MS4610 Introduction to Data Analytics

Project Report

Loan Default Prediction

Machine learning model development for classification

submission by

Group 12

CE17B026 Ashish Soni

CE17B037 Jai Kedia

CE17B057 Shrirang Vaidya

ME17B018 Kaivalya Rakesh Chitre

ME17B071 Surjeet Kumar Verma

INDIAN INSTITUTE OF TECHNOLOGY MADRAS
CHENNAI 600036, INDIA.

July-Nov 2020

1 Abstract

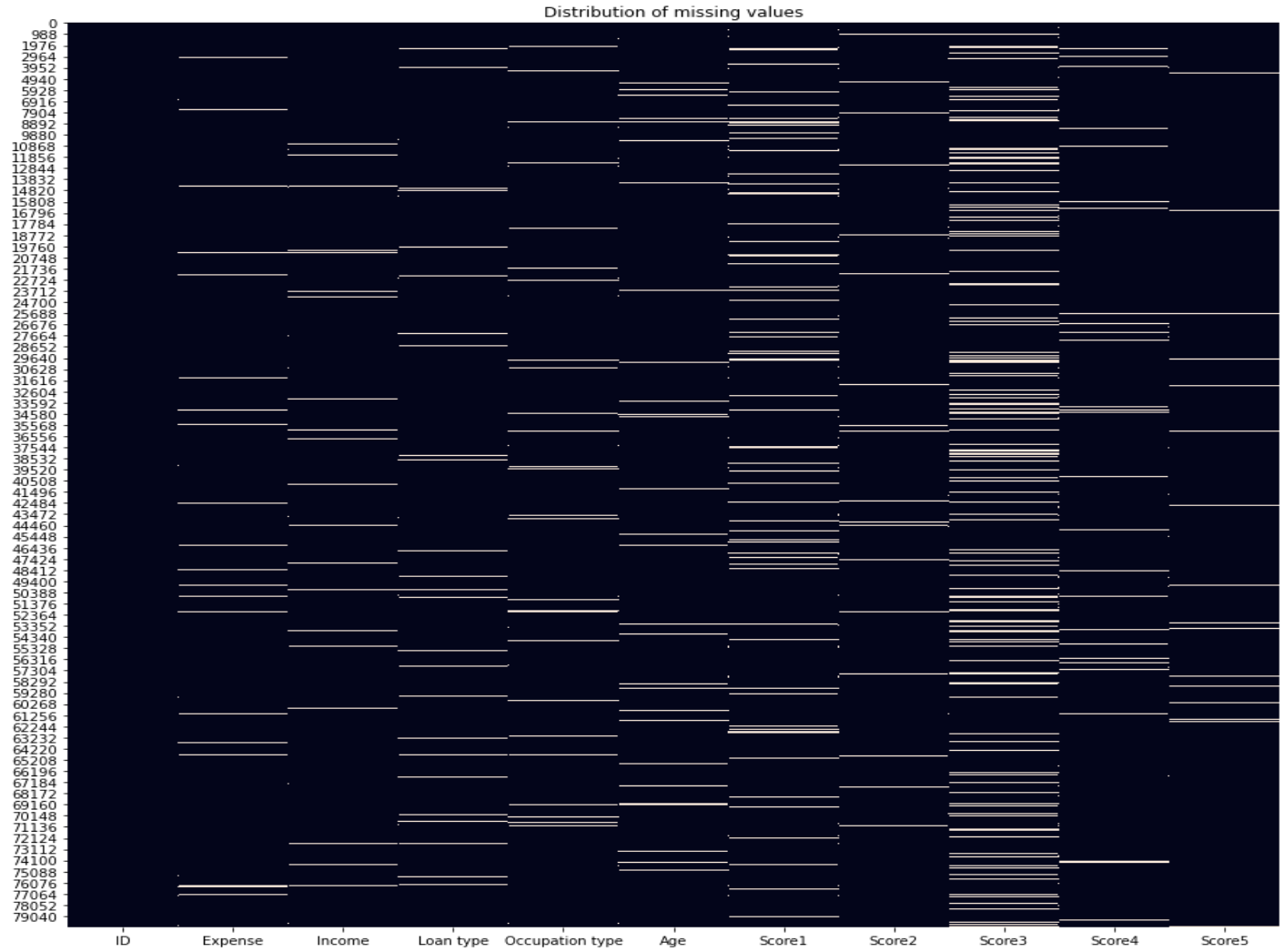
We have analysed a small database of customers and built a classification model to predict whether a loan will go default or not.
In our Analysis: We first pre-processed the data (like handling missing values, using categorical data…) to make them suitable for our machine learning models. Then we have done some Exploratory data analysis to understand the data better. Finally, we trained a model on the dataset and predict the test_y for the test data.

2 Data Pre-processing

2.1 Missing Values:

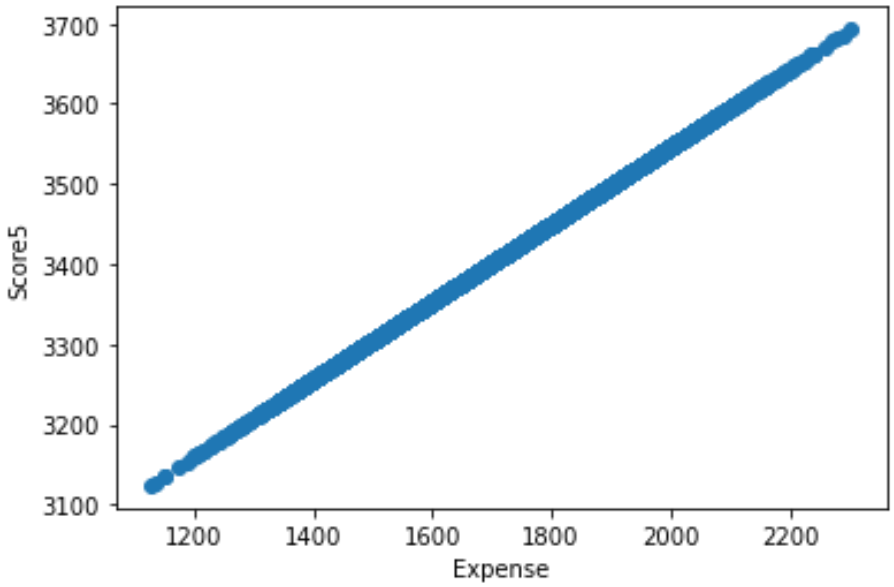
This is the most visible problem the dataset faces. Most models cannot deal with missing values directly, which calls for imputation. However, imputation has some potential drawbacks.
Values added by imputation methods can be very different from actual data, adding noise and unnecessary variation to the data. This can be detrimental to our classifier's performance.
Generally, imputing data in columns with more than 20% of their data missing is considered unhealthy.

2.1.1 Heat-Map of missing values of our dataset:



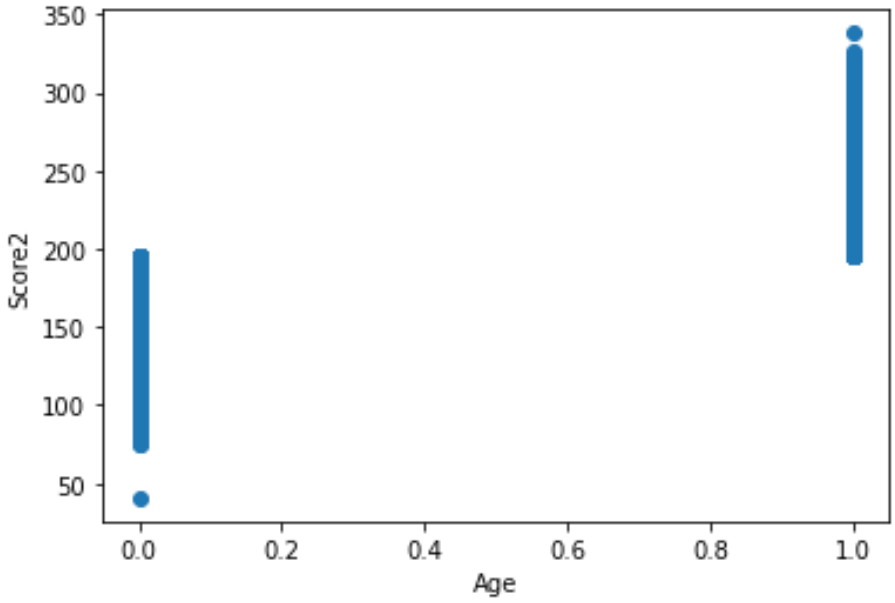
2.1.2 Handling Numerical value

- **Zero and garbage imputation.** Here, some missing values are replaced with zeros while some others are replaced with unlikely values like -9999. This helps the classifier recognize anomalies in the data, while also filling in for absent information.
- **Relation between *Score5* and *Expense***



Score5 and expense showed linear relationship. We used this equation to fill the missing values.

- **Relation between *Score2* and *Age***



Score2 and Age showed an interesting relation.
If score2>194.9 then age=1 else age=0. We used this relation to fill missing Age values.

2.1.3 Handling Categorical Data:

XGBoost model can handle missing categorical value so we did not replace them.

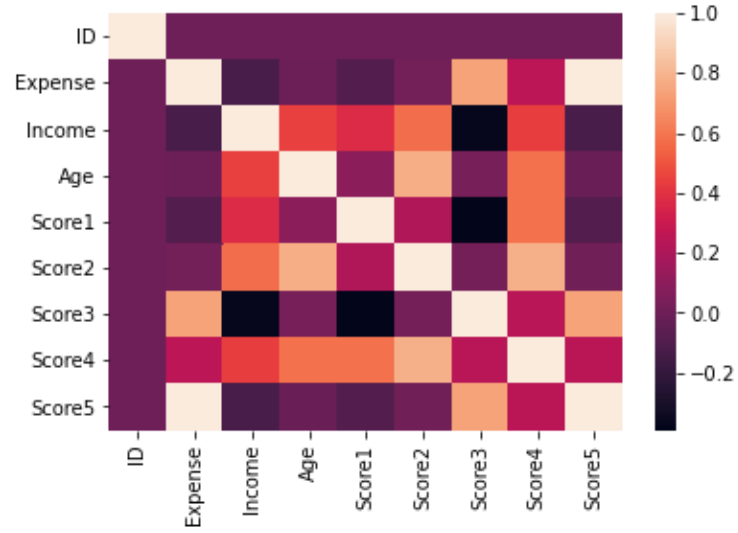
2.2 One-Hot encoding

We performed One-Hot encoding on categorical variables *'Loan type'* and *'Occupation type'*

3 Exploratory Data Analysis

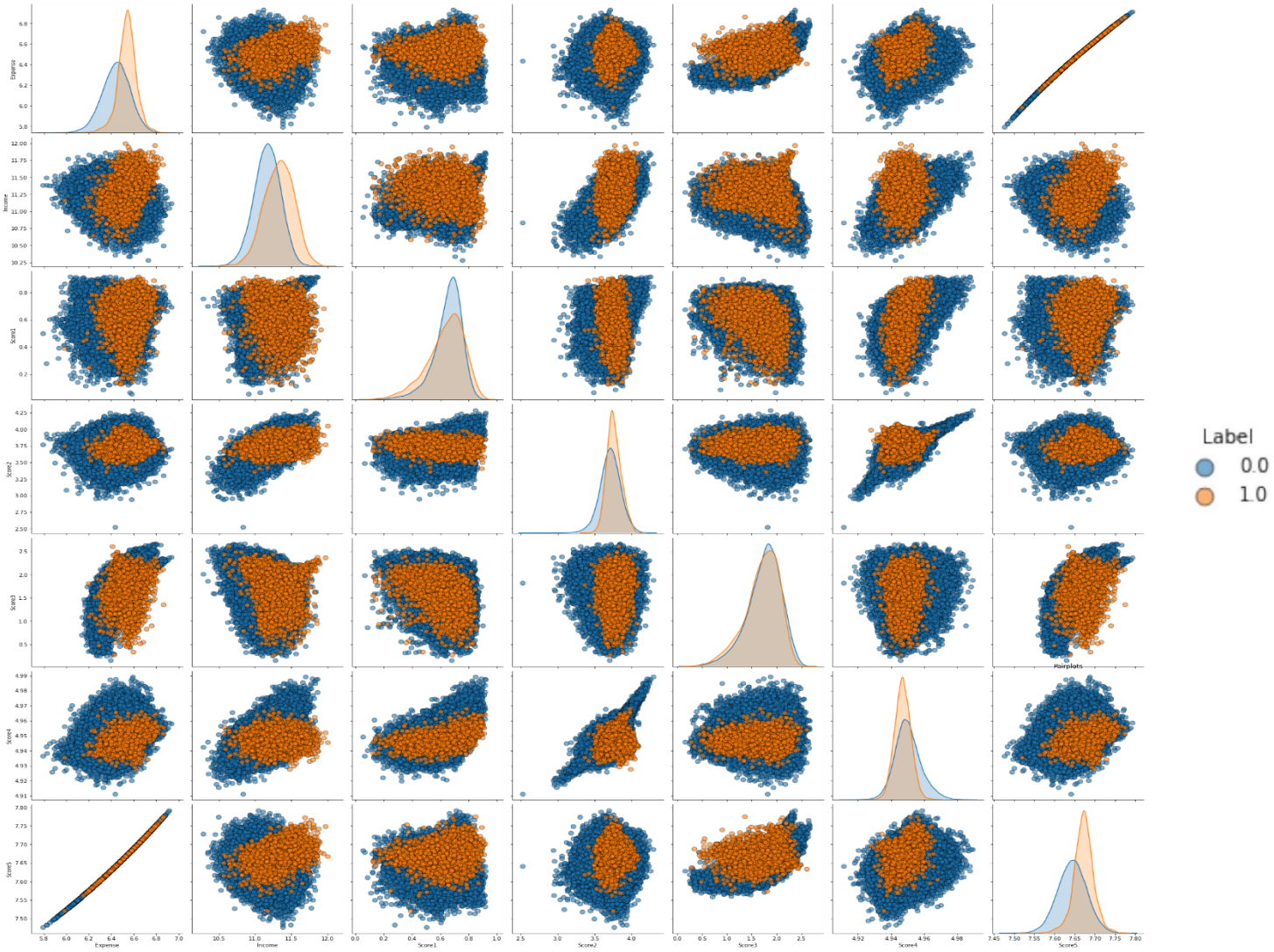
We have created several plots to understand the dataset.

Correlation Heatmap

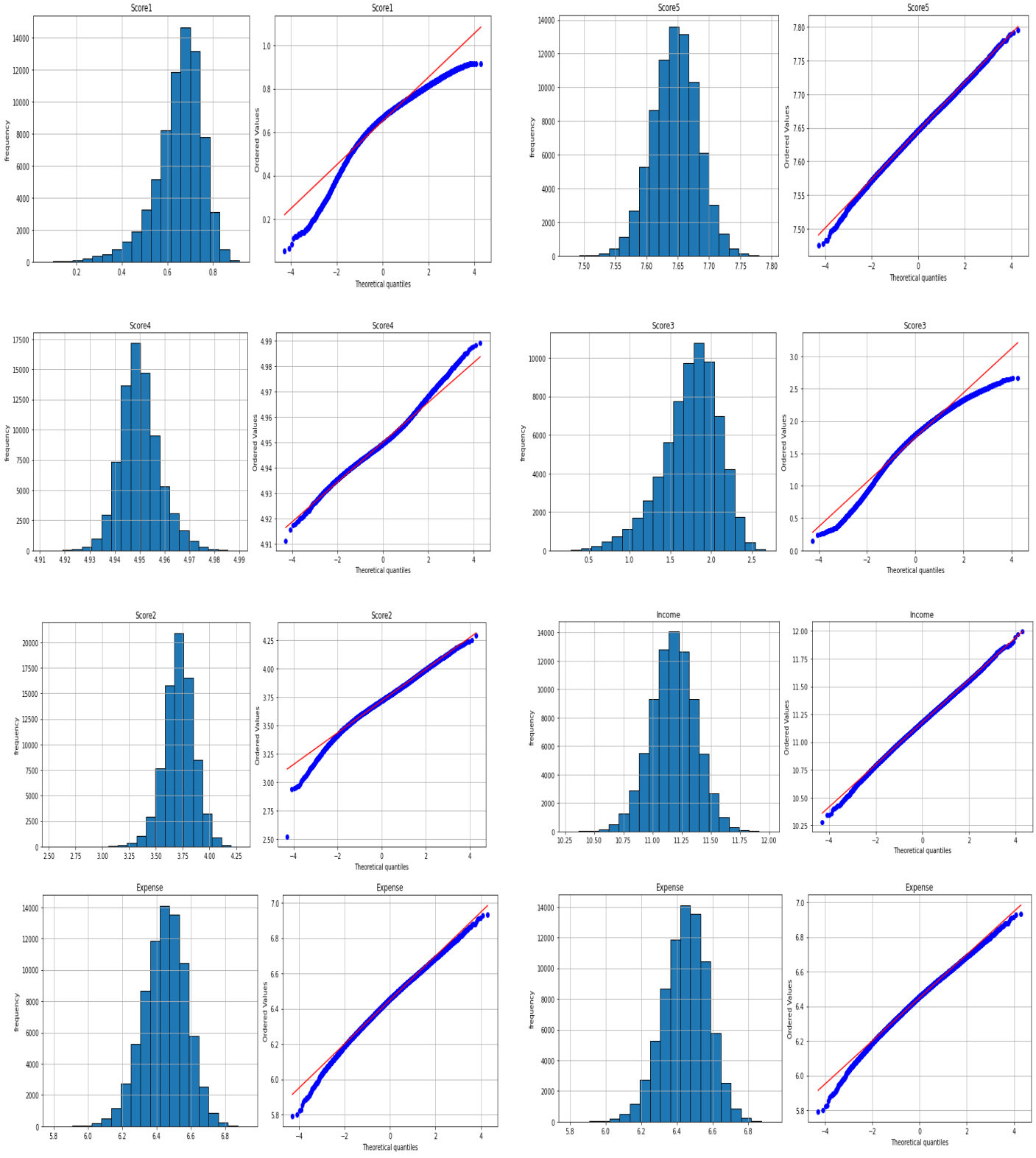


This heat map shows linear correlation between Expense and score5

Pair-plot between *'Expense','Income','Score1','Score2',"Score3",'Score4','Score5'*



Data distributions and Prob-plots

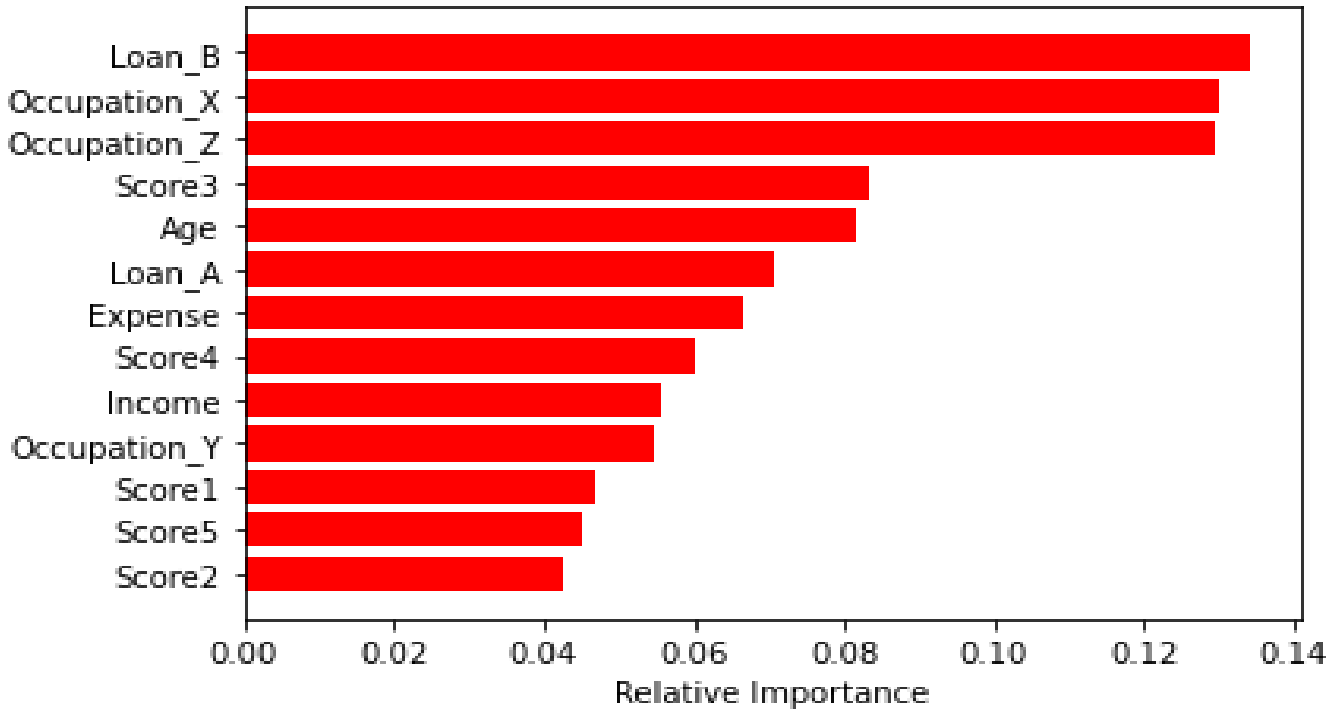


These plot shows the distributions of all the numerical variables

Feature Importance:

Finally, we adopt a more direct approach in determining the predictors important for our model: by asking our model which predictors helped it the most. Since our submission is based off a gradient boosted classification tree, we train a XGboostclassifier on the data and extract feature importances from the trained model using it model.feature_importance_ attribute.

Feature importance plot:

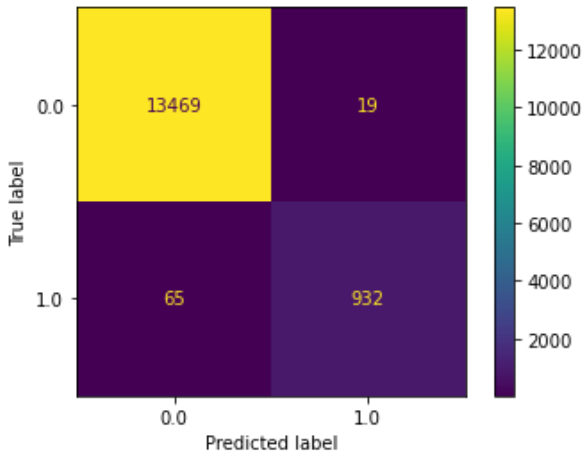


The variables Loan_b, occupation_x ,occupation_z leave the other variables behind in importance by a huge margin. These, in a way, are the most important factors in deciding whether a loan will be default or not. (for this dataset).

4 Model Selection

In the final section we tried out different models. Best score was generated by XGBoost, lightGBM and random forest. We decided to go ahead with XGBoost because of its highest f1-score.

We have trained a xgboost classifier with a 5-fold cross-validation to optimize the F1 score using randomized search for hyperparameter optimization. The final confusion matrix and classification report for test split is as follow



	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	13534
1.0	0.93	0.98	0.96	951
accuracy			0.99	14485
macro avg	0.97	0.99	0.98	14485
weighted avg	0.99	0.99	0.99	14485

5 Important Links

Our complete work has been added in a GitHub repository, whose link we have provided below.

- Project link
- Final Notebook link
- EDA Notebook link

END OF REPORT