System Used:
MacBook Pro - 2.8 GHz Intel Core i7, 16 GB 2133 MHz LPDDR3

Preprocessing of Data:

1. Collapsed the target variable (loan_status) into two factor level – Default & Fully Paid by changing 'Charged Off' into 'Default'
2. Imputed the NA values to 999 for numerical predictors
3. Imputed NA values for categorical variables using the library call fct_explicit_na() from the library forcats. The formula used for this column_number*100
4. Removed the predictor 'title' from the data set before training (Column number 15)
5. Removed the predictor 'emp_title' from the data set before training
6. Removed the predictor 'zip_code' from the data set before training (Column number 15)
7. Removed the predictor 'earliest_cr_line' from the data set before training
8. Transformed the variable 'loan_status' into numeric value 0 & 1, instead of categorical 'Default' & 'Fully Paid'
9. Employed one hot encoding technique for handling the categorical variables.

Model Used:

Xgboost is used to train in 'binary:logistic' mode.
Trained the model by tuning the parametrs max_depth, eta, nthread and nrounds for different combination of values.
Best result achieved with values: max_depth = 5, eta = 0.2, nthread = 2, nrounds = 250

Accuracy:

Split the loan data into three train and test split by using the id from the file Project3_test_id.csv
Created 3 csv files with the corresponding true labels for accuracy measurement.

Code finally generates three submission files, one for each test id, used for the evaluation.

On testing the model on 3 generated train and test split achieved following accuracy:

	Test1	Test2	Test3	Average
Model(xgboost)	0.4406513	0.4423714	0.4410044	0.4413424

Runtime of the code:

17.79546 minutes