TECHNICAL REPORT ON PREDICTION OF LOAN GRADE USING MACHINE LEARNING TECHNIQUES

BY

DORTIMIARIYE MAXWELL ANGALABIRI

INTRODUCTION

Loan grading is an aspect of the credit screening and approval processes that is part of a lending institution's loan review or credit risk system. Loan grading is a categorization method that assigns a ranking or a designation to a loan based on the borrower's credit history, the quality of the collateral, and the possibility of principal and interest payments. Predicting the grade of a loan is an important activity conducted by lending institutions to determine the risk associated with a loan and the likelihood of repayment. This project explores the use of machine learning techniques in automating the process of loan grading.

PROJECT OBJECTIVE

The goal of this project is to carry out a detail exploratory data analysis and develop a predictive model to predict loan grades based on the information available before a loan is approved.

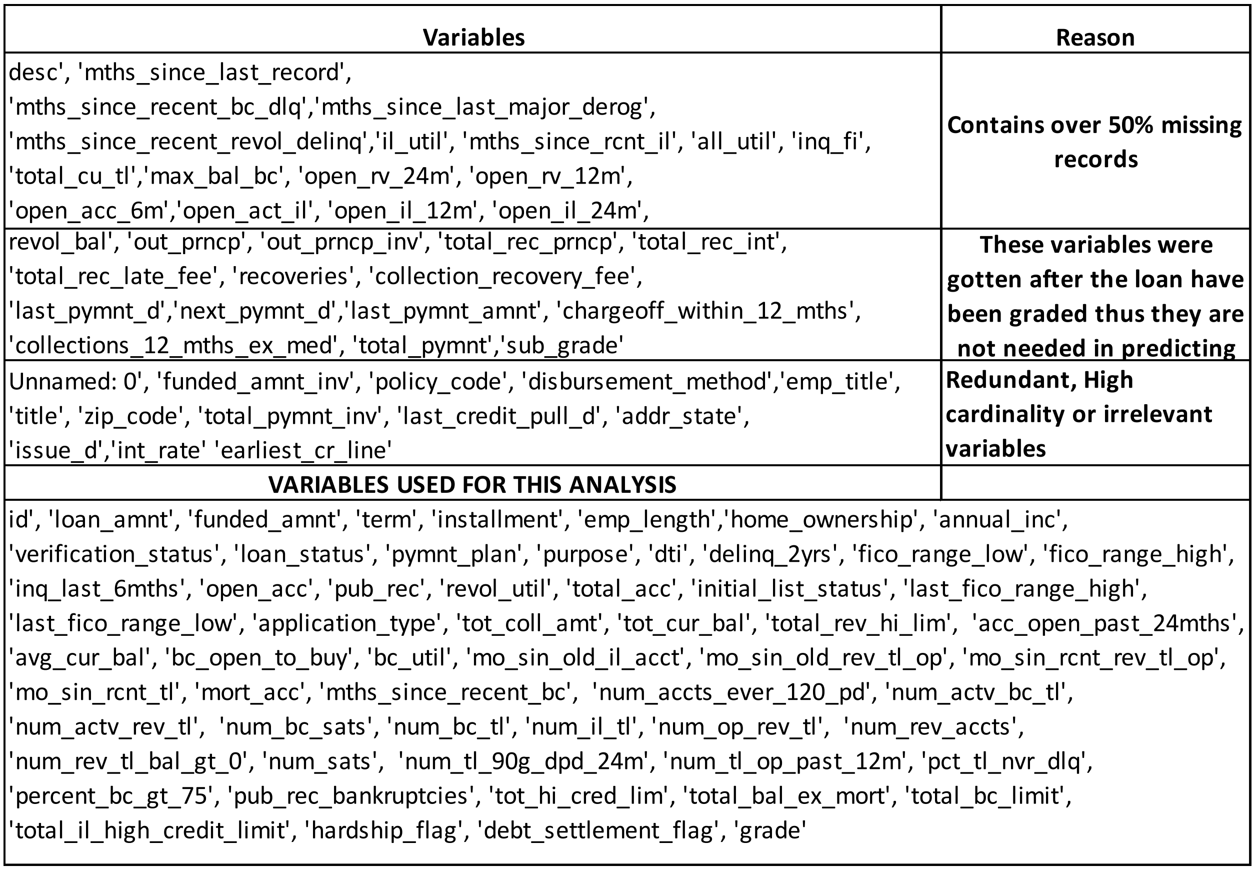
BUSINESS UNDERSTANDING

A few points deduced from the data and research about the business model are as follow:

- Borrowers applies for loans on the platform. During the application stage certain key information about the borrower are provided and company runs a credit check to assess the borrower using previous loan history, outstanding debts, records from credit bureaus, etc. before a loan is approved. Based on this a grade is assigned to every loan.
- Approved loans are part or fully funded by the company and investors. These are listed on the platform for investors to assess and invest.
- Borrowers are expected to repay the loans in either 36 or 60 months in installments. A loan is classified as being defaulted or charged-off if the borrower fails to repay or miss installments.

DATA UNDERSTANDING

The data provided consists of 75000 rows and 114 columns. Every variable in the data was carefully studied using the data dictionary and research to ascertain the features important for loan grade prediction and based on findings the variables were classed and all variables not needed for this analysis were dropped to reduce the complexity of the data. The table below gives the variables used for this analysis, the dropped variables and reason why they were dropped.



Variables	Reason
desc', 'mths_since_last_record', 'mths_since_recent_bc_dlq','mths_since_last_major_derog', 'mths_since_recent_revol_delinq','il_util', 'mths_since_rcnt_il', 'all_util', 'inq_fi', 'total_cu_tl','max_bal_bc', 'open_rv_24m', 'open_rv_12m', 'open_acc_6m','open_act_il', 'open_il_12m', 'open_il_24m',	**Contains over 50% missing records**
revol_bal', 'out_prncp', 'out_prncp_inv', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_d','next_pymnt_d','last_pymnt_amnt', 'chargeoff_within_12_mths', 'collections_12_mths_ex_med', 'total_pymnt','sub_grade'	**These variables were gotten after the loan have been graded thus they are not needed in predicting**
Unnamed: 0', 'funded_amnt_inv', 'policy_code', 'disbursement_method','emp_title', 'title', 'zip_code', 'total_pymnt_inv', 'last_credit_pull_d', 'addr_state', 'issue_d','int_rate' 'earliest_cr_line'	**Redundant, High cardinality or irrelevant variables**
VARIABLES USED FOR THIS ANALYSIS	
id', 'loan_amnt', 'funded_amnt', 'term', 'installment', 'emp_length','home_ownership', 'annual_inc', 'verification_status', 'loan_status', 'pymnt_plan', 'purpose', 'dti', 'delinq_2yrs', 'fico_range_low', 'fico_range_high', 'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_util', 'total_acc', 'initial_list_status', 'last_fico_range_high', 'last_fico_range_low', 'application_type', 'tot_coll_amt', 'tot_cur_bal', 'total_rev_hi_lim', 'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy', 'bc_util', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl', 'mort_acc', 'mths_since_recent_bc', 'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq', 'percent_bc_gt_75', 'pub_rec_bankruptcies', 'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit', 'hardship_flag', 'debt_settlement_flag', 'grade'	

The variables used for this analysis were further grouped into a few classes to help explore the data in chunks which will aid better data understanding. The groups are as follow:

VARIABLES	CATEGORY
open_acc', 'total_acc', 'acc_open_past_24mths', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op', 'mort_acc', 'mo_sin_rcnt_tl', 'mths_since_recent_bc', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_sats', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl','num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_tl_op_past_12m'	**ACCOUNTED RELATED**
'inq_last_6mths', 'mths_since_recent_inq'	**INQUIRY RELATED**
'tot_cur_bal', 'avg_cur_bal', 'total_bal_ex_mort','total_bc_limit', 'revol_util', 'total_rev_hi_lim', 'bc_util', 'bc_open_to_buy'	**BALANCE RELATED**
loan_amnt', 'term', 'grade', 'installment', 'purpose', 'application_type','initial_list_status', 'funded_amnt'	**GRADE RELATED**
fico_range_low', 'fico_range_high', 'tot_hi_cred_lim', 'total_il_high_credit_limit', 'pct_tl_nvr_dlq', 'percent_bc_gt_75', 'pub_rec_bankruptcies', 'pub_rec','last_fico_range_high', 'last_fico_range_low','hardship_flag', 'debt_settlement_flag','tax_liens', 'loan_status'	**SCORING RELATED**
emp_length', 'home_ownership', 'annual_inc', 'verification_status', 'dti', 'tax_liens'	
delinq_2yrs', 'mths_since_last_delinq', 'acc_now_delinq', 'tot_coll_amt', 'delinq_amnt','num_accts_ever_120_pd', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m'	**DELINQUENCY RELATED**

EXPLORATORY DATA ANALYSIS

The data was explored to understand the data better, spot data quality issues and create initial insights from the data. This will help identify the steps needed to make the data suitable for model, identify and select important variables, identify trends in the data and select the appropriate modeling technique. Key aspects explored are:

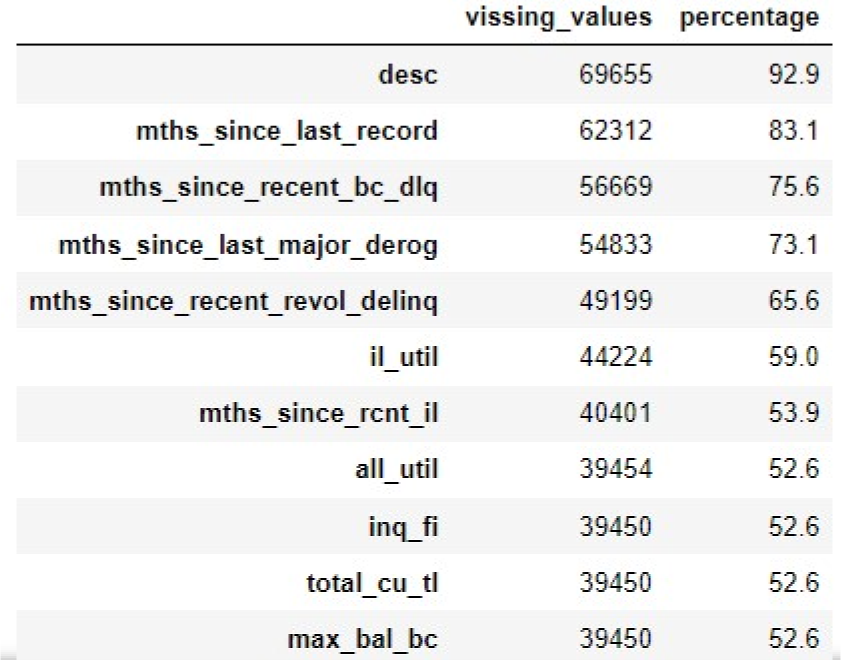
- Descriptive statistics of the data
- Distribution of the data (Univariate Analysis)
- Relationships between variables (multi-variate analysis)
- Data quality concerns like:
    - Identification of missing values
    - Identification of redundant or unwanted variables
    - Identification of outliers
    - Identification of multicollinearity.

Details, figures and thought process for this exploratory data analysis are shown in script “LOAN DATA EXPLORATORY DATA ANALYSIS.ipynb”.

KEY FINDING FROM EDA

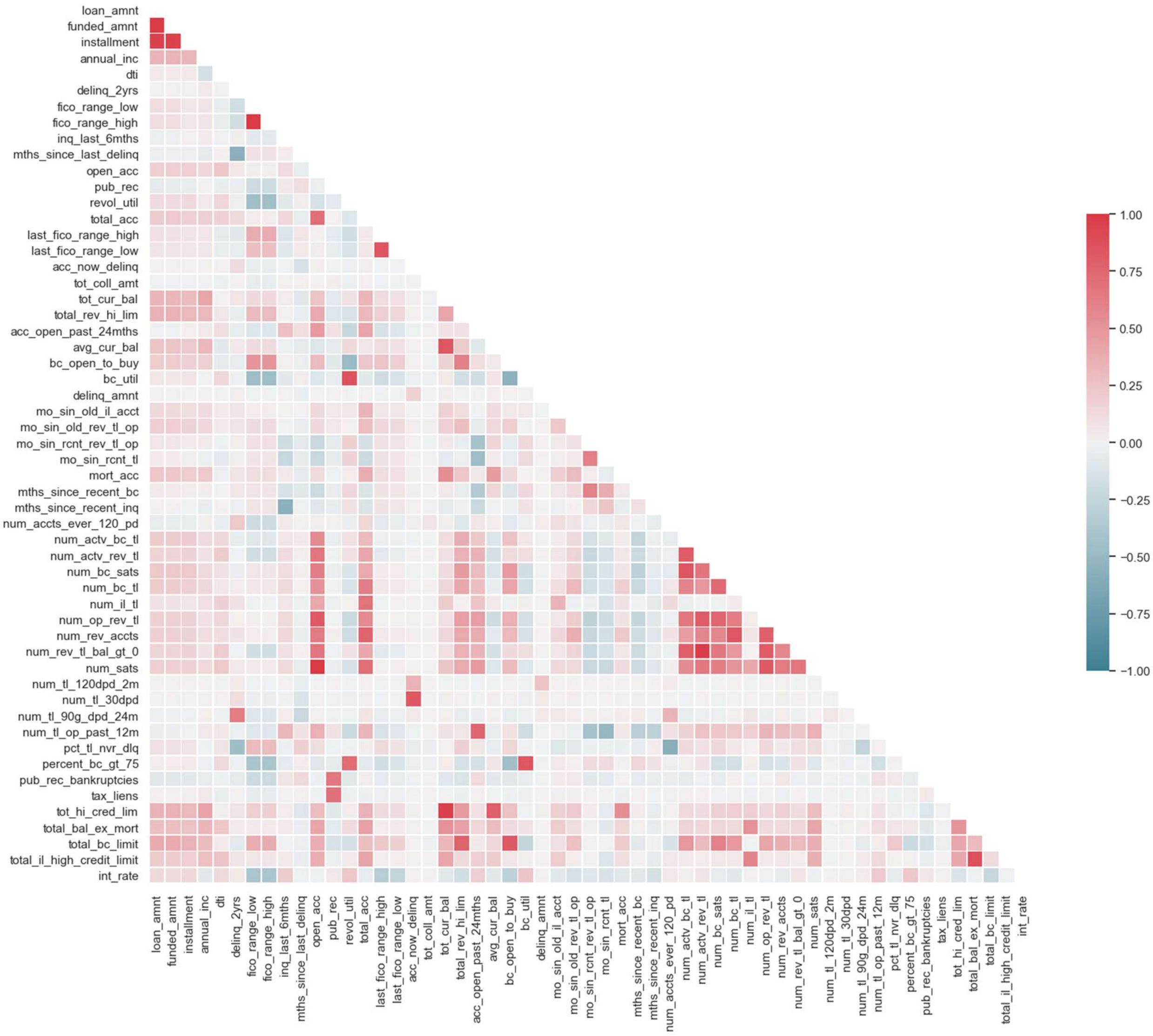
- The dimension of the data is 75000 rows and 114 columns.
- The distribution of the numeric variables shows that some variables do not follow gaussian distribution thus efforts need to be made to correct this using data transformation techniques like boxcox or log transformation before modeling.
- The target variable consists of 7 classes. These classes are not uniformly distributed thus this is a multi-class, imbalance classification problem. This greatly influence the modeling technique to be employed. Although Linear models can be adapted to deal with multi-class problems, tree-based models and neural network models will be more suitable for this problem.
- The distribution of the target variable is: Grade C, B, A, D, E, F, G with a count of 22465, 22465, 12264, 10844, 5136, 1701, 476 records respectively.
- The predominant class for each categorical feature is given below:
  - Term: 36 months
  - Employment length: 10 years plus
  - Home ownership: Mortgage
  - Income source verification status: Source verified
  - Loan Status: Current loan
  - Repayment plan already setup: No
  - Purpose: Debt consolidation
  - Application type: Individual
  - Debt settlement flag: No
  - Hardship flag: No
  - Grade: C
- The cardinality of the following variables is high thus needs to be reduced for better analysis:
  - Home ownership
  - Purpose
  - Loan Status
- A total 66 variables contain missing values. It was also observed that some variables contain 4% missing values across board and further investigation showed that these missing values were present across the rows of the data.

Such rows need to be dropped. Few variables with missing values and the percentage of missing values are summarized below:

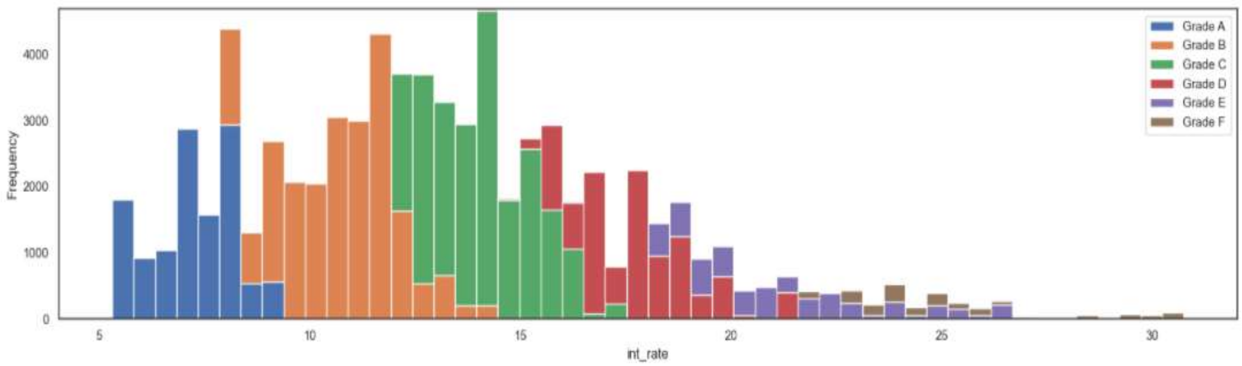


	vissing_values	percentage
desc	69655	92.9
mths_since_last_record	62312	83.1
mths_since_recent_bc_dlq	56669	75.6
mths_since_last_major_derog	54833	73.1
mths_since_recent_revol_delinq	49199	65.6
il_util	44224	59.0
mths_since_rcnt_il	40401	53.9
all_util	39454	52.6
inq_fi	39450	52.6
total_cu_tl	39450	52.6
max_bal_bc	39450	52.6

- Outliers were observed in the data. This will be corrected before modeling.
- Some variables were observed to have a strong positive or strong negative correlation. The correlation between variables is summarized below:



- The distribution of the loan grades with respect to interest rate shows a very clear relationship. The interest rate increases with respect to loan grade such that grade A loans have the lowest interest rate and grade F loans have the highest interest rate.



- Multicollinearity check using variable inflation factor shows very high collinearity in some variables. Although the predictive power or reliability of machine learning algorithms is generally not affected by the multicollinearity of variables, the importance of variables with high collinearity will be offset by each other, thereby affecting the overall interpretability of the predictor variables. This greatly affects linear models thus its more suitable to use tree-based models for this use case or perform some sort of dimensionality reduction to cancel the effect of multicollinearity or total removal of collinear variables.

DATA PRE-PROCESSING

The following steps were carried out to preprocess the data:

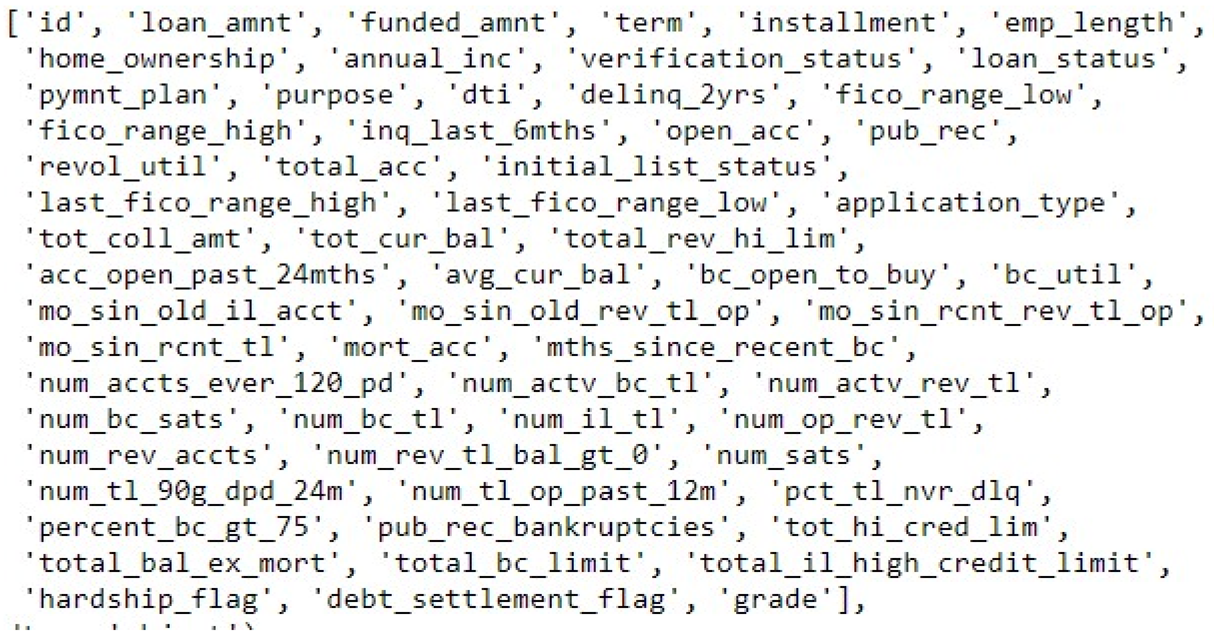
- Delete columns containing either 50% or more than 50% missing Values.
- Drop columns that are not needed, redundant or have high cardinality.
- Delete columns that will only be available after a loan have been graded.
- Drop all rows with 4% missing values across all columns.
- Drop columns in which missing values cannot be imputed. The two columns dropped are: 'mths_since_last_delinq','mths_since_recent_inq'
- Drop Duplicated and Redundant records.
- Reduce cardinality of categorical variables with high cardinality and format the classes appropriately.
- Fill missing values of numerical variables with the mean values of each variable.
- Fill missing values of categorical variable with the mode.
- Handle outliers by setting the upper and lower boundary of values in each variable to the 95th and 5th percentile respectively. This will automatically set

values above or below the upper and lower boundary for each variable to the value of the upper and lower boundary respectively.

- Convert the data type of all categorical, continuous and ordinal variables to object, float and integer data type respectively.
- Encode the labels of the target variables using ordinal encoding
- Split data into training and validation set in a ratio of 80:20
- Scale numeric variables
- Encode categorical variables using label encoder.

Modeling using XGBoost needs an extra step which involves the convertion of the training dataframe to a DMatrix.

After data preprocessing the dimension of the data was reduced to 72003 rows and 60 columns as oppose 75000 rows and 114 columns in the original data. The variables selected for modeling task are as follow:



```
['id', 'loan_amnt', 'funded_amnt', 'term', 'installment', 'emp_length',
 'home_ownership', 'annual_inc', 'verification_status', 'loan_status',
 'pymnt_plan', 'purpose', 'dti', 'delinq_2yrs', 'fico_range_low',
 'fico_range_high', 'inq_last_6mths', 'open_acc', 'pub_rec',
 'revol_util', 'total_acc', 'initial_list_status',
 'last_fico_range_high', 'last_fico_range_low', 'application_type',
 'tot_coll_amt', 'tot_cur_bal', 'total_rev_hi_lim',
 'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy', 'bc_util',
 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op',
 'mo_sin_rcnt_tl', 'mort_acc', 'mths_since_recent_bc',
 'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl',
 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl',
 'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats',
 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq',
 'percent_bc_gt_75', 'pub_rec_bankruptcies', 'tot_hi_cred_lim',
 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit',
 'hardship_flag', 'debt_settlement_flag', 'grade'],
```

MODELING

Tree based models will be experimented in this project due to their suitability in handling multi-class prediction problems and the multicollinearity observed in the data. The following models were experimented:

- Decision Tree Classifier

- Random Forest Classifier
- XGBoost Classifier

The performance of the models was evaluated using the following metrics

- Accuracy score
- Precision: How many percent of times our model can identify a positive class correctly.
- Recall: What proportion of the actual positive class were identified correctly
- F1-Score
- Misclassification Rate (Confusion Matrix)

The models will be evaluated by on the accuracy on the train and test set to assess if the model is overfitting and how it will perform on unseen data.

The precision and recall score will also be used to evaluate the models. The precision and recall scores for a good model should be high.

These models were experimented using all features in the data and a subset of the data obtain using recursive feature elimination technique. The models were fit using cross validation techniques to avoid overfitting.

PRESENTATION OF RESULTS

The figure below shows the result and confusion matrix obtained from each model using the entire features and a subset obtained using recursive feature elimination.

DECISION TREE CLASSIFIER

A. ALL FEATURES

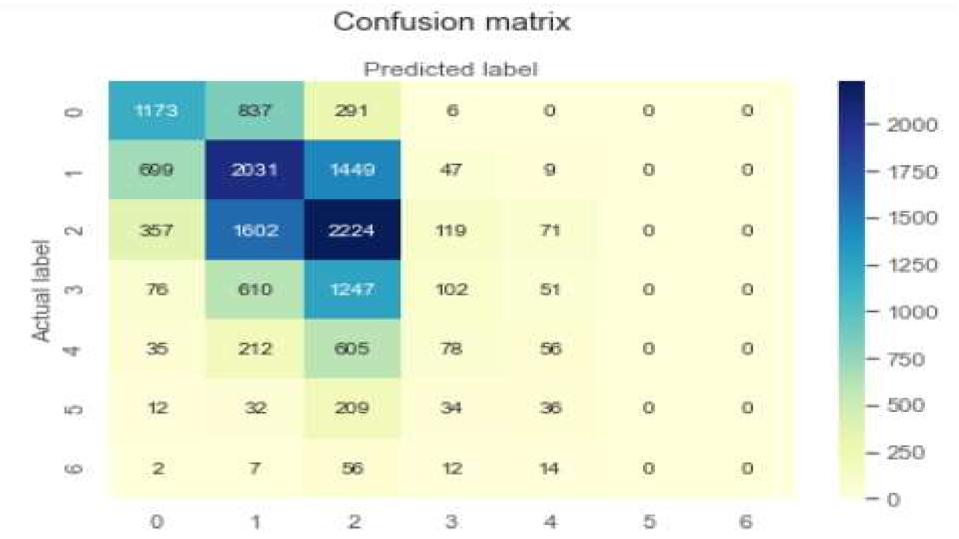
Train data accuracy score: 0.395 (39.5%)

Test data accuracy score: 0.388 (38.8%)

Precision: 0.23

Recall: 0.25

F1-Score: 0.22



	0	1	2	3	4	5	6
0	1173	837	291	6	0	0	0
1	699	2031	1449	47	9	0	0
2	357	1602	2224	119	71	0	0
3	76	610	1247	102	51	0	0
4	35	212	605	78	56	0	0
5	12	32	209	34	36	0	0
6	2	7	56	12	14	0	0

B. Recursive Feature Elimination

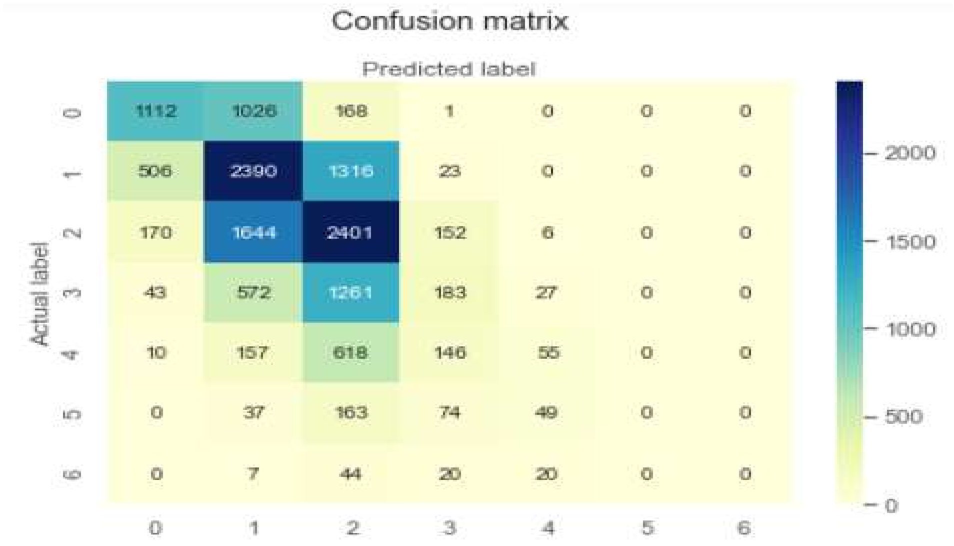
Train data accuracy score: 0.438 (43.8%)

Test data accuracy score: 0.426 (42.6%)

Precision: 0.25

Recall: 0.3

F1-Score: 0.24



	0	1	2	3	4	5	6
0	1112	1026	168	1	0	0	0
1	506	2390	1316	23	0	0	0
2	170	1644	2401	152	6	0	0
3	43	572	1261	183	27	0	0
4	10	157	618	146	55	0	0
5	0	37	163	74	49	0	0
6	0	7	44	20	20	0	0

RANDOM FOREST CLASSIFIER

Random Forest Classifier (All Features)

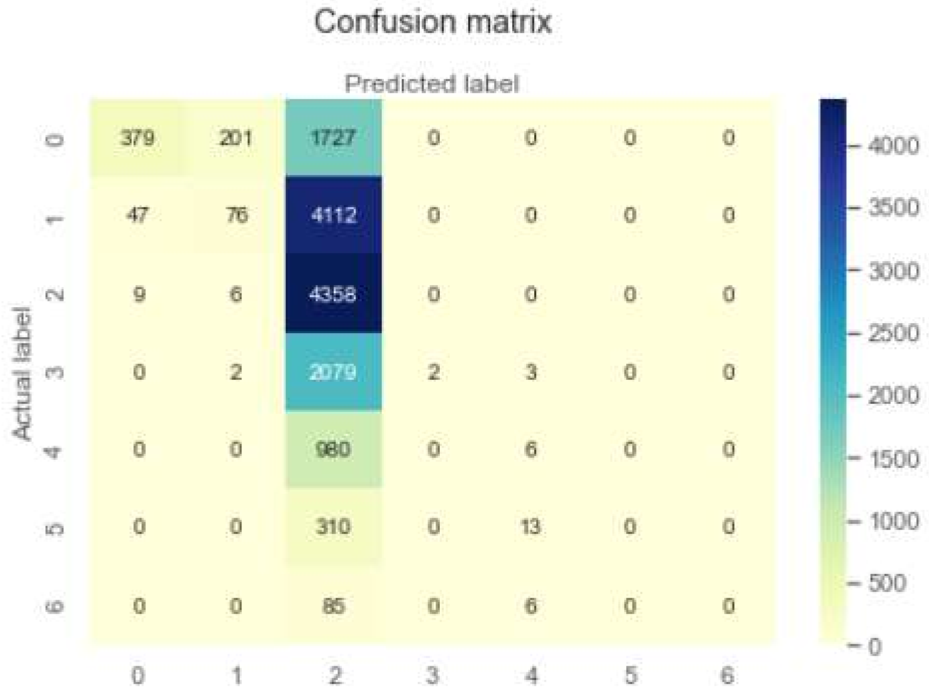
Train data accuracy score: 0.336 (33.6%)

Test data accuracy score: 0.335 (33.5%)

Precision: 0.17

Recall: 0.38

F1-Score: 0.12



Random Forest Classifier (Recursive feature Elimination)

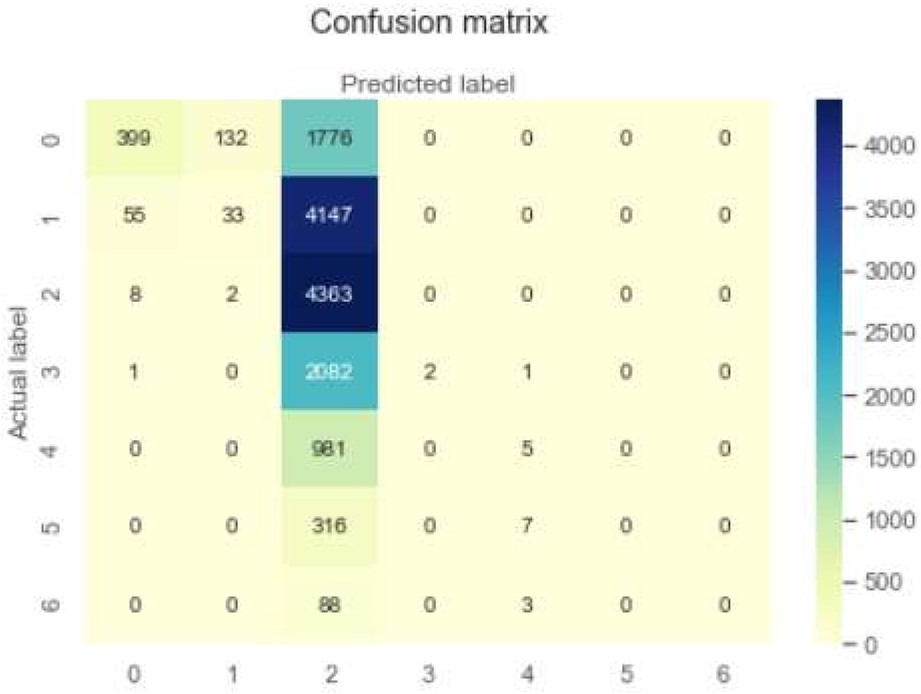
Train data accuracy score: 0.334 (33.4%)

Test data accuracy score: 0.333 (33.3%)

Precision: 0.17

Recall: 0.38

F1-Score: 0.11



XGBOOST CLASSIFIER

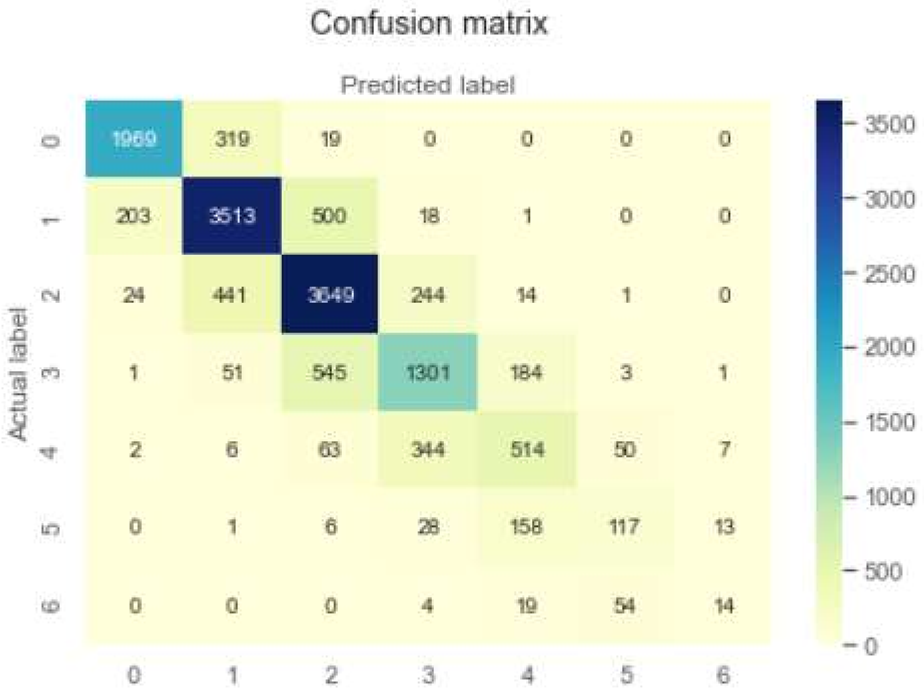
Train data accuracy score: 0.868 (86.8%)

Test data accuracy score: 0.769 (76.9%)

Precision: 0.6

Recall: 0.66

F1-Score: 0.62



DISCUSION

- Decision tree and random forest model failed to predict classes with little number of record and the accuracy score, precision score, recall score, and misclassification rate of these models is very poor.
- The results clearly shows that XGBoost model outperforms all other model both in terms of all metrics and it predicted all 7 classes with a great deal of accuracy thus XGBoost was adopted as the final model.
- A precision value of 0.6 obtained using XGBoost model indicates that the model can make correct predictions 60% of times.
- A recall value of 0.66 obtained using XGBoost indicates that 66% of the actual positive classes were identified correctly by the model.
- The recall and precision values for the XGBoost model is above the 0.5 threshold and they are both close. This indicates a good model. However, there is room for further improvement of the model.

CONCLUSIONS

This project provides a present a few machine learning models that can be used to predict loan grade and concludes that XGBoost model is the most suitable model for this use case and dataset. However, the model can be further improved by exploring further preprocessing steps like transformation of skewed variables using boxcox transformation or log transformation, experimenting on different encoding techniques, dimensionality reduction using PCA etc. Also, its will be a good idea to explore deep learning techniques.