Grade Prediction and Default Loan Detection Project

By: Ensieh Bahrami

Supervised by: Alison Cossette

Contents

Problem Identification

Lending Club is an American peer-to-peer lending company, one of its services is offering loans to the individuals. They have collected data about 55 features for 10 thousand borrowers. In fact, data set only represents the loans made.

Based on dataset, each loan has been classified in one of the 7 grades. 'A' grade loans represent the lowest risk while 'G' grade loans are the riskiest. Loan grading is a classification system that involves assigning a quality score to a loan based on a borrower's financial history and present the likelihood of repayment of the principal and interest.
While majority of the loans were in the process of being paid back, there were 7 loans out of 10 thousand which were in default.

There is an opportunity to recognize the most important features in grade classification and also defaulting loans to greatly hedge the risk of the lender. Building a predictive model to simply interpret the results is the main goal of this project.

Data Cleaning & Understanding and EDA

Dataset namded 'loans_full_schema' had 10k rows and 55 features. It included 36 numerical, 16 categorical and 3 Boolean variables. One of the categorical variables was employment title which was string and separately cleaned, analyzed, and used for prediction.
I performed data wrangling and exploration with two approaches, once considering loan grade as target variable and another time being target or not as independent variable.
Predicting the loan grade was a multiclassification problem, while detecting the default loans was a binary classification problem.
Dataset source: https://www.openintro.org/data/index.php?data=loans_full_schema
Dataset was about 10k unique borrowers whose application for a loan had been accepted. Based on borrowers' characteristics, each borrower received a loan with different grade from A to G. Now few of the loans were default, majority of them are in the process of reimbursement and some have delay in their reimbursement.

My findings during data wrangling and exploration are as follows:

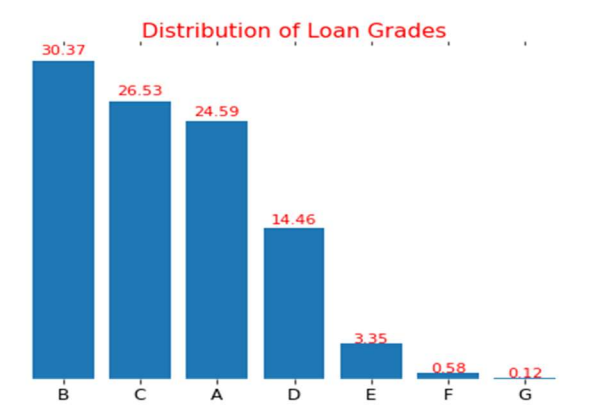
1. Grade and Interest Rate

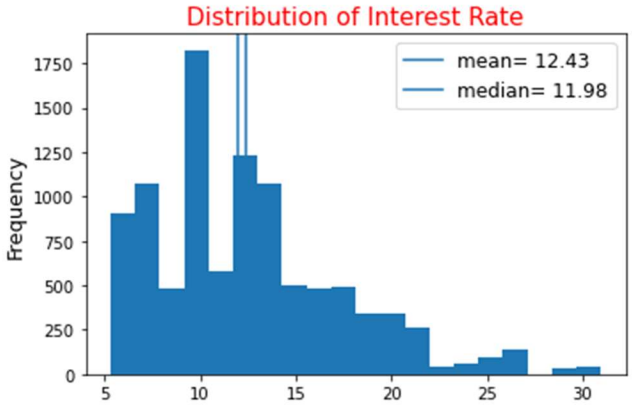
More than 80% of borrowers had grade A, B, or C and about 4.5% had grade E, F, or G.

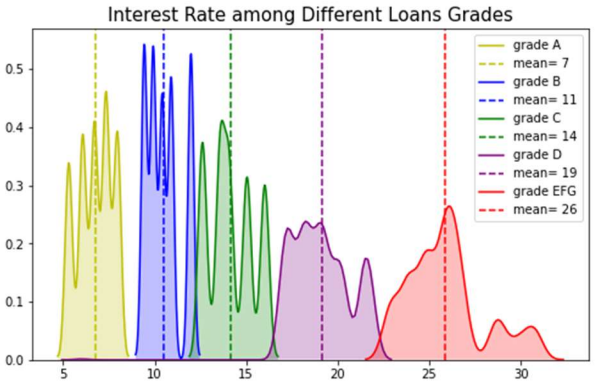
Distribution of interest rate is skewed to the left. There are loans with interest rate between 20% and 31%, while 50% of loans have less than 12% interest rate.

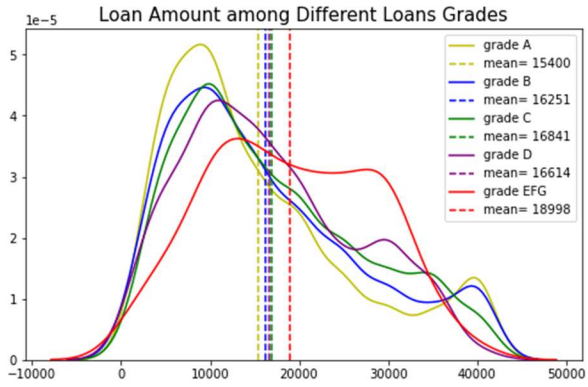
Interest rate is totally dependent to the grade of loan. Mean of interest rate for loans with low risk (grade A) is 7%, whereas loans with high risk (EFG) had higher mean of interest rate of 26%.

Average amount of loan with grade A is 15.4k $, which is less than average amount of high-risk loans (grades E, F or G) 18998$.







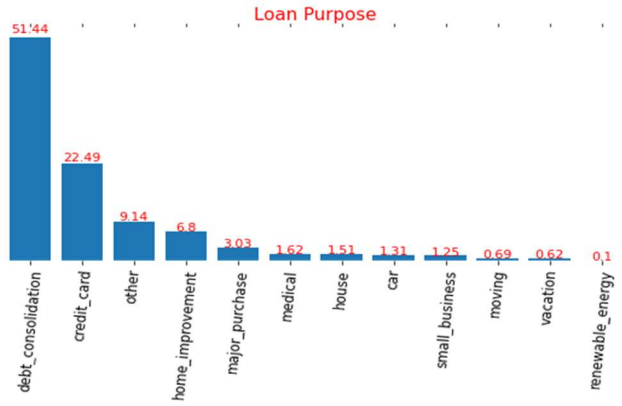


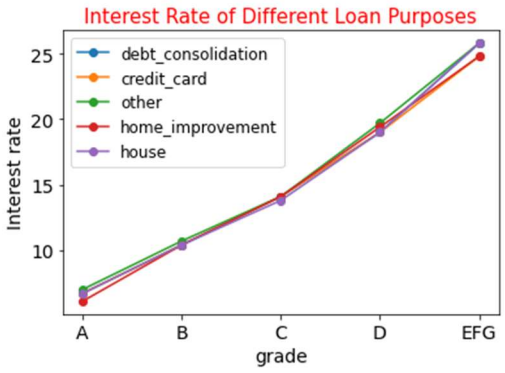
2. Grade, Loan Purpose, and Loan Amount

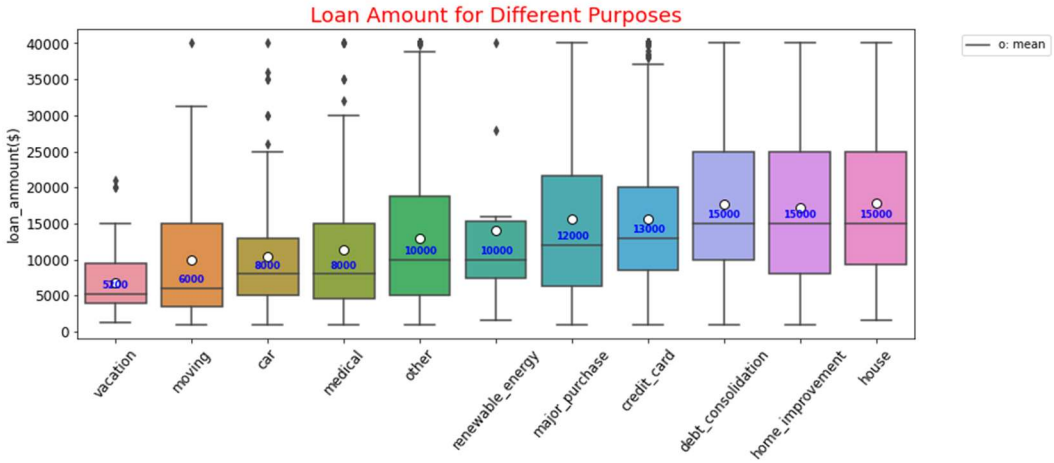
About 60% of loans used for debt consolidation, home improvement and buying house when 50% of them were 15k $ or less.

More than 22% of loans used to pay off the credit card with the median of 13k $ or less.

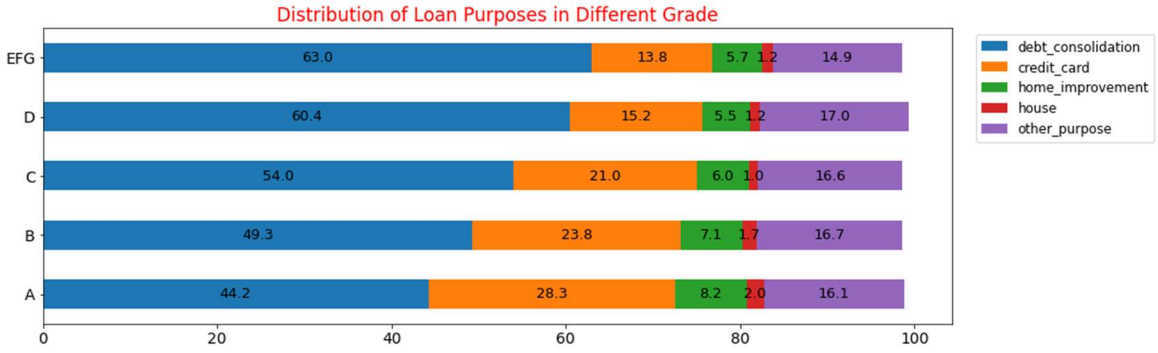
Interest rate was highly dependent to loan grade rather than its purpose.







Higher risk loans used more often for debt consolidation, and lower risk loans commonly used for credit card.

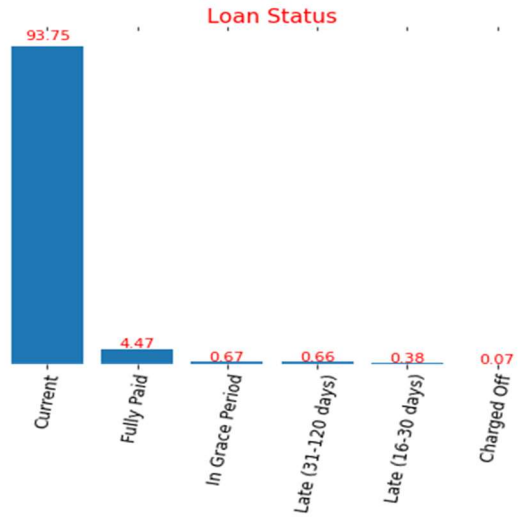


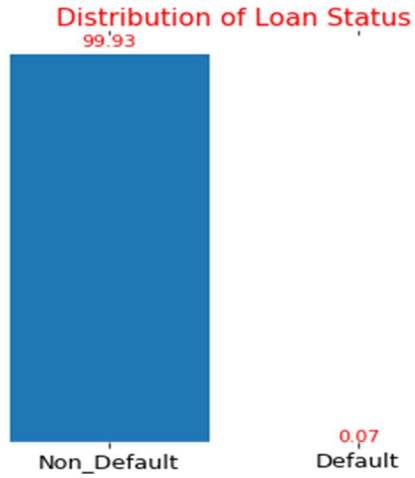
3. Default, Loan Status, Grade, and Interest Rate

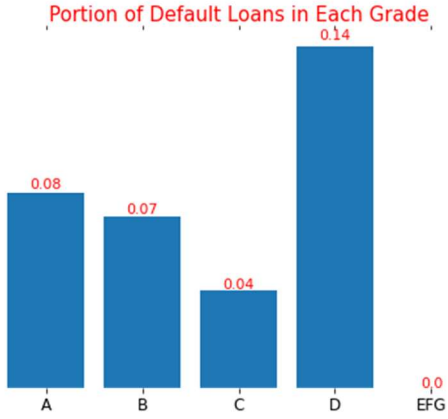
About 1.11% of loans were late from 16 to 120 days, or were charged off. And, only 0.07% of loans were default (7 default loans from 10k loans).

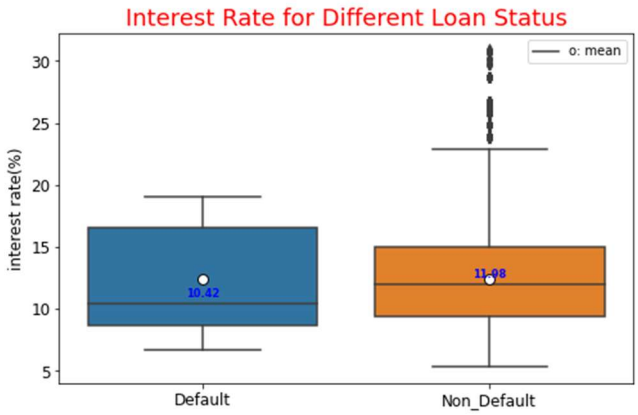
Highest portion of default loans were in grade D loans.

Median of interest rate of default loans were 0.66% lower than non-default loans.





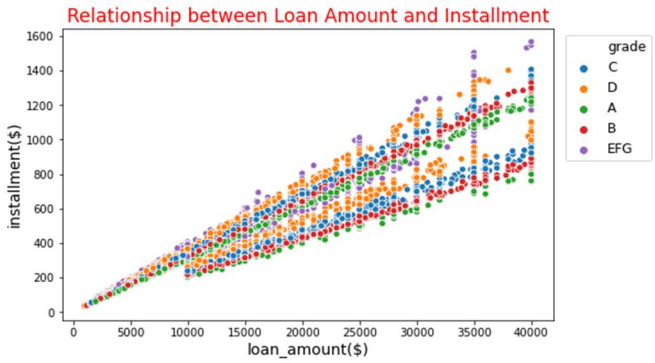


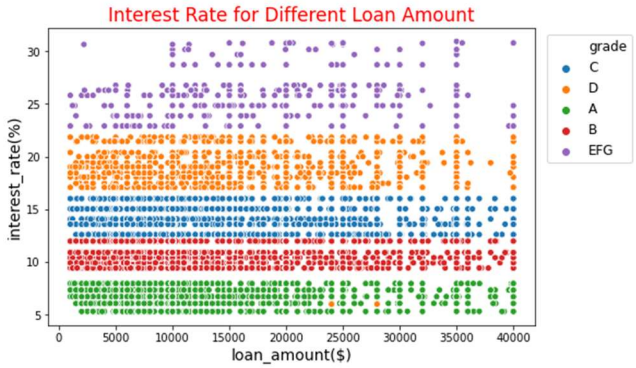


4. Default, Loan Amount, and Loan Purposes

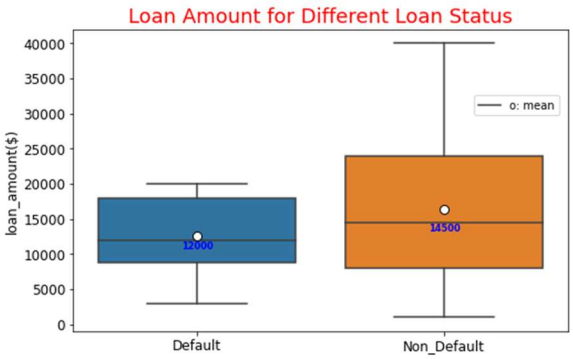
Installment is positively related to the loan amount, and it seems to be independent from grade.

Interest rate was highly dependent to loan grade rather than loan amount.

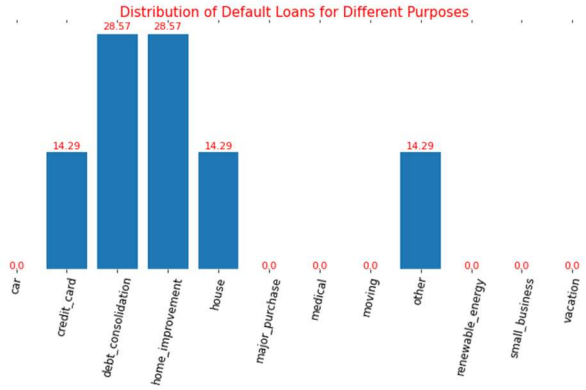


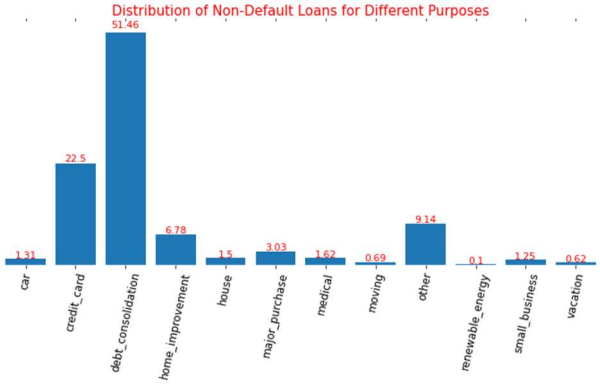


Median of loan amount of defaults was 2.5k dollars lower than non-default ones.



Default loans got more for home improvement and house than non-default loans. And non-default loans received more for credit card pay-off and debt consolidation.



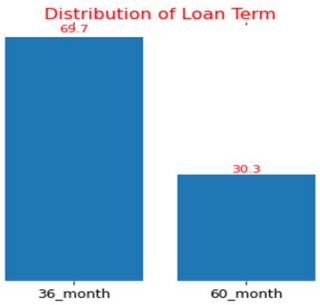


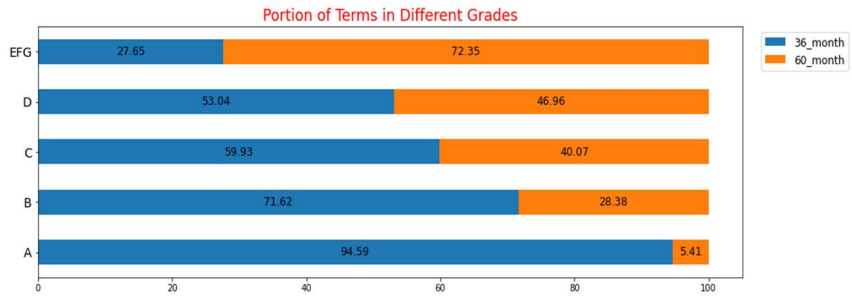
5. Grade and Term

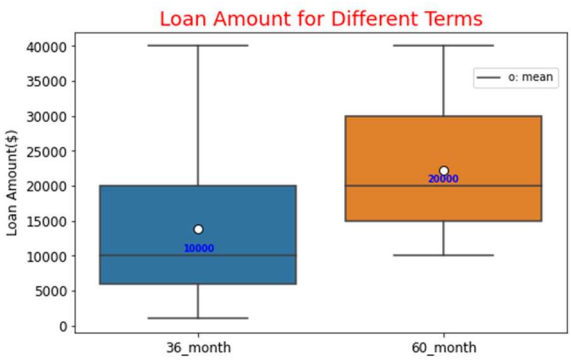
About 70% of loans had term of 36 month and remaining 30% had 60 month.

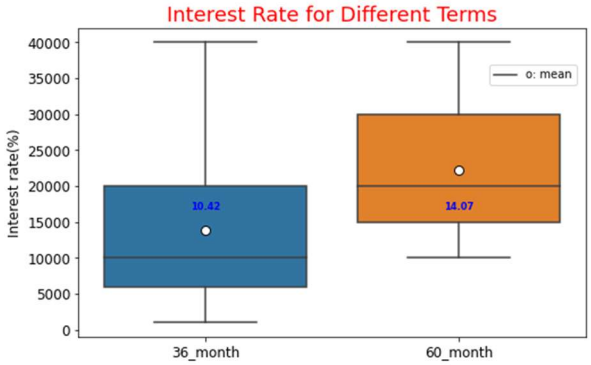
Majority of low risk loans had term of 36 month; while majority of high risk loans had term of 60 month.

Loan amount and interest rate for loans with term of 60 month were higher than loans with term of 36 month. Median of loan amout for longer term had twice of 36 month loans; and median of interest rate for long term loans was 3.6% higher than shorter term.

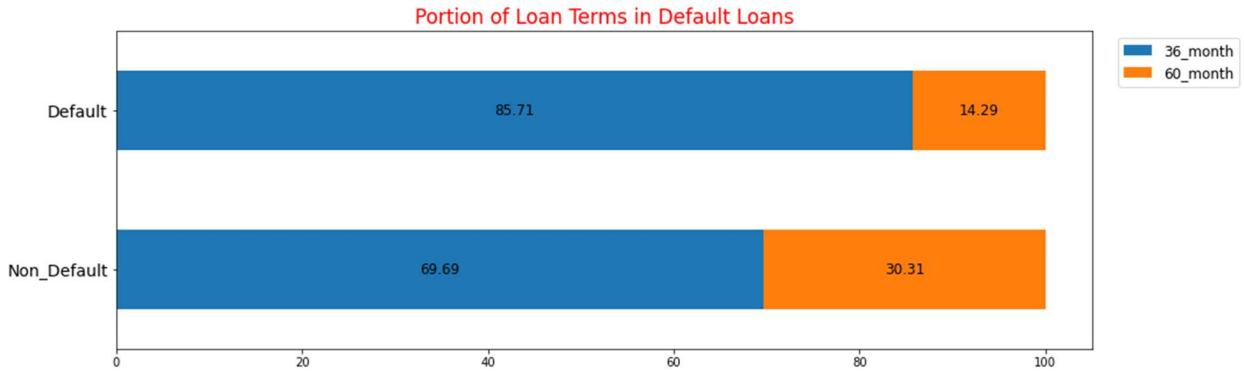








Percentage of 60-month loans among default loans were half of the corresponding percentage among non-default loans.

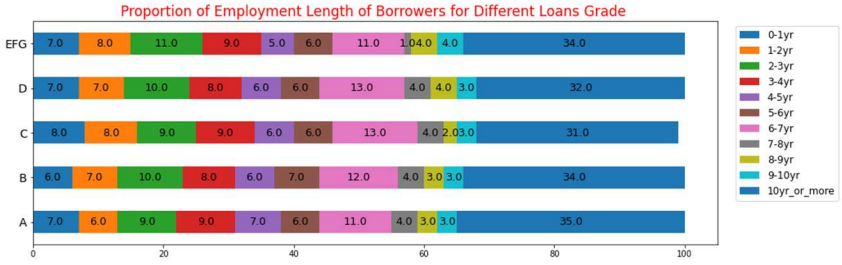


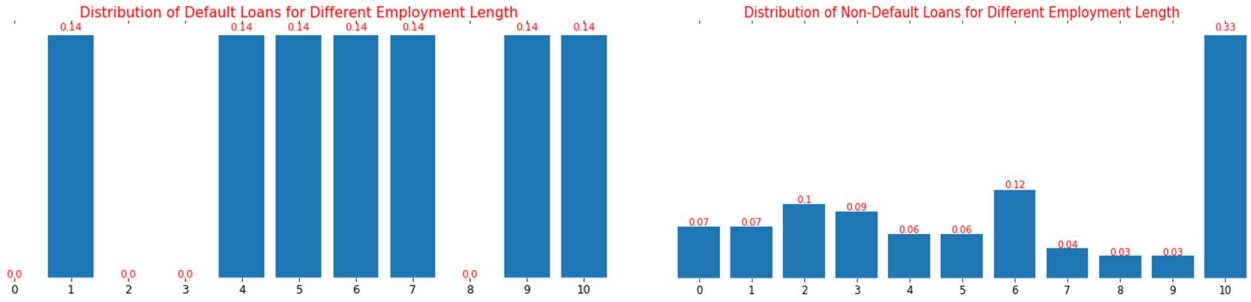
6. Grade and Employment Length

More than 80% of borrowers had grade A, B, C, and 33% of borrowers were with employment length of 10 years or more. Majority of the borrowers with 10 or more years of experience received the loan with grade A, B or C.

Borrowers with 10 or more years of experience have had mor money to spend and used more credit card, which might impact credit card score positively. But Borrowers with 1 to 2 years of experience got fewer low risk loans.

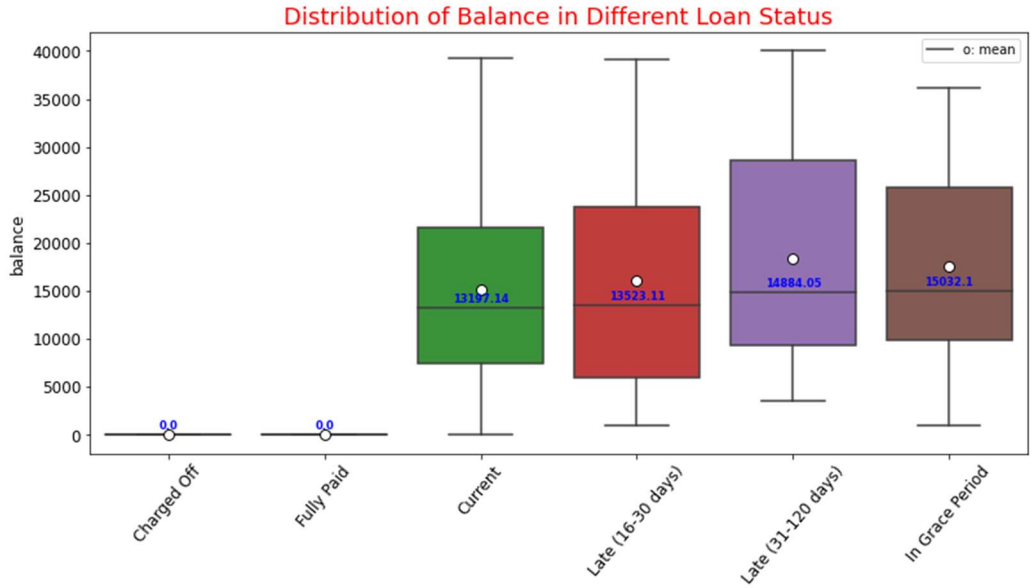
There weren't any default loans received by borrowers with less than 1, 1-2 ,2-3 and 8-9 years of employment.





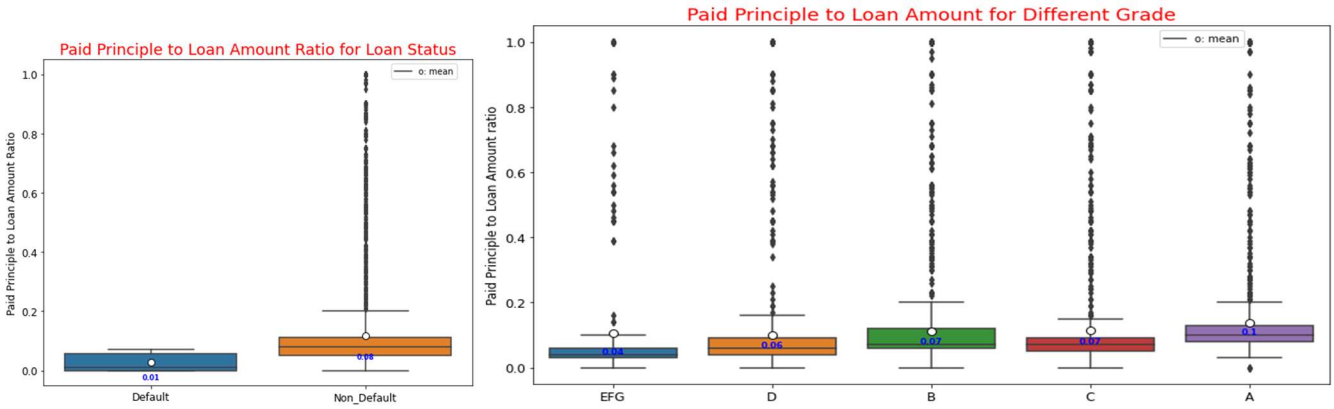
7. Default and Balance

Balance of fully paid and default loans were zero, which makes sense.



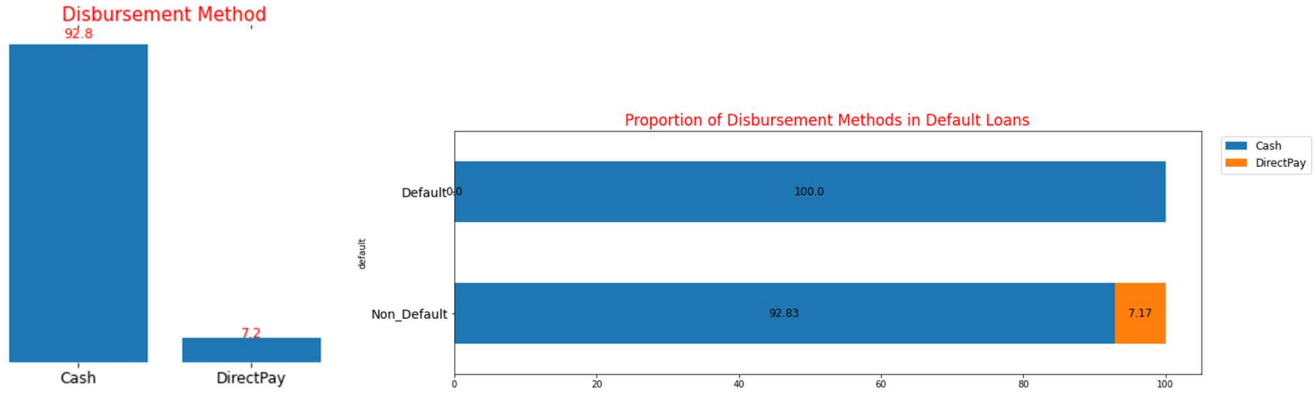
8. Default and Paid Principle

Default loans were not principal paid more than 0.07% of loan amount.



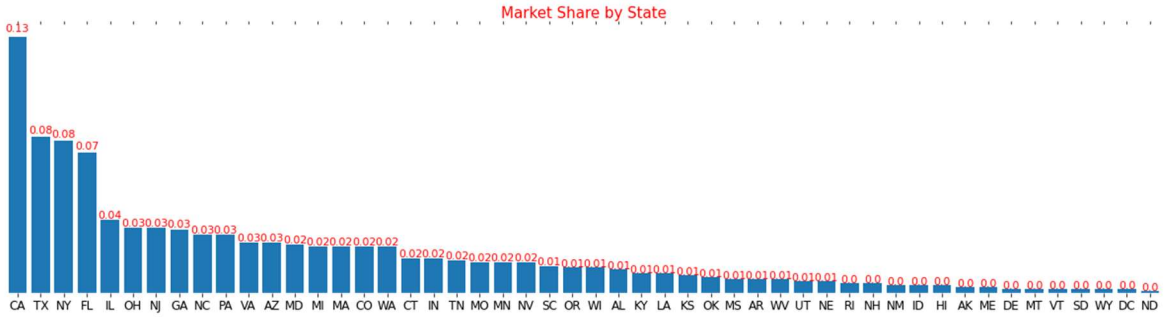
9. Default and Disbursement Method

All default loans were disbursed through cash.



10. Grade, State and Loan Status

I calculated market share based on the number of the loans in each state. Only CA, TX, NY and FL had the market share higher than 5%.

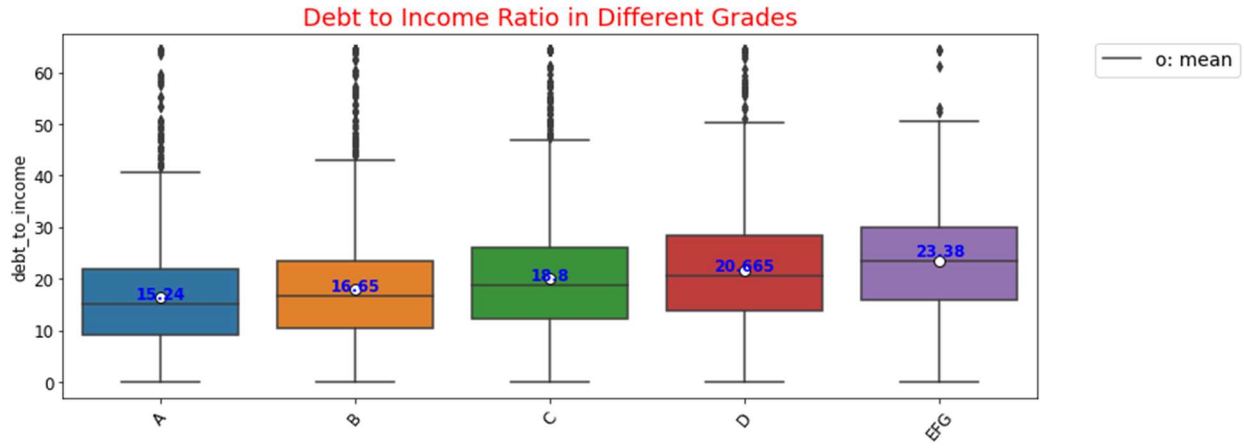


There were 10 states with market share higher than 3%. I considered those states as top markets.

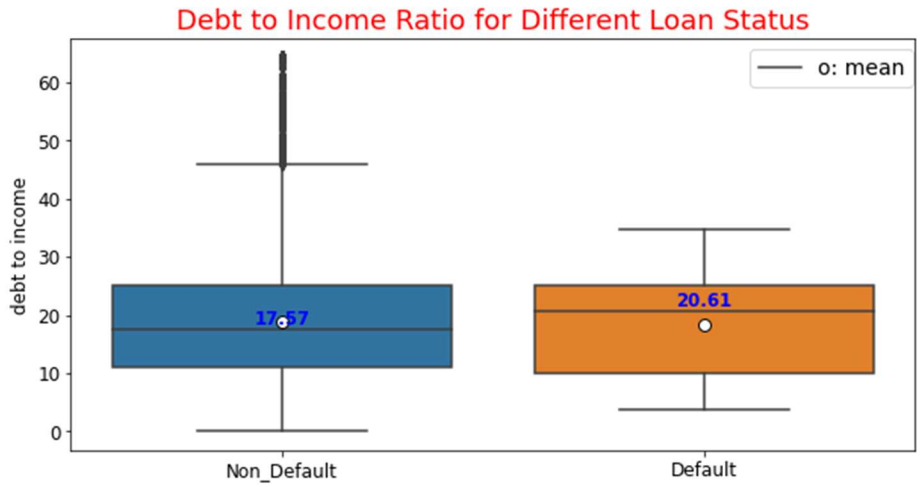
Default loans were only in CA, FL, TX, MD, MI, NV and TN.

11. Default, Grade, and Debt to Income Ratio

As mean and median of debt to income increased, the borrowers received loans with higher risk.



Median of Debt to income for default loans was about about 3% higher that median for non-default.

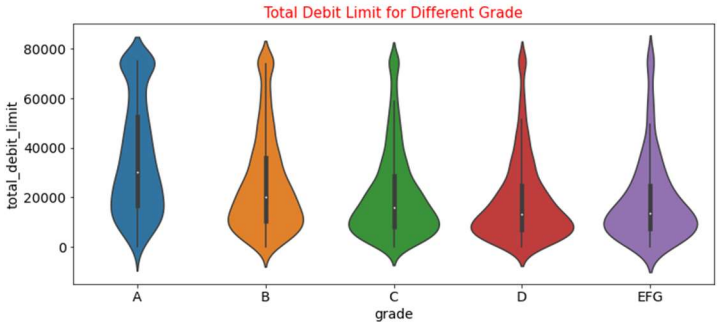


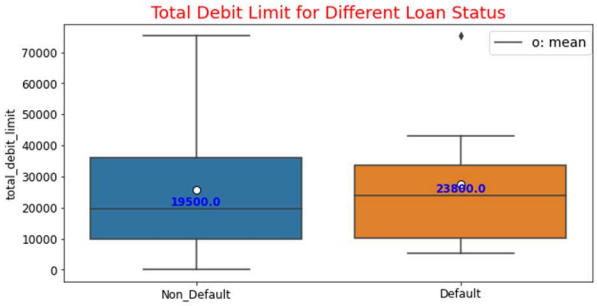
12. Grade and Total Debit Limit

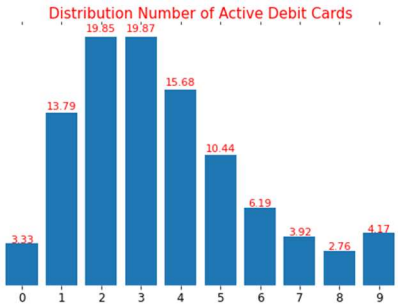
Median of total debit limit for borrowers received lower risk loans were higher than those with higher risk loan, while most of them had fewer number of active debit cards.

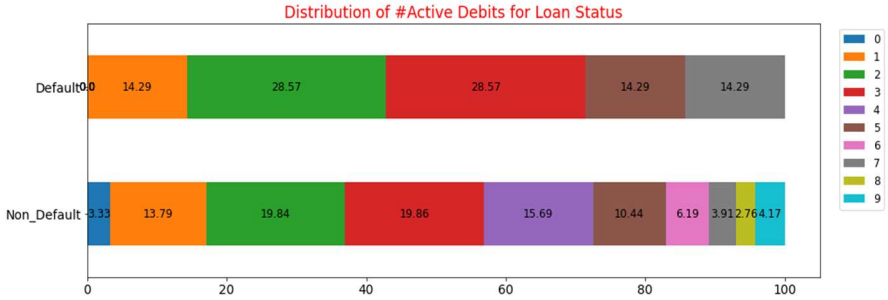
Median of Total Debit limit for borrowers with default loans was 4.3k dollars higher that median for borrowers with non-default.

More than 55% of borrowers with default loans had 2 or 3 active debit cards.



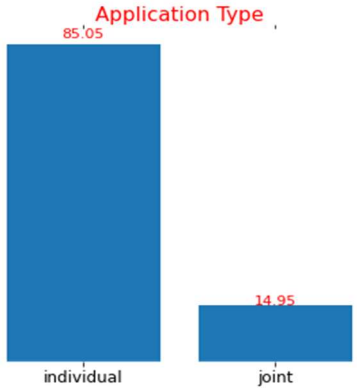


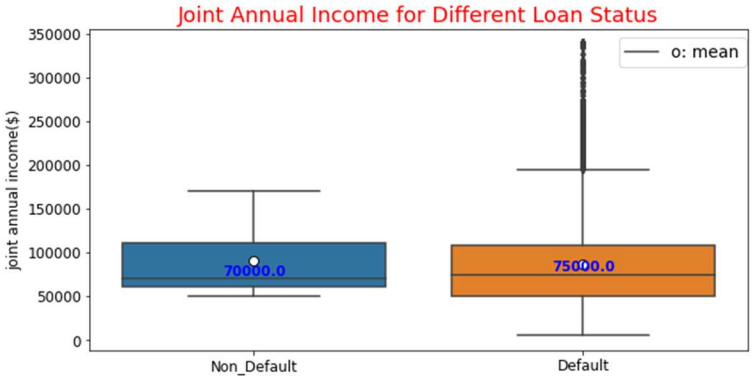




13. Default and Joint

About 15% of borrowers had co-signer. The median of joint annual income for Borrowers with default loans was by 5k dollars higher than borrowers with non-default loans.

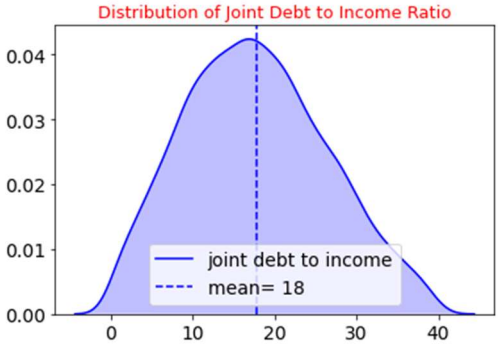


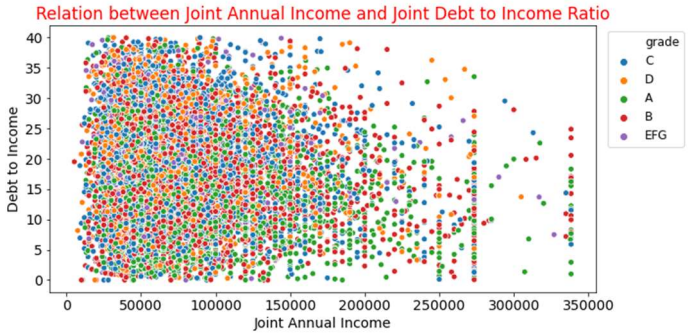


14. Debt to Income Joint

Those 85% of single borrowers had Null value in column of debt_to_income_joint, which makes sense to replace null values with the borrowers' debt_to_income'.

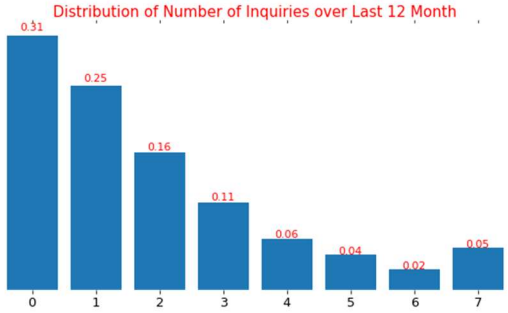
Column name of debt_to_income was replaced with individual_debt_to_income, and column name of debt_to_income_joint was replaced with joint_debt_to_income.

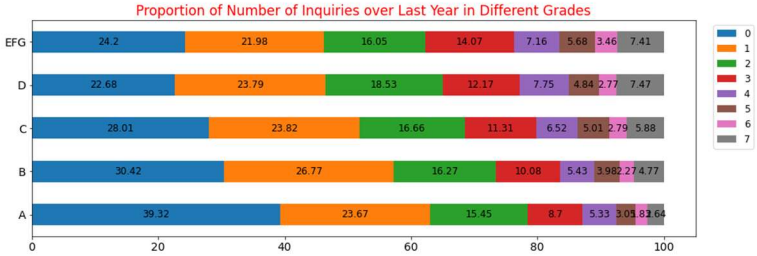




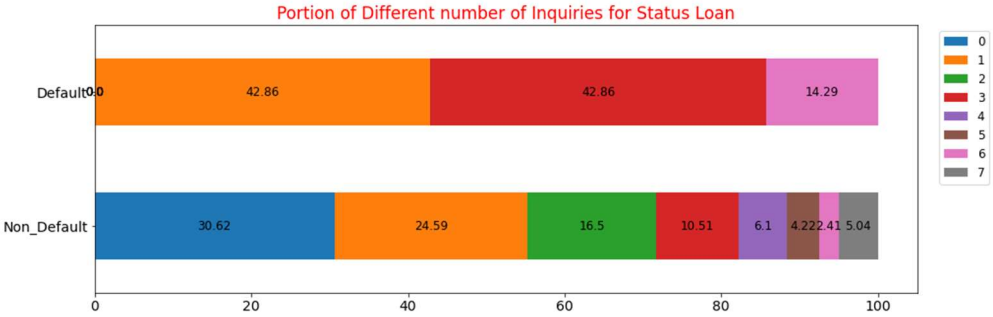
15. Grade and Inquiries over Last 12 Month

Majority of borrowers with zero inquiries over last year recieved low risk loans. In contrast, majority of borrowers with 2 or more inquiries got higher risk loans.



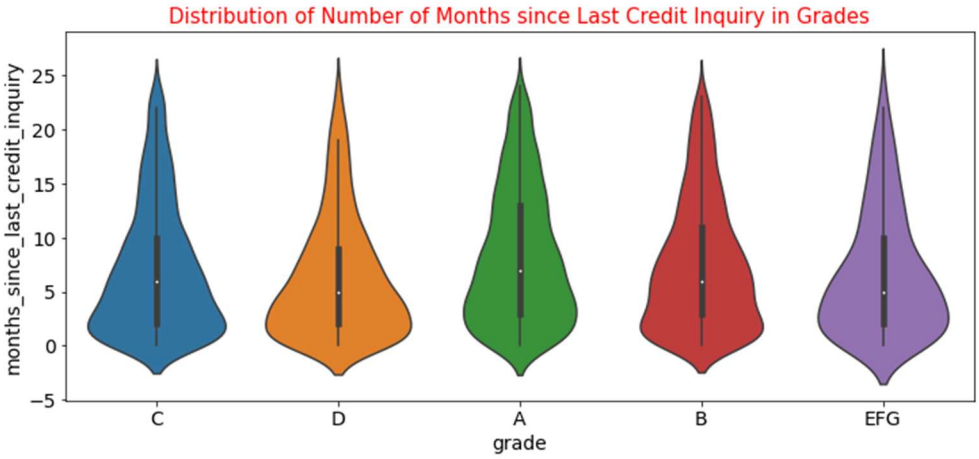


Majority of borrowers with default loans had one or three Inquiries over last 2 years.



16. Grade and Number of Months since Last Credit Inquiry

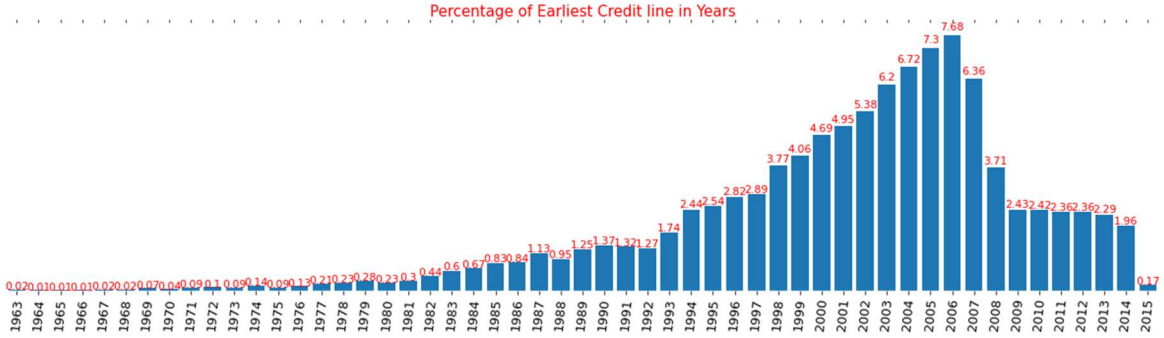
Fewer number of borrowers with positive number of months since last credit inquiry got grade A loans.



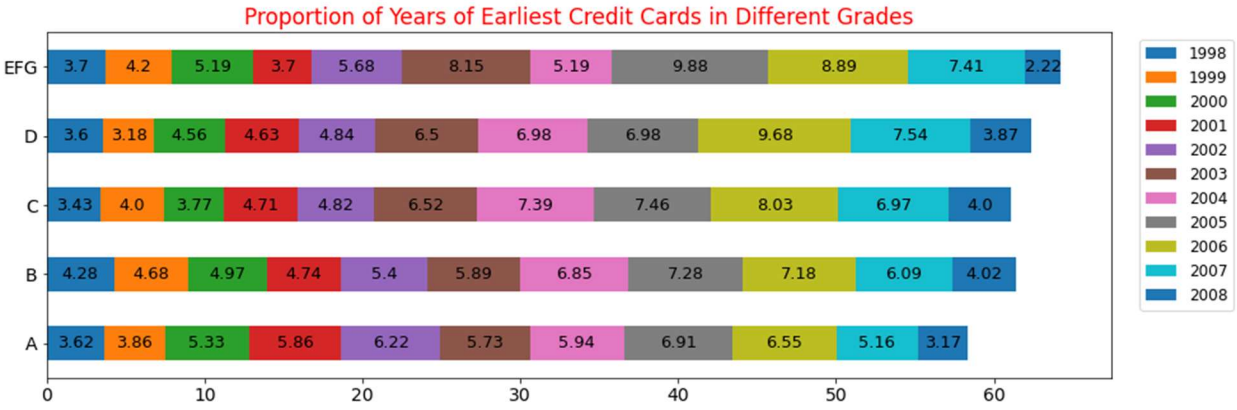
17. Earliest Credit Line

Percentage of earliest credit lines only between 1998 and 2008 were greater than 3%.

Percentage of earliest credit lines from 1992 continuously increased until 2006; then the portion decreased from 2006 to 2014. Recession between the end of Dec-2007 and Jun-2009 might be the reason of drop of portion in those years.



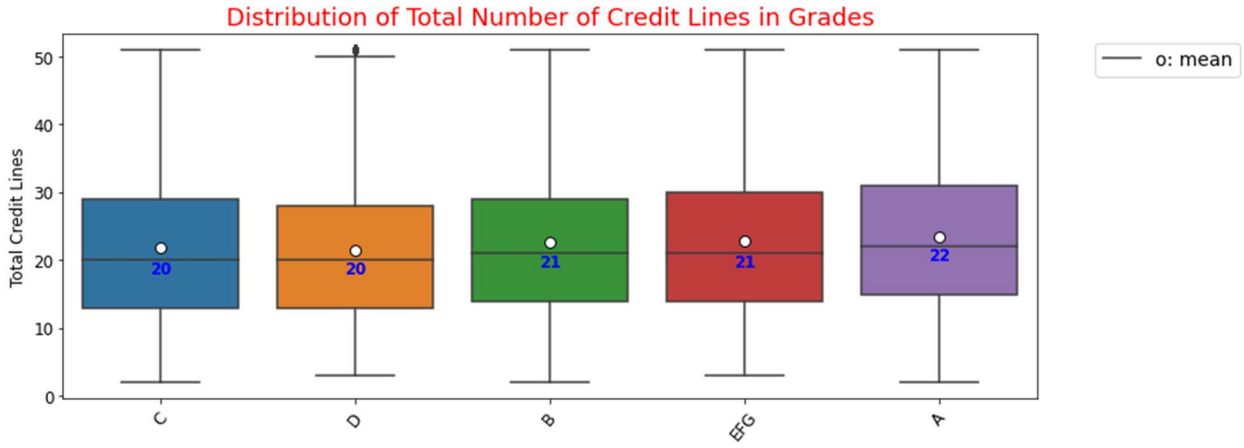
Borrowers who their earliest credit cards issued in 2003, 2006 and 2007 had bigger portion in high-risk loans than low risk ones.

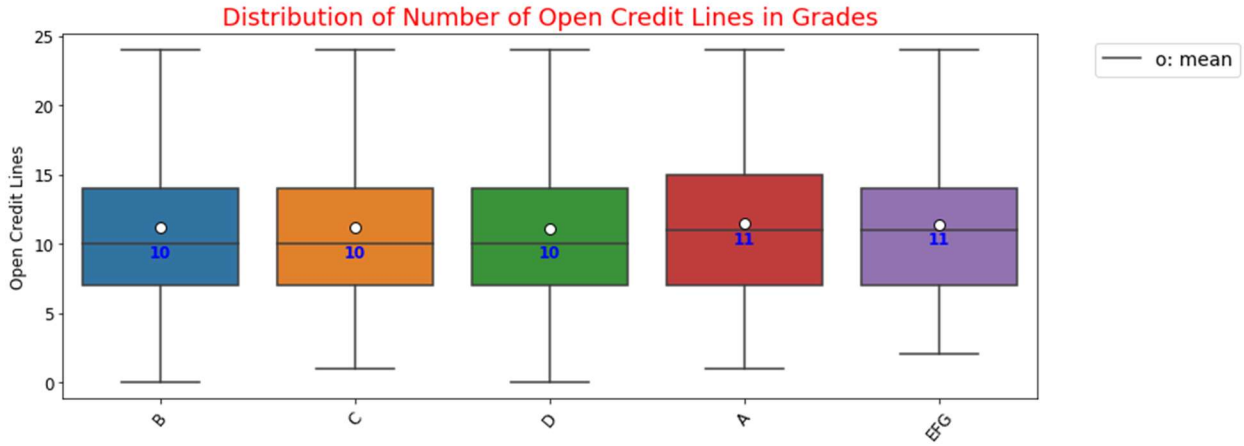


18. Grade, Total Number of Credit lines, and Open Credit Lines

Distribution of total number of credit lines for different grades are similar. Only among grade A loans, the median is greater by one compared with other grades.

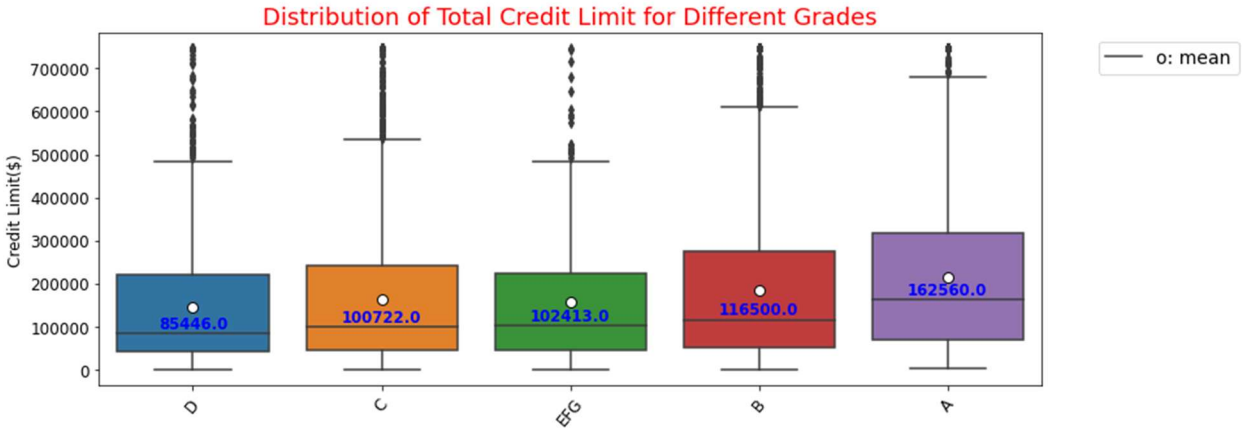
Almost 50% of total credit lines were open. Borrowers with grade A loans, had few more open credits than other grades.



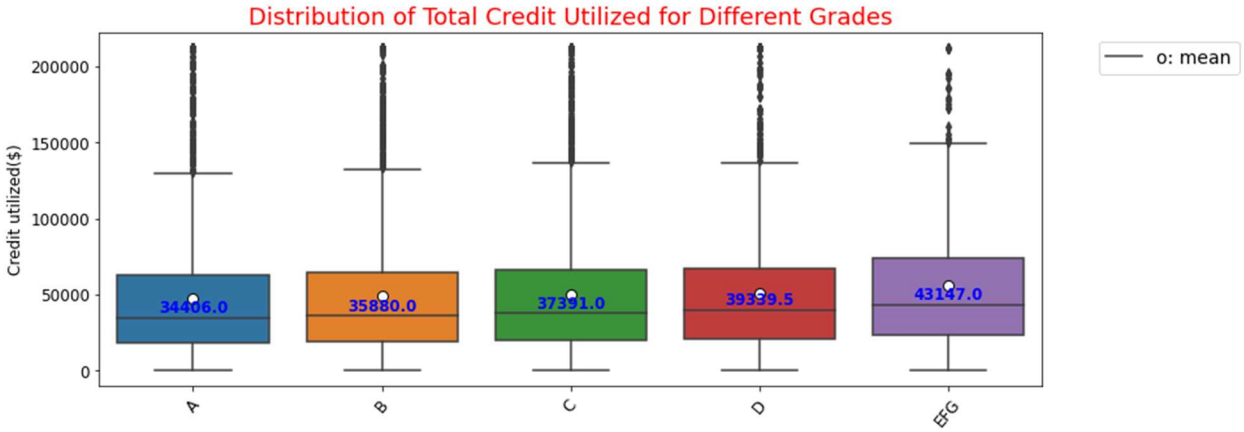


19. Grade and Total Credit Limit, and Total Credit Utilized

Median of total credit limit for grade A was higher that other grades by at least 64k dollars.



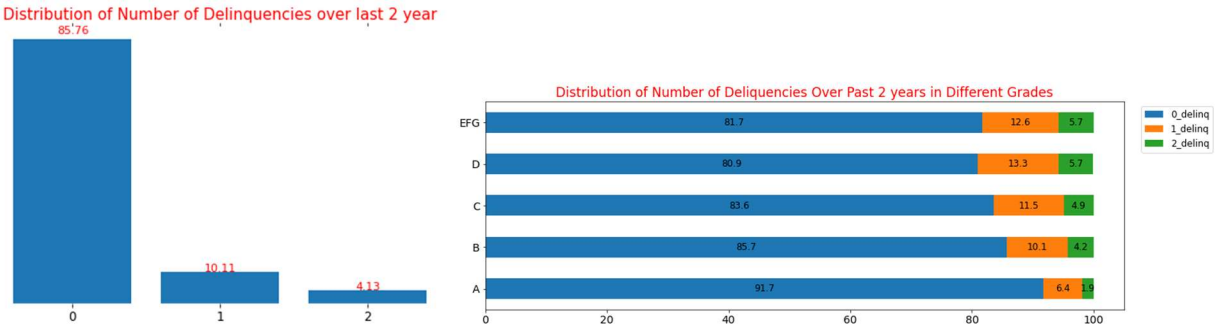
The lowest median of total credit utilized belonged to borrowers of loans with grade A. Borrowers of loans with grade EFG had median of credit utilized equaled to 43,147 $ which was at least 3800 $ greater than other groups.



20. Grade and Number of Delinquencies over Past 2 Years

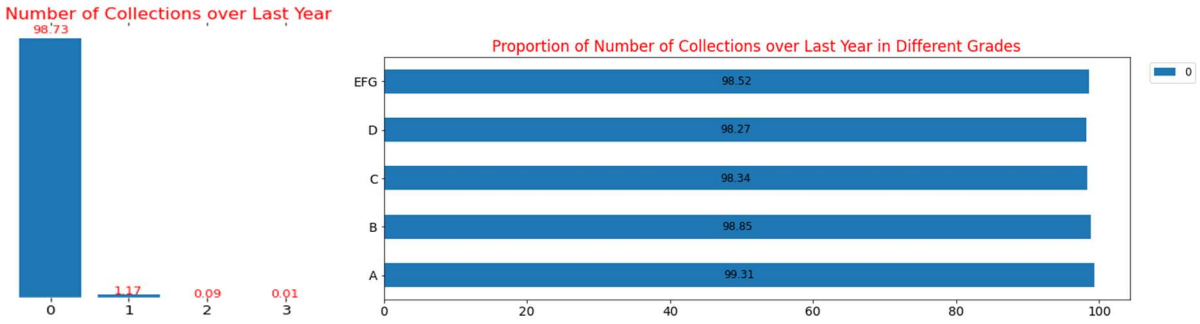
Only 4.13% of borrowers had delinquencies for 2 to 13 times.

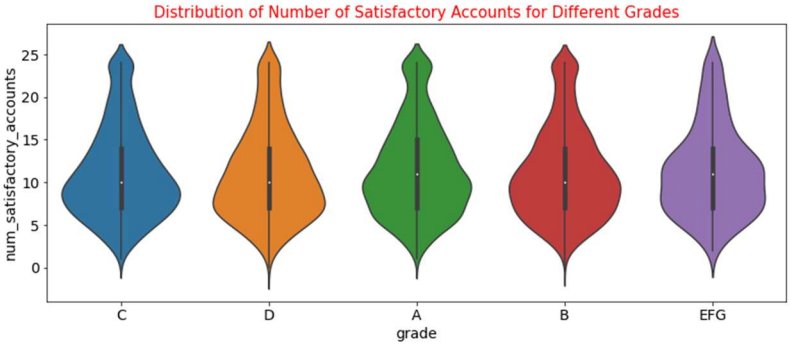
Borrowers with zero number of delinquencies had higher portions in receiving low risk loans. Borrowers with positive number of delinquencies had higher portions in receiving high risk loans.



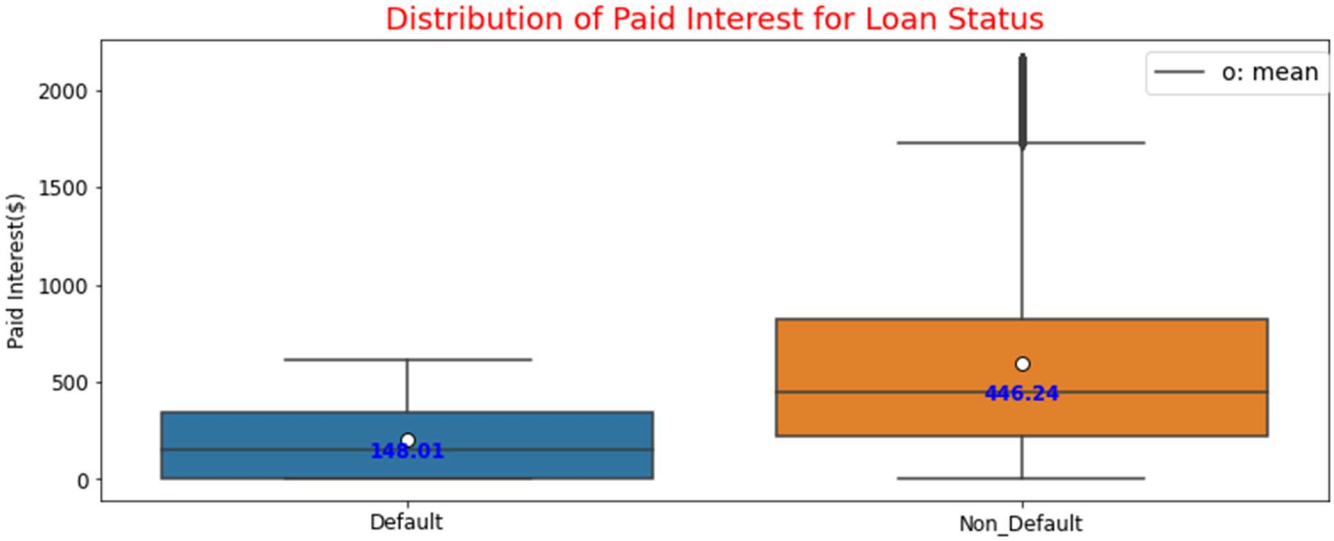
21. Grade and Number of Collections over Last 12 month

It doesn't seem that number of collections over last year impacted the grade, since almost 99% of borrowers had zero collections in that period.

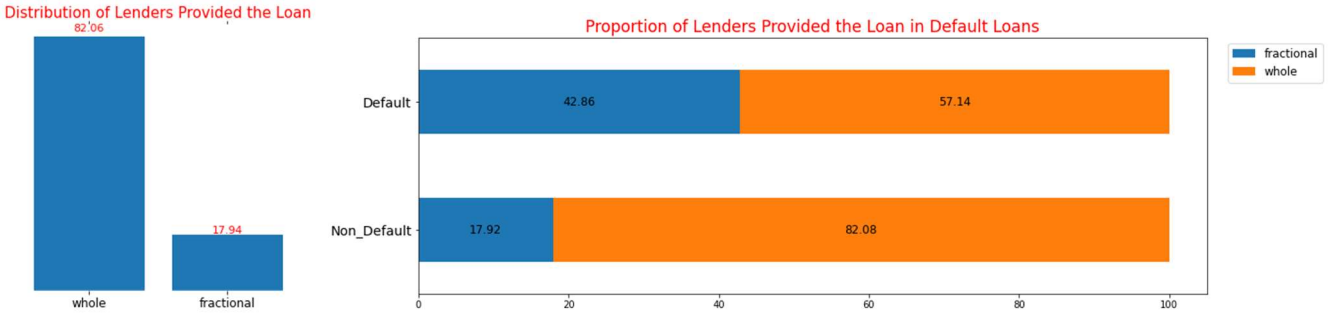




22. Default and Paid Interest



23. Default and initial_listing_status



24. Employment Title

This dataset had one column of emp_title. After cleaning and preprocessing about 8500 job title, I performed chunk noun to extract the root of title. There were 965 unique job titles.

After count vectorizing, Tfidf transforming, using SMOTE method on noun roots , two models MultinomialNB and SVC were tuned. None of models had good performance on predicting grades. So epm_title was not an important feature.

SVC

Accuracy on train set: 57.57

Confusion Matrix on train set

	predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G
A	810	246	136	468	143	22	2
B	384	609	95	525	192	18	4
C	420	275	436	491	190	13	2
D	314	186	106	1042	167	7	5
E	110	83	45	372	1204	13	0
F	29	57	3	168	125	1445	0
G	5	3	0	0	3	0	1816

Accuracy on test set: 22.48

Confusion Matrix on test set

	predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G
A	183	154	56	186	58	4	3
B	196	199	74	243	65	5	1
C	172	166	72	197	68	6	1
D	80	91	29	123	31	5	2
E	18	31	11	32	2	0	0
F	2	2	1	4	1	0	0
G	1	0	0	0	1	0	0

Modeling and Evaluation

First some categorical variables encoded through One-Hot encoding method. Then X, y as a set of predictors and an independent variable were set.

- Default Detection: Default Loan was set as y1 and other variables excluding loan status, balance and subgrades were set as X1. After recognizing a loan as default, the balance would be zero, so balance couldn't be a predictor for detecting default loan.
- Loan Grade Classification: Loan grade was set as y2 and other variables excluding sub grade, balance, paid interest, paid principle, total paid and paid principle to loan amount ratio were set as X2, another time default was set as y and other variables as X2.

Note:

- Grade and subgrade variables were explaining the same thing, so I drop subgrade.
- Some variables were associated to default detection not loan grade classification, since there were related to the period that loan received by borrowers. These variables were balance, paid interest, paid principle, total paid and paid principle to loan amount ratio which excluded from X2 in grade prediction and only used for default detection.

X1 and y1, also X2 and y2 were split 70/30 as train and test set. Binary encoding helped to encode categorical variables particularly with high dimensionality like state.

Since the dataset was highly imbalanced, using SMOTE method helped to oversample the minority classes and made the train set balanced. Then, I adopted the MinMax Scaler since some of independent variables were to some extend skewed.

Loan Grade Classification

The hyper parameters tunned for Multinomial Logistic Regression and Random Forest Classifier. Hyper parameter tunned through Grid Search while it used Repeated Stratified K Fold to make sure train sets and validation sets were balanced.

- **Multinomial Logistic Regression**

Multinomial had 65.02% accuracy on train set and 47.97% accuracy on test set.

```
Accuracy Score on Train Set : 65.02

Classification Report on Train Set
              precision    recall  f1-score   support

           A       0.67      0.73      0.70      2126
           B       0.47      0.48      0.47      2126
           C       0.43      0.40      0.41      2126
           D       0.52      0.45      0.48      2126
           E       0.70      0.63      0.66      2126
           F       0.77      0.88      0.82      2126
           G       0.94      0.98      0.96      2126

    accuracy                           0.65     14882
   macro avg       0.64      0.65      0.64     14882
weighted avg       0.64      0.65      0.64     14882

Confusion Matrix on  Train Set
```

	predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G
A	1548	457	89	32	0	0	0
B	467	1021	530	89	15	2	2
C	194	557	850	408	89	25	3
D	88	144	427	958	333	150	26
E	3	8	74	299	1340	359	43
F	0	0	4	55	149	1866	52
G	0	0	0	2	1	30	2093

```
Accuracy Score on Test Set : 47.97

Classification Report on Test Set
              precision    recall  f1-score   support

           A       0.66      0.68      0.67       738
           B       0.46      0.47      0.47       911
           C       0.42      0.46      0.44       796
           D       0.36      0.28      0.31       434
           E       0.24      0.20      0.22       100
           F       0.22      0.24      0.23        17
           G       0.00      0.00      0.00         4

    accuracy                           0.48      3000
   macro avg       0.34      0.33      0.33      3000
weighted avg       0.47      0.48      0.48      3000

Confusion Matrix on  Test Set
```

	predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G
A	504	189	42	3	0	0	0
B	182	426	251	44	4	1	3
C	61	233	364	118	14	4	2
D	17	67	180	121	42	6	1
E	2	3	29	43	20	3	0
F	0	1	1	8	3	4	0
G	0	1	0	1	2	0	0

- **Random Forest Classifier**

Random Forest Classifier had 100% accuracy on train set and 48% accuracy on test set.

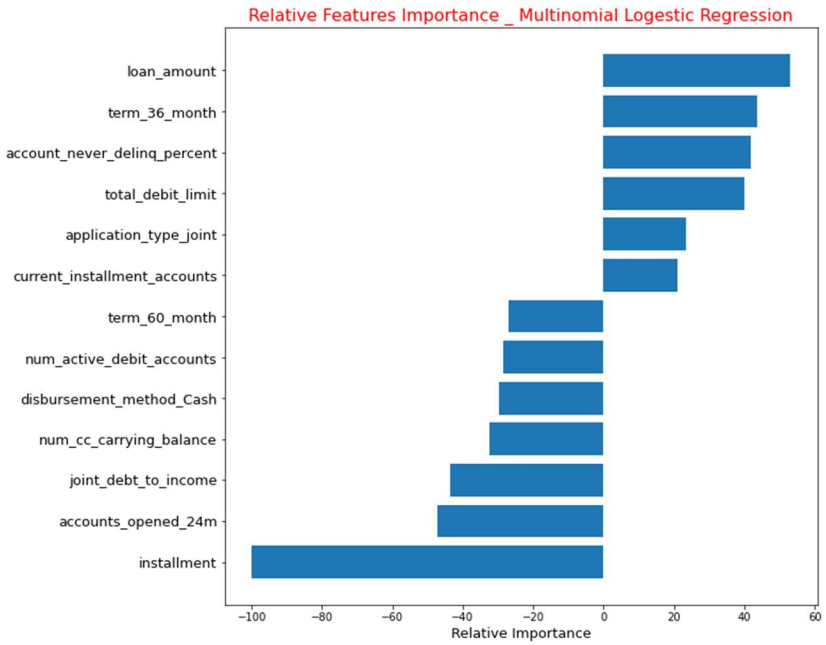
```
Accuracy Score on Train Set : 100.0

Classification Report on Train Set
              precision    recall  f1-score   support

           A       1.00      1.00      1.00      2126
           B       1.00      1.00      1.00      2126
           C       1.00      1.00      1.00      2126
           D       1.00      1.00      1.00      2126
           E       1.00      1.00      1.00      2126
           F       1.00      1.00      1.00      2126
           G       1.00      1.00      1.00      2126

    accuracy                           1.00     14882
   macro avg       1.00      1.00      1.00     14882
weighted avg       1.00      1.00      1.00     14882

Confusion Matrix on  Train Set
```

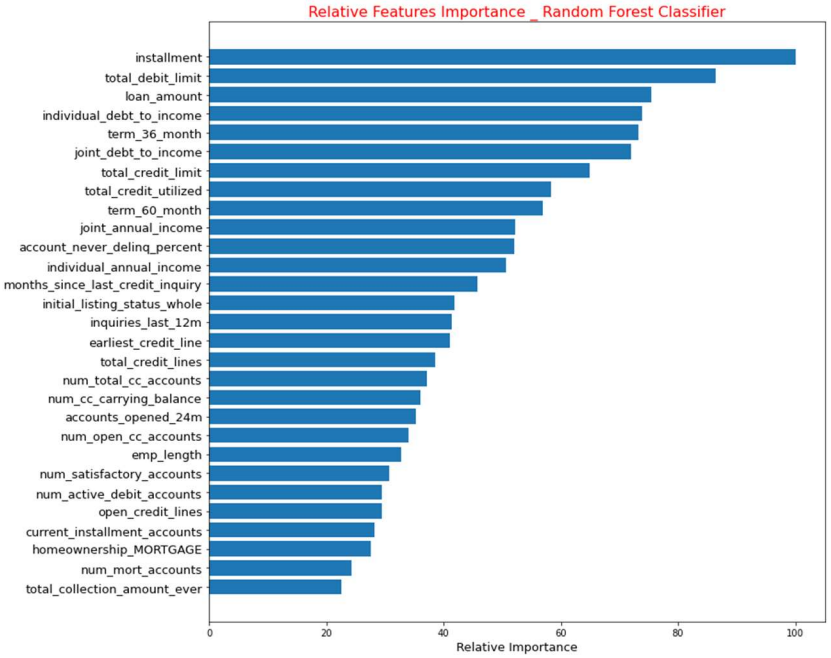
	predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G
A	2126	0	0	0	0	0	0
B	0	2126	0	0	0	0	0
C	0	0	2126	0	0	0	0
D	0	0	0	2126	0	0	0
E	0	0	0	0	2126	0	0
F	0	0	0	0	0	2126	0
G	0	0	0	0	0	0	2126

```
Accuracy Score on Test Set : 48.0

Classification Report on Test Set
              precision    recall  f1-score   support

           A       0.61      0.76      0.67       738
           B       0.45      0.44      0.45       911
           C       0.44      0.44      0.44       796
           D       0.36      0.29      0.32       434
           E       0.16      0.08      0.11       100
           F       0.00      0.00      0.00        17
           G       0.00      0.00      0.00         4

    accuracy                           0.48      3000
   macro avg       0.29      0.29      0.28      3000
weighted avg       0.46      0.48      0.47      3000

Confusion Matrix on  Test Set
```

	predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G
A	558	144	34	2	0	0	0
B	235	400	216	55	5	0	0
C	87	236	347	114	12	0	0
D	38	89	158	127	21	1	0
E	2	12	33	43	8	2	0
F	0	2	4	8	3	0	0
G	0	2	0	2	0	0	0

Random Forest Classifier had better performance in predicting the grades. The relative feature importance (greater than 4%) showed as I explored the most important features are:





The most important features which contribute to grade classification were:

- Installment
- Loan Amount
- Term 36 month
- Accounts opened over last 2 years
- Joint debt to income

Default Loan Detection

The hyper parameters tunned for Random Forest Classifier, Logistic Regression, and Light Gradient Boosting model through Grid Search and Bayesian Optimization. RF classifier, LGB model and Logistic Regression had recall rate of 1 on default loans (No False Negative).

Random Forest Classifier

Random Forest Clf on Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	6995
1	1.00	1.00	1.00	6505
accuracy			1.00	13500
macro avg	1.00	1.00	1.00	13500
weighted avg	1.00	1.00	1.00	13500

Random Forest Clf on Test

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2998
1	1.00	1.00	1.00	2
accuracy			1.00	3000
macro avg	1.00	1.00	1.00	3000
weighted avg	1.00	1.00	1.00	3000

	predicted_Non_Default	predicted_Default
Non_Default	6995	0
Default	0	6505

	predicted_Non_Default	predicted_Default
Non_Default	2998	0
Default	0	2

Light Gradient Boosting Model

lightgbm on Train

	precision	recall	f1-score	support
0	1.00	0.96	0.98	6995
1	0.96	1.00	0.98	6505
accuracy			0.98	13500
macro avg	0.98	0.98	0.98	13500
weighted avg	0.98	0.98	0.98	13500

lightgbm on Test

	precision	recall	f1-score	support
0	1.00	0.97	0.98	2998
1	0.02	1.00	0.04	2
accuracy			0.97	3000
macro avg	0.51	0.98	0.51	3000
weighted avg	1.00	0.97	0.98	3000

	predicted_Non_Default	predicted_Default
Non_Default	6740	255
Default	0	6505

	predicted_Non_Default	predicted_Default
Non_Default	2898	100
Default	0	2

Logistic Regression

Logistic Regression on Train

	precision	recall	f1-score	support
0	1.00	0.88	0.94	6995
1	0.89	1.00	0.94	6505
accuracy			0.94	13500
macro avg	0.94	0.94	0.94	13500
weighted avg	0.95	0.94	0.94	13500

Logistic Regression

	precision	recall	f1-score	support
0	1.00	0.88	0.94	2998
1	0.01	1.00	0.01	2
accuracy			0.88	3000
macro avg	0.50	0.94	0.47	3000
weighted avg	1.00	0.88	0.94	3000

	predicted_Non_Default	predicted_Default
Non_Default	6173	822
Default	0	6505

	predicted_Non_Default	predicted_Default
Non_Default	2642	356
Default	0	2

Linear SVC

LinearSVC on Train

	precision	recall	f1-score	support
0	0.78	0.96	0.86	6995
1	0.94	0.72	0.81	6505
accuracy			0.84	13500
macro avg	0.86	0.84	0.84	13500
weighted avg	0.86	0.84	0.84	13500

LinearSVC on Test

	precision	recall	f1-score	support
0	1.00	0.95	0.98	2998
1	0.01	1.00	0.03	2
accuracy			0.95	3000
macro avg	0.51	0.98	0.50	3000
weighted avg	1.00	0.95	0.98	3000

	predicted_Non_Default	predicted_Default
Non_Default	6714	281
Default	1851	4654

	predicted_Non_Default	predicted_Default
Non_Default	2859	139
Default	0	2

Ridge Classifier

Ridge Classifier on Train

	precision	recall	f1-score	support
0	1.00	0.94	0.97	6995
1	0.94	1.00	0.97	6505
accuracy			0.97	13500
macro avg	0.97	0.97	0.97	13500
weighted avg	0.97	0.97	0.97	13500

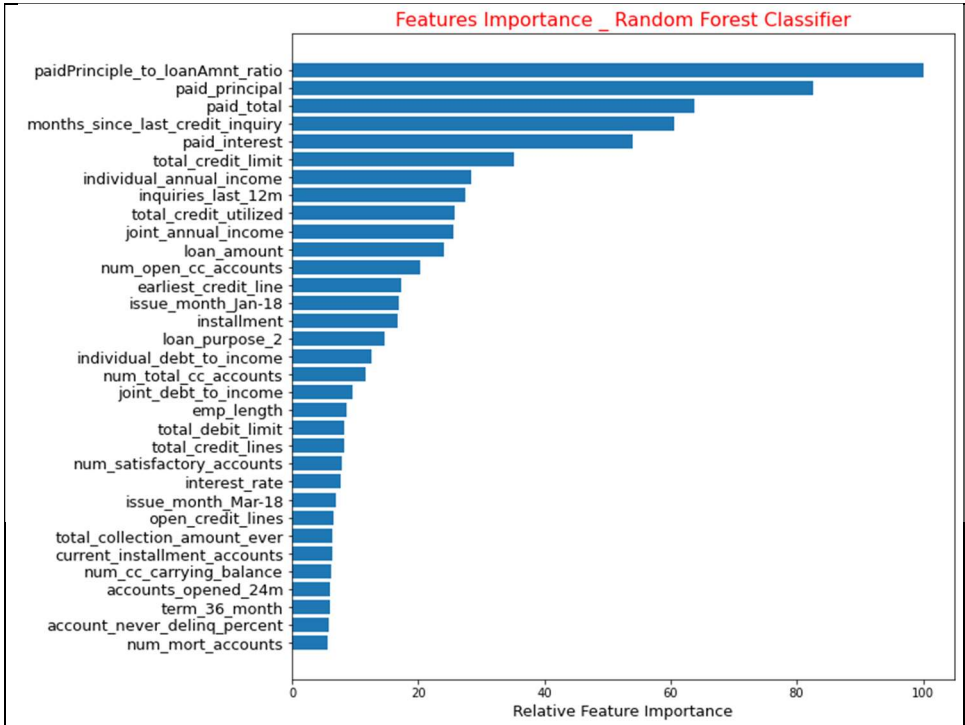
Ridge Classifier on Test

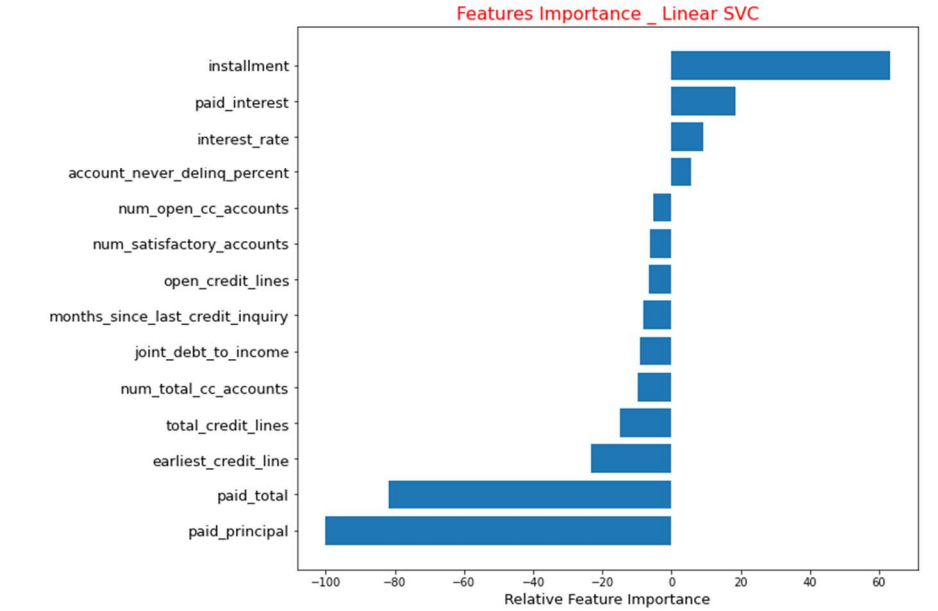
	precision	recall	f1-score	support
0	1.00	0.94	0.97	2998
1	0.01	1.00	0.02	2
accuracy			0.94	3000
macro avg	0.51	0.97	0.49	3000
weighted avg	1.00	0.94	0.97	3000

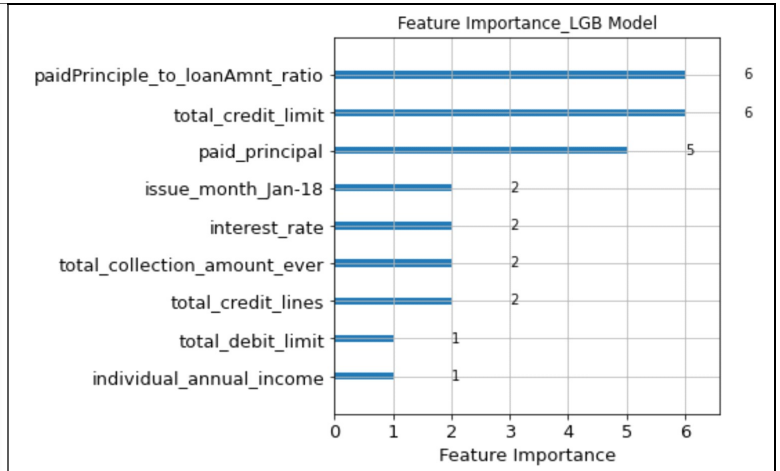
	predicted_Non_Default	predicted_Default
Non_Default	6556	439
Default	0	6505

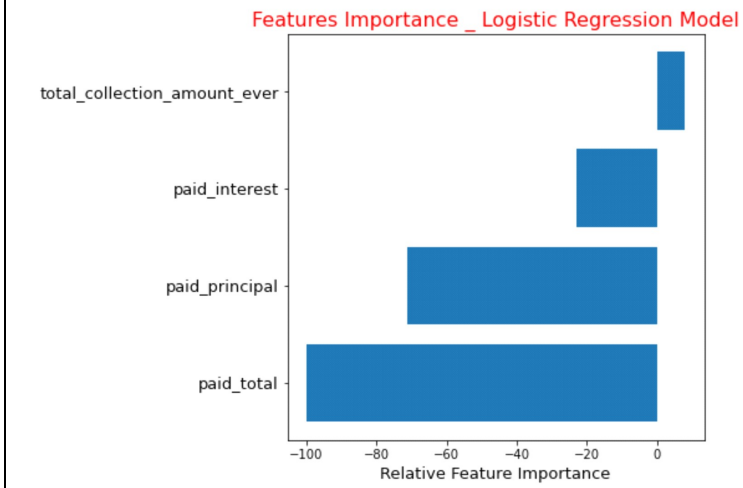
	predicted_Non_Default	predicted_Default
Non_Default	2804	194
Default	0	2

	model	Accuracy	Precision	Recall	F1-score	AUC
0	Random Forest Clf on Test	1	1	1	1	1
0	Random Forest Clf on Train	1	1	1	1	1
0	lightgbm on Train	0.98	0.98	1	0.98	0.98
0	lightgbm on Test	0.97	1	1	0.04	0.98
0	Logistic Regression on Train	0.94	0.95	1	0.94	0.94
0	Logistic Regression	0.88	1	1	0.01	0.94
0	LinearSVC on Train	0.84	0.86	0.72	0.81	0.84
0	LinearSVC on Test	0.95	1	1	0.03	0.98
0	Ridge Classifier on Train	0.97	0.97	1	0.97	0.97
0	Ridge Classifier on Test	0.94	1	1	0.02	0.97


Features Importance _ Random Forest Classifier
paidPrinciple_to_loanAmnt_ratio
paid_principal
paid_total
months_since_last_credit_inquiry
paid_interest
total_credit_limit
individual_annual_income
inquiries_last_12m
total_credit_utilized
joint_annual_income
loan_amount
num_open_cc_accounts
earliest_credit_line
issue_month_Jan-18
installment
loan_purpose_2
individual_debt_to_income
num_total_cc_accounts
joint_debt_to_income
emp_length
total_debit_limit
total_credit_lines
num_satisfactory_accounts
interest_rate
issue_month_Mar-18
open_credit_lines
total_collection_amount_ever
current_installment_accounts
num_cc_carrying_balance
accounts_opened_24m
term_36_month
account_never_delinq_percent
num_mort_accounts
0
20
40
60
80
100
Relative Feature Importance


Features Importance _ Linear SVC
installment
paid_interest
interest_rate
account_never_delinq_percent
num_open_cc_accounts
num_satisfactory_accounts
open_credit_lines
months_since_last_credit_inquiry
joint_debt_to_income
num_total_cc_accounts
total_credit_lines
earliest_credit_line
paid_total
paid_principal
−100
−80
−60
−40
−20
0
20
40
60
Relative Feature Importance


Feature Importance_LGB Model
paidPrinciple_to_loanAmnt_ratio
total_credit_limit
paid_principal
issue_month_Jan-18
interest_rate
total_collection_amount_ever
total_credit_lines
total_debit_limit
individual_annual_income
6
6
5
2
2
2
2
1
1
0
1
2
3
4
5
6
Feature Importance


Features Importance _ Logistic Regression Model
total_collection_amount_ever
paid_interest
paid_principal
paid_total
−100
−80
−60
−40
−20
0
Relative Feature Importance

The most important features in detecting default loans were:

- Paid principle to loan amount ratio
- Paid principle and Paid interest and Paid total,
- Total credit limit,
- Months since last inquiry,

Finding and Recommendation

Problem:

Minority classes in grade classification like grade E, F and G had very few sample

Minority class in default detection, default loan had only 7 sample out of 10 thousand samples

Even hyper tunned models couldn't train well on minority classes with such few samples

Next Step:

Gathering more sample from minority classes,

The more balance class, the better training the models and more reliable classifiers