CREDIT RISK ANALYSIS

The assignment that was given was to determine whether a loan applicant is a higher risk applicant or not.
A higher risk applicant can be referred to a customer whose chances to not be able to repay the loan is very high.
An organization who is in the business of providing loans needs to know the chances of receiving the loan back.
As data scientists, we study the data of different loan applicants and train a model on that data, so that later on it can predict whether a certain applicant with a certain amount of data will be at risk or not.

TASK 1- EXPLORATORY DATA ANALYSIS

The first task was to explore the data that was provided to us. We had 2 sets of csv files, namely, loan and applicant. We merged the two datasets to make it easier to study the merged data.
The merged data provided us with the following inferences-
1.Higher risk of not getting back the loan in females than in males because according to the visualization.(more than 50% of females have higher_risk)
2.Separated and divorced people also have higher risk.
3.People who live for free or on rent have higher risk.
4.People who are unskilled-non-resident/unemployed have high risk.
5.People who are employed for 0 or 1 year have high risk.
6.People who have a savings account balance as 'low' or ' medium' are also high risk applicants.
7.People wanting loans for education,new vehicles and repair costs are high risk applicants.
8.People taking a loan for a car or other are high risk applicants.
9.The age column is right skewed.
10.People with age between 25-40 years are high risk applicants.

11.Foreigner workers are also high risk applicants.
12.Applicants who have loan duration less than 25 are at higher risk.
13.Applicants with high EMI rate have higher risk.
14.People who have no co-applicant or guarantor are at a higher risk.
15.High_risk_applicant has high positive correlation with months_loan_taken_for and Principal_loan_amount
16.The principal loan amount has a high positive correlation with months loan taken for as if the principal loan amount increases the time needed to repay it will also increase.

TASK 2-TRAINING THE MODEL

The two models that I have used to train this model are Logistic Regression and RandomForestClassifier.
These two models have been used because they are used in classification and predictions of discrete values(0 and 1) in this case.
The missing values were handled and taken care of in Task 1.
The categorical features are expanded in various columns using one-hot-coding.
The prediction of the models along with the test score,train score and confusion matrix are shown in the notebook.

I have not used hyperparameter tuning as I don't have enough knowledge of that, but will surely in a few days.