original dataset, we have 6.684 percent defaulters. This is an imbalanced data set. So, we smote the data set for building a model to identify defaulters. In the smoted dataset, the percent of defaulters has been raised to 48.35 percent. Also, total number of samples have been increased to 216744 from 150000.

We divide the dataset into training set and test set in 80:20 fashion and build our supervised classification models on the training set and validate our results using the AUC score on the test set.

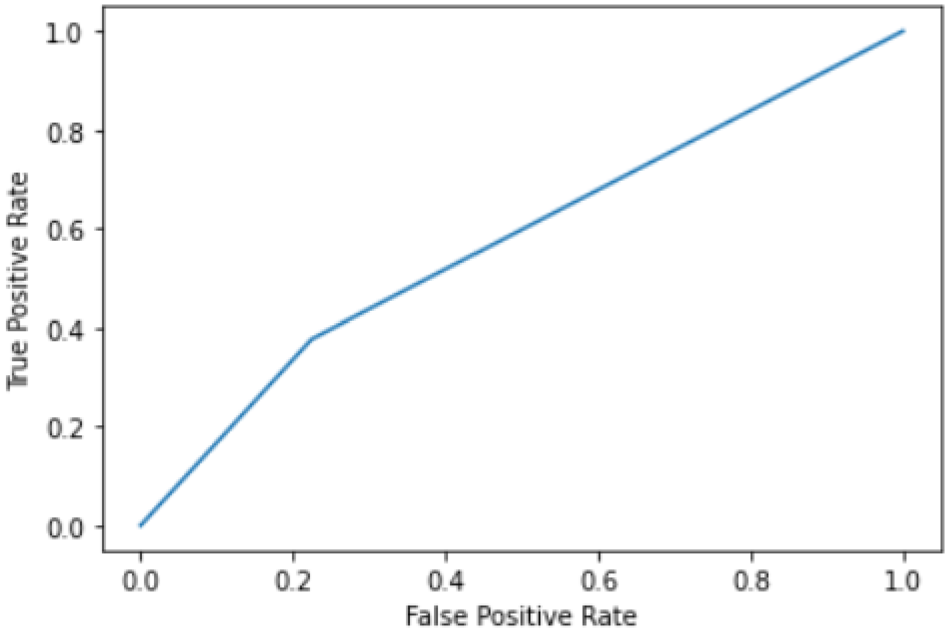
4 Supervised Models:

1. **Random Forest Classifier:**We built a random forest classifier with 500 decision trees built using the bootstrap method and we considered Gini's Coefficient for measuring the gain in information. The hyper parameters yielding the best AUC were chosen by Grid Search Cross

Validation Technique. The performance of the classifier on the test set was observed as below:

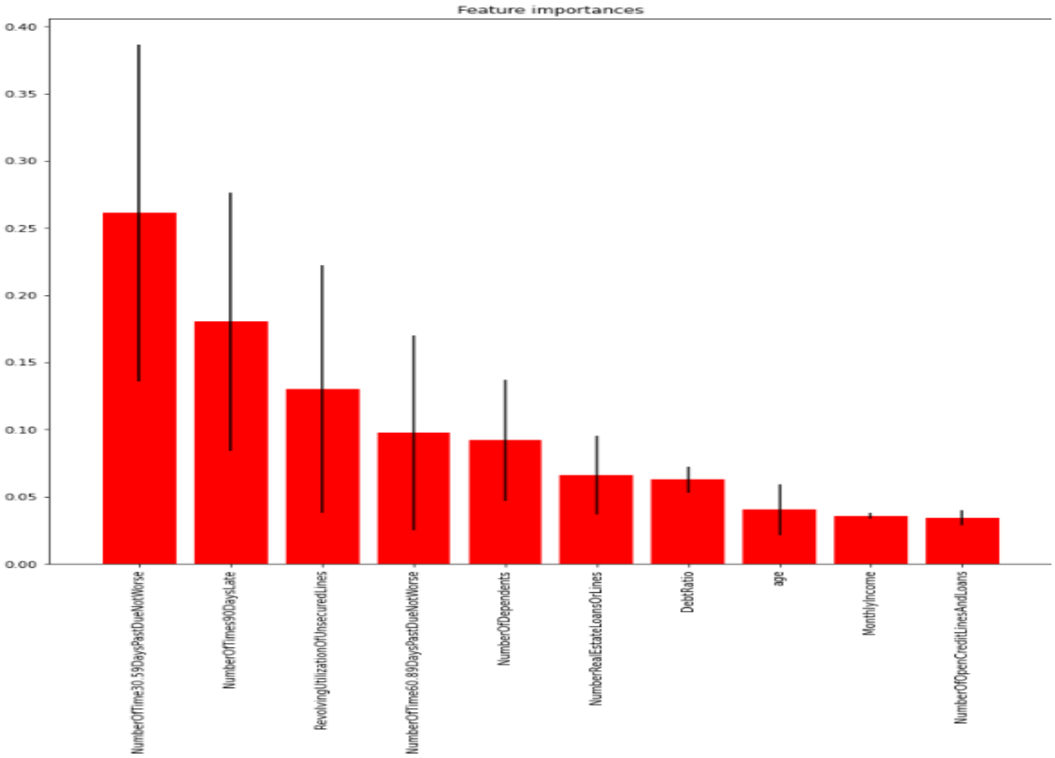
Report	Precision	Recall	f1-Score	support
0	0.95	0.78	0.85	28036
1	0.11	0.38	0.16	1964

The ROC Curve along with the AUC value is presentes as below:



AUC:0.576
Accuracy:0.75

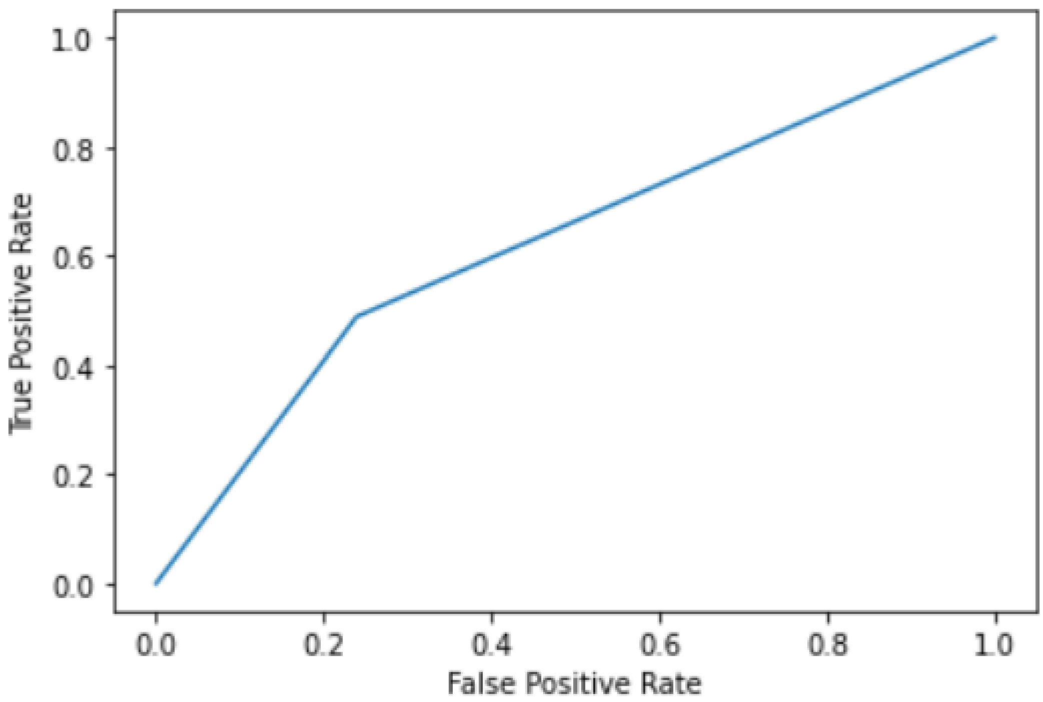
The importance of the features from random forest model are observed as below:



The three least significant features namely "Age","MonthlyIncome","NumberOfOpenCreditLinesAndLoans" are removed from the dataset and another random forest model is bult on the modified data set with the same hyper parameters as mentioned above.The performance of the modified classifier on the test set was observed as below:

Report	Precision	Recall	f1-Score	support
0	0.95	0.76	0.85	28036
1	0.13	0.49	0.20	1964

The ROC Curve along with the AUC value is presentes as below:

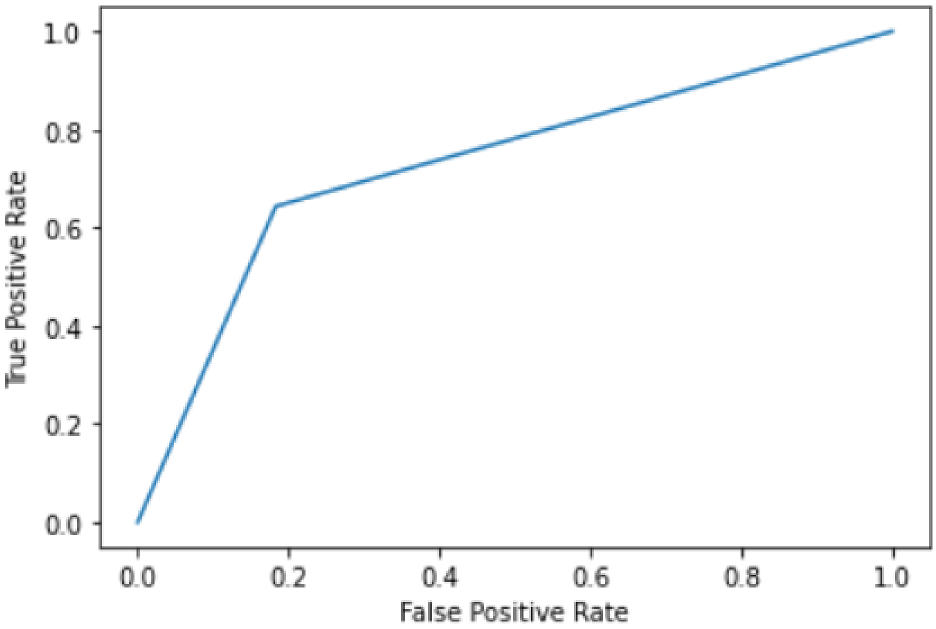


AUC:0.624
Accuracy:0.74

2. **Logistic Classifier**: We standartized the training data and built a logistic classifier model on the standartized training set. The threshold for classification was set as 0.7. The performance of the logistic classifier on the test set was observed as below:

Report	Precision	Recall	f1-Score	support
0	0.97	0.82	0.89	28036
1	0.20	0.64	0.30	1964

The ROC Curve along with the AUC value is presentes as below:

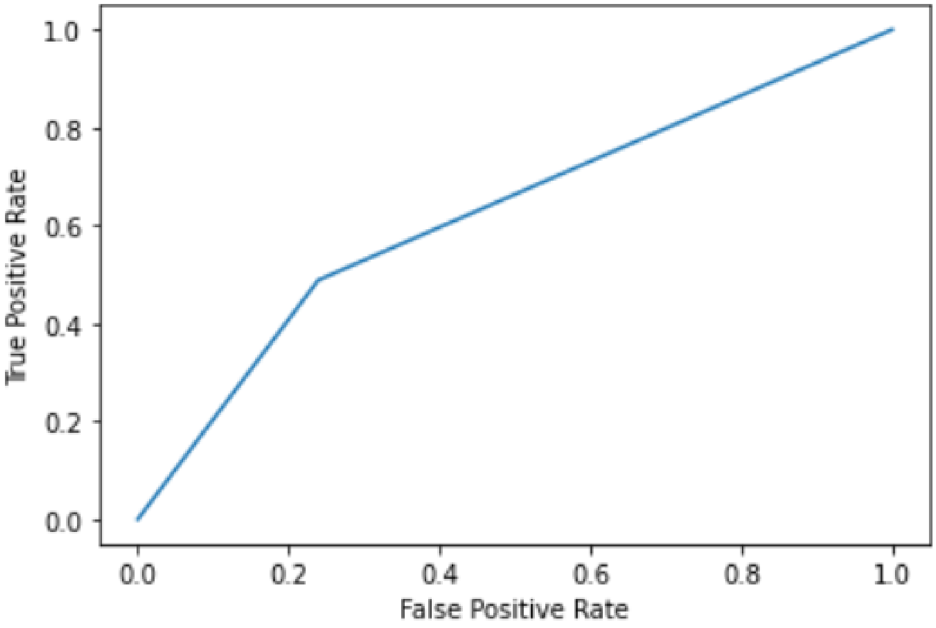


AUC:0.73
Accuracy:0.81

3. **Decision Tree Classifier:**We built decision tree classifier with maximum depth,10 and minimum samples per leaf, 5 and we considered Gini's Coefficient for measuring the gain in information. The hyper parameters yielding the best AUC were chosen by Grid Search Cross Validation Technique. The performance of the decision tree classifier on the test set was observed as below:

Report	Precision	Recall	f1-Score	support
0	0.95	0.78	0.85	28036
1	0.11	0.39	0.17	1964

The ROC Curve along with the AUC value is presentes as below:

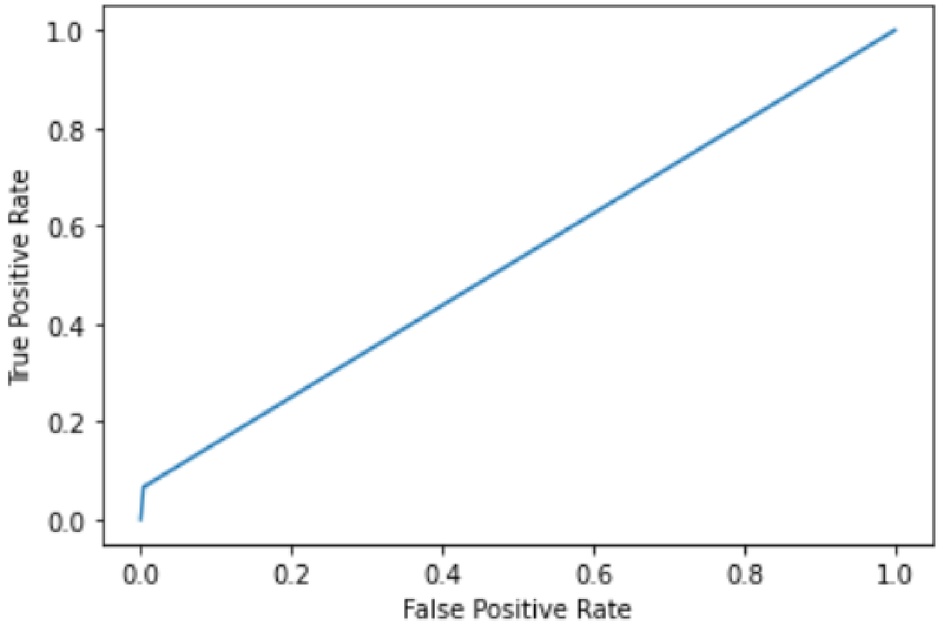


AUC:0.624
Accuracy:0.75

4. **Naive Bayes Classifier:** We built a Naive Bayes Classifier to classify defaulters accurately. The performance of the Naive Bayes classifier on the test set was observed as below:

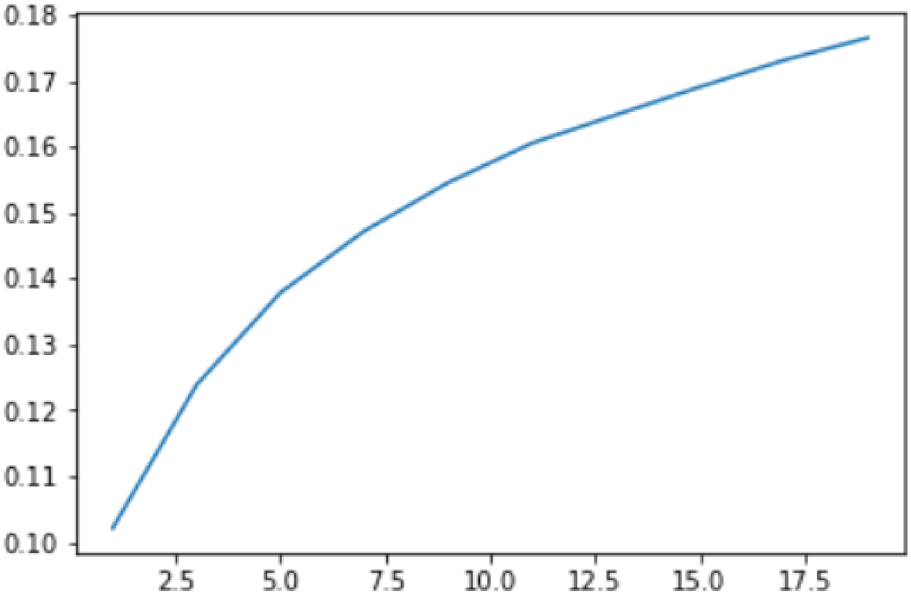
Report	Precision	Recall	f1-Score	support
0	0.94	0.18	0.97	28036
1	0.55	0.07	0.12	1964

The ROC Curve along with the AUC value is presentes as below:



AUC:0.531
Accuracy:0.94

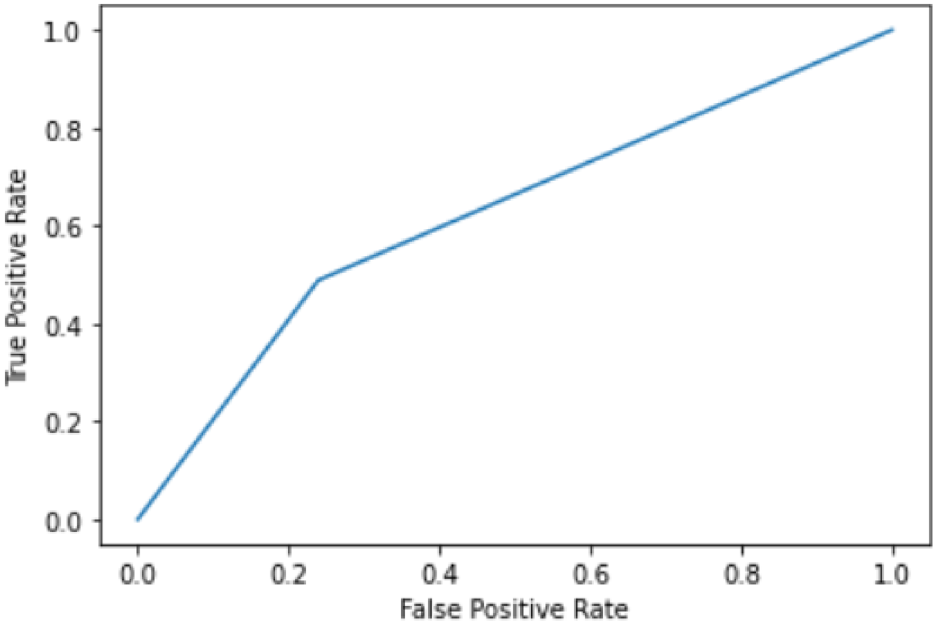
5. **KNN Classifier:** We plotted the miss classification error for different values of neighbours.



From the plot, it is evident that as number of neighbours increase, miss classification error also increses. So, we built KNN 1 classifier. The performance of the Naive Bayes classifier on the test set was observed as below:

Report	Precision	Recall	f1-Score	support
0	0.96	0.68	0.80	28036
1	0.11	0.56	0.18	1964

The ROC Curve along with the AUC value is presentes as below:



AUC:0.624
Accuracy:0.68

5 Conclusion:

All the supervised classification models with their corresponding AUC scores are listed as below:

Classification Model	AUC
Random Forest	0.576
Modified Random Forest	0.624
Logistic	0.73
Decision Tree	0.624
Naive Bayes	0.531
KNN	0.624

The AUC value of the Logistic Classifier is 0.73 which is the highest among all the supervised classifiers considered.Also, the f1-Score for Defaulters in the Logistic model was 0.64 which suggests that the Logistic Classifier classifies the actual defaulters in the training set with 64 percent success rate and this success rate is the highest among all the other classifiers. So the Logistic Classifier should be used to predict a borrower as defaulter or not based on the portfolio.

6 Refernce:

1. **Code:** Github Code
2. **Dataset:** Github Dataset
3. **Kaggle Competition:** Kaggle Link