

Grade and Default Loan Prediction Report

Prepared by: Ensieh Bahrami

Mentor: Alison Cossette

Contents

About Dataset	3
Wrangling and Exploration Report	3
1. Grade and Interest Rate	4
2. Grade, Loan Purpose, and Loan Amount	4
3. Default, Loan Status, Grade, and Interest Rate	5
4. Default, Loan Amount, and Loan Purposes	6
5. Grade and Term	7
6. Grade and Employment Length.....	8
7. Default and Balance.....	9
8. Default and Paid Principle	9
9. Default and Disbursement Method	10
10. Grade, State and Loan Status	10
11. Default, Grade, and Debt to Income Ratio	10
12. Grade and Total Debit Limit	11
13. Default and Joint.....	12
14. Debt to Income Joint.....	12
15. Grade and Inquiries over Last 12 Month	13
16. Grade and Number of Months since Last Credit Inquiry	13
17. Earliest Credit Line	14
18. Grade, Total Number of Credit lines, and Open Credit Lines	15
19. Grade and Total Credit Limit, and Total Credit Utilized	15
20. Grade and Number of Delinquencies over Past 2 Years.....	16
21. Grade and Number of Collections over Last 12 month	17
22. Default and Paid Interest.....	17
23. Default and initial_listing_status	18
24. Employment Title.....	18
Modeling	19
Loan Grade Classification	19
Default Loan Detection	22

About Dataset

Dataset of 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 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.

Wrangling and Exploration Report

My findings during data wrangling and exploration are as follows:

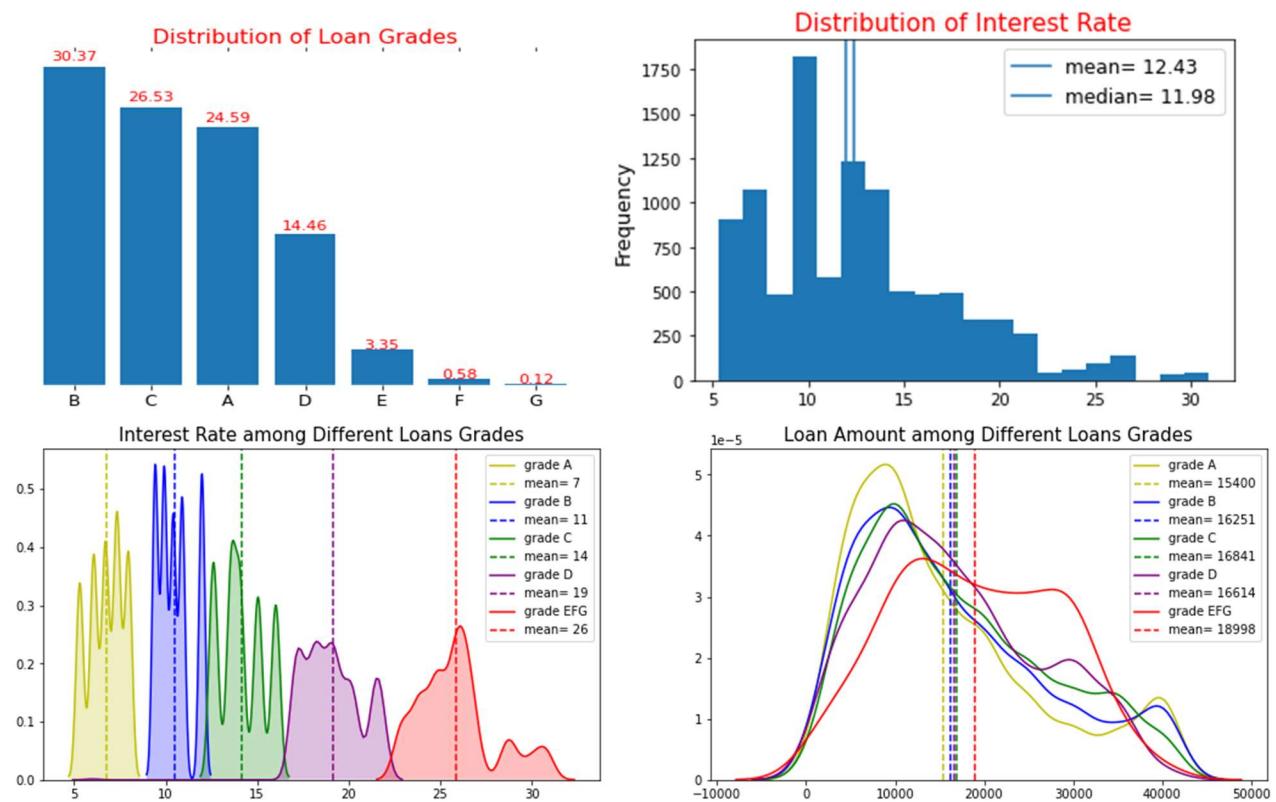
1. Grade and Interest Rate

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\$.

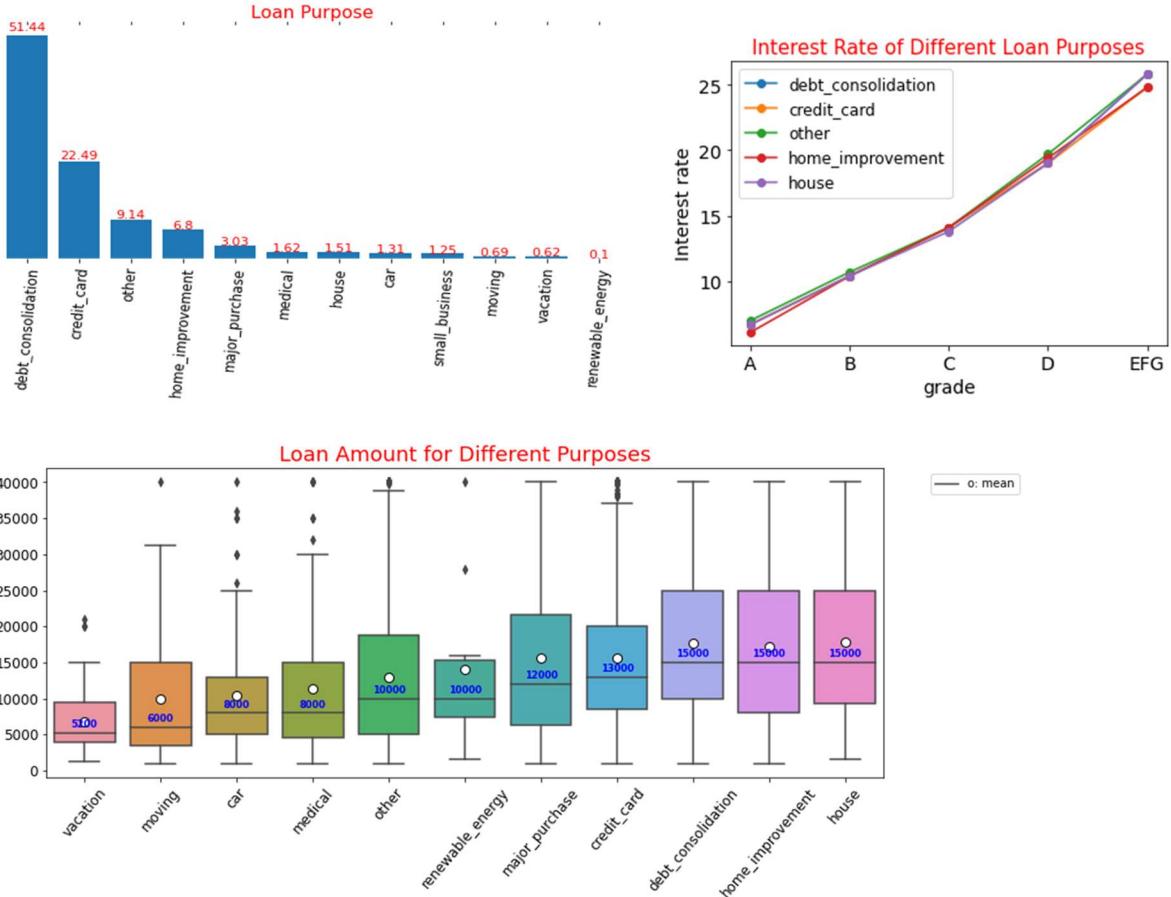


2. Grade, Loan Purpose, and Loan Amount

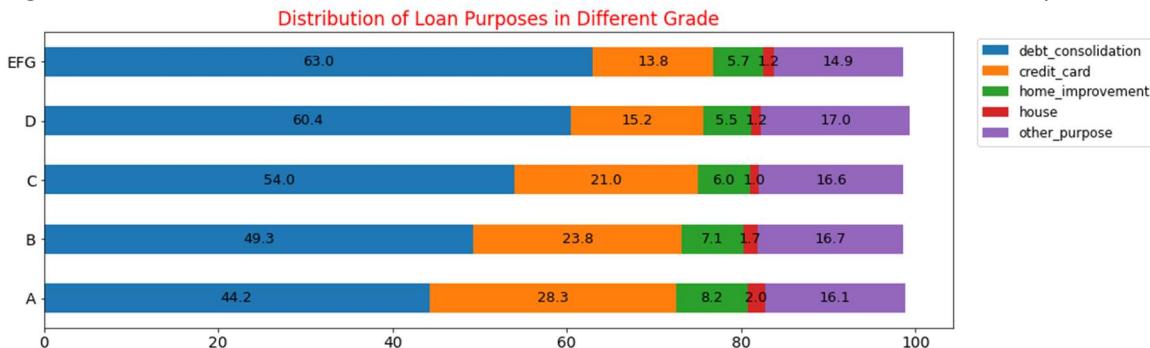
About 60% of loans used for debt consolidation, home improvement and buying house when 50% of them were 15k \$ or less.

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.

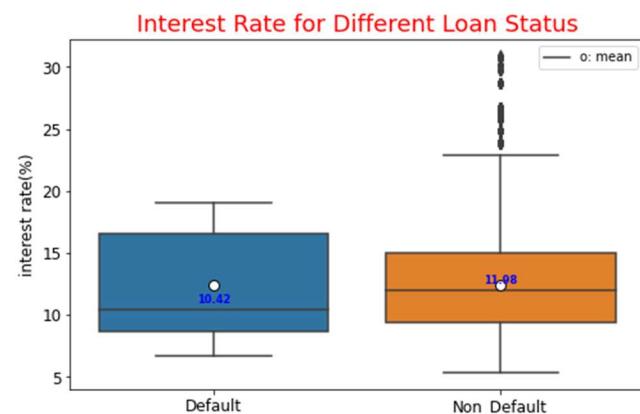
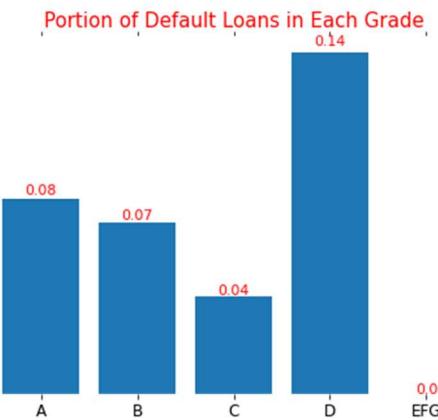
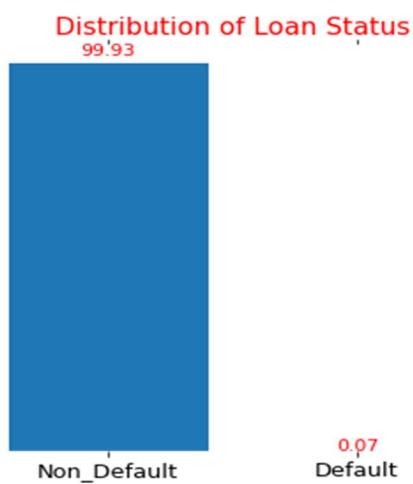
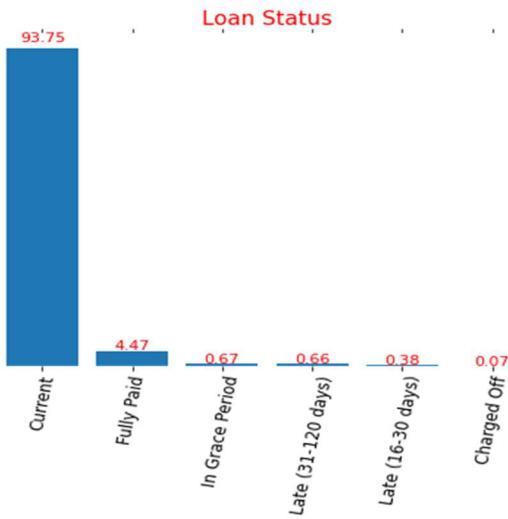


3. Default, Loan Status, Grade, and Interest Rate

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.

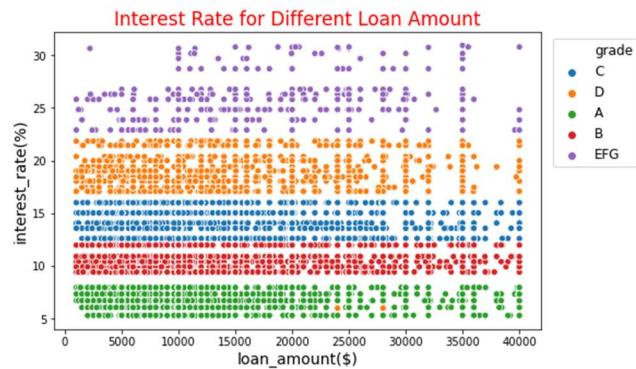
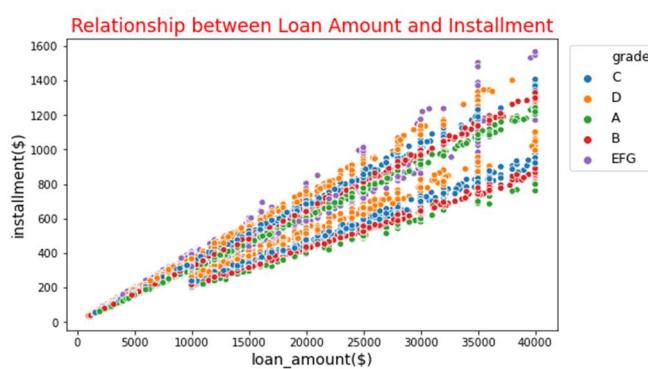
Median of interest rate of default loans were 0.66% lower than non-default loans.



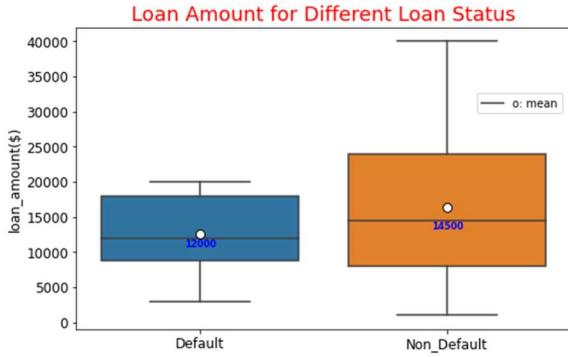
4. Default, Loan Amount, and Loan Purposes

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

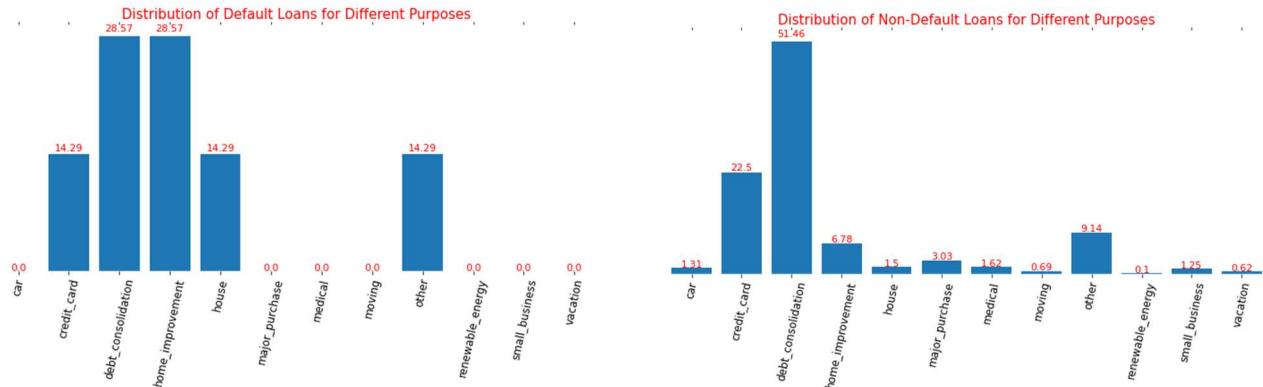
Interest rate was highly dependent to loan grade rather than loan amount.



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.

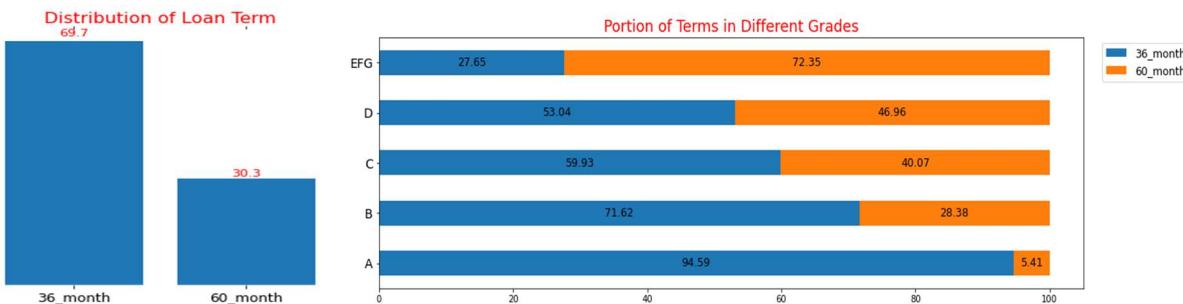


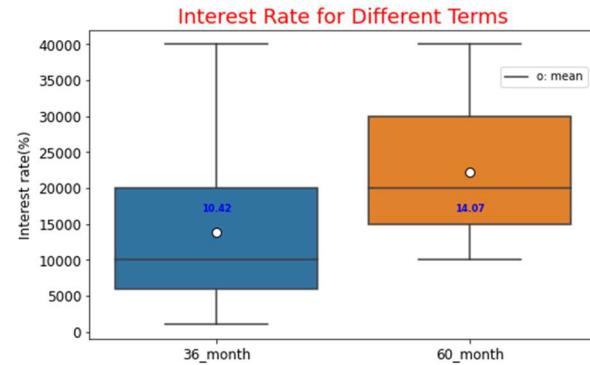
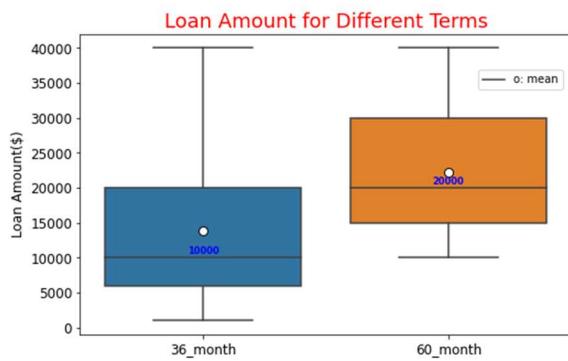
5. Grade and Term

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

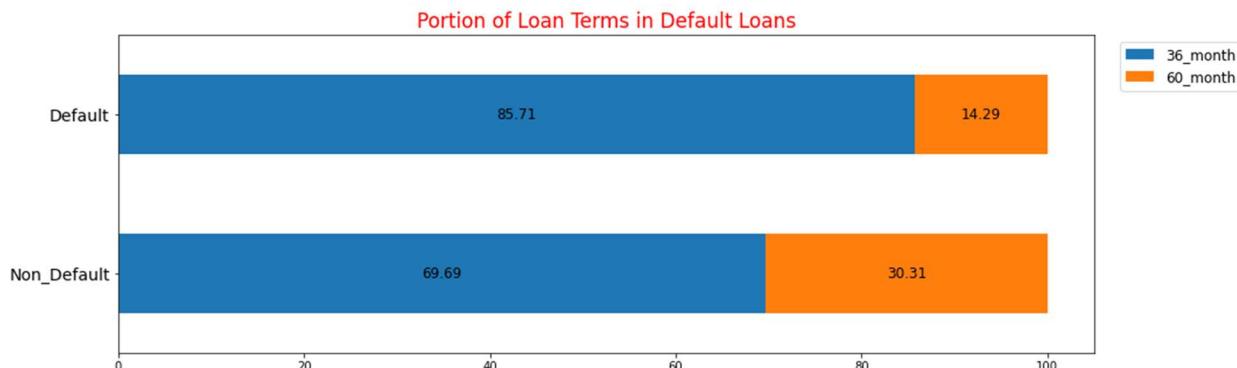
Majority of low risk loans had term of 36 month; while majority of high risk loans had term of 60 month.

Loan amount and interest rate for loans with term of 60 month were higher than loans with term of 36 month. Median of loan amount 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.

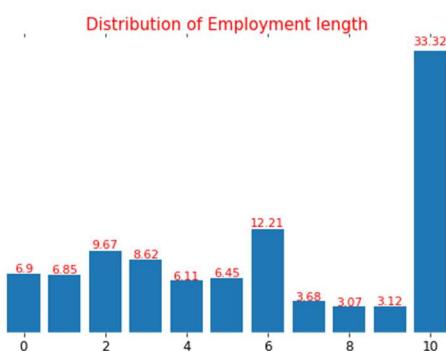


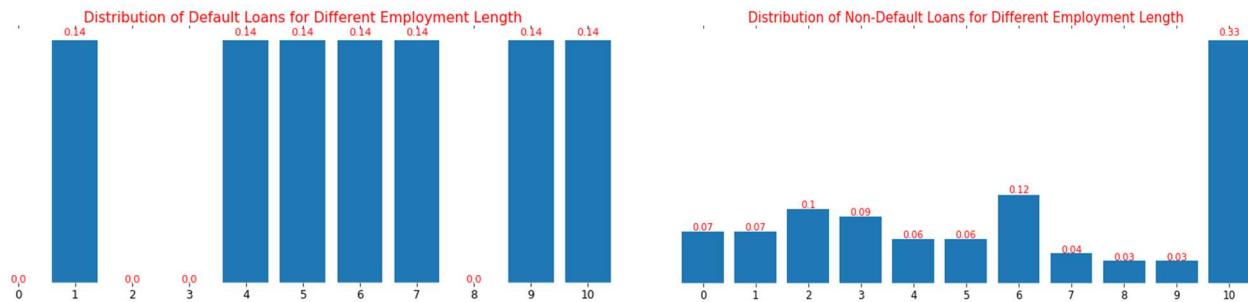
6. Grade and Employment Length

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 more 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.

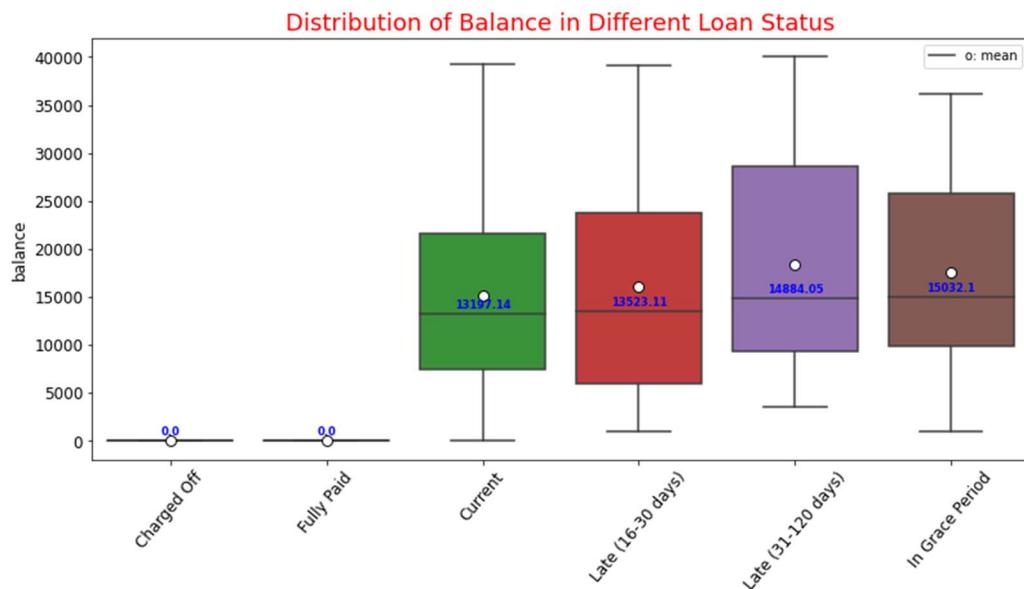
There weren't any default loans received by borrowers with less than 1, 1-2 ,2-3 and 8-9 years of employment.





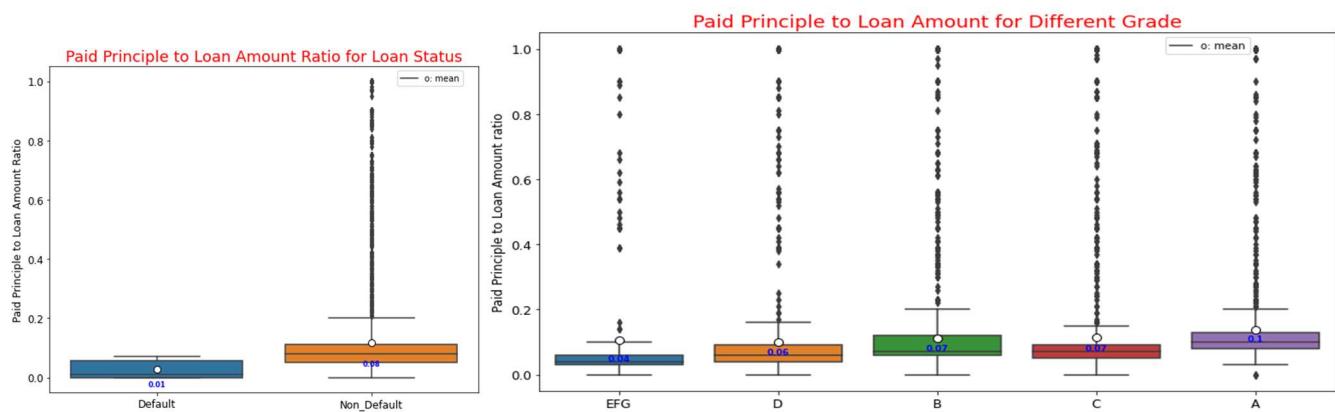
7. Default and Balance

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



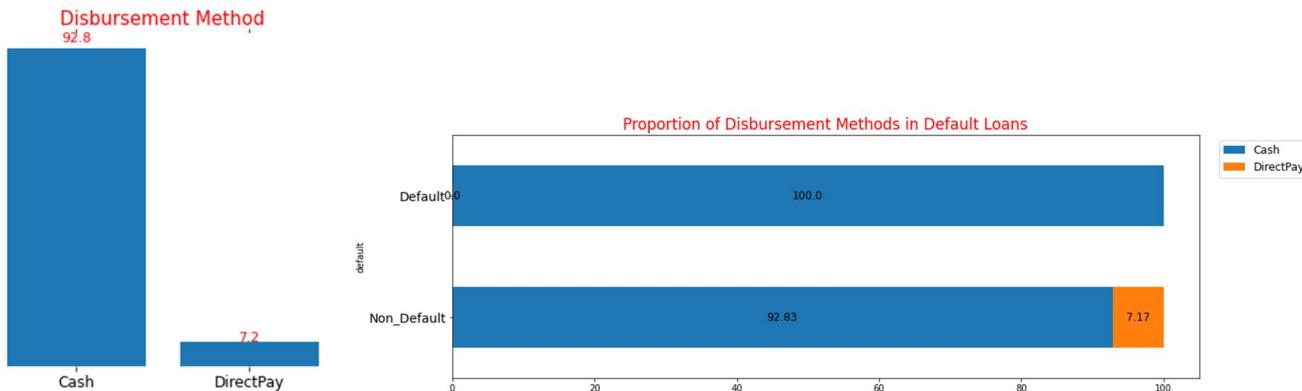
8. Default and Paid Principle

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



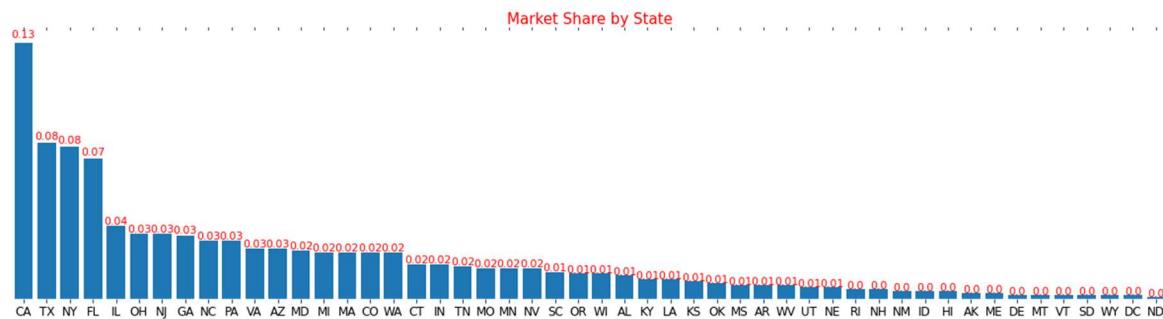
9. Default and Disbursement Method

All default loans were disbursed through cash.



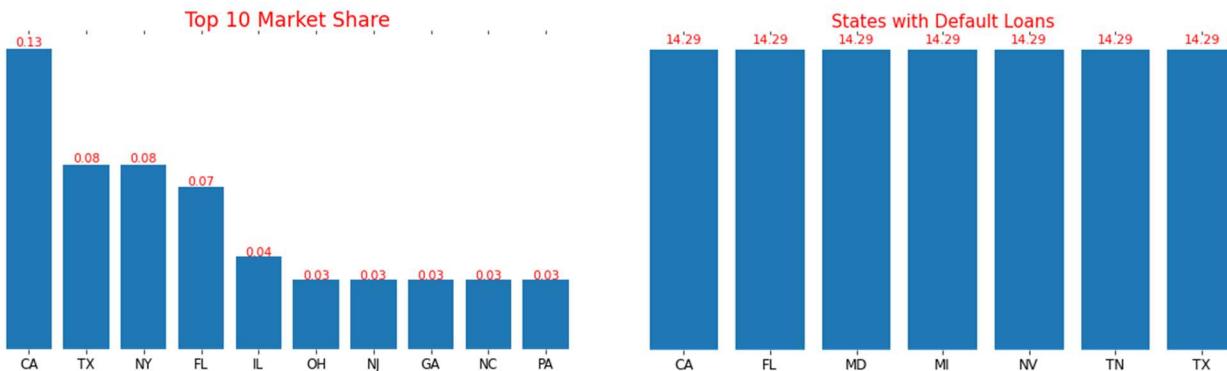
10. Grade, State and Loan Status

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%.



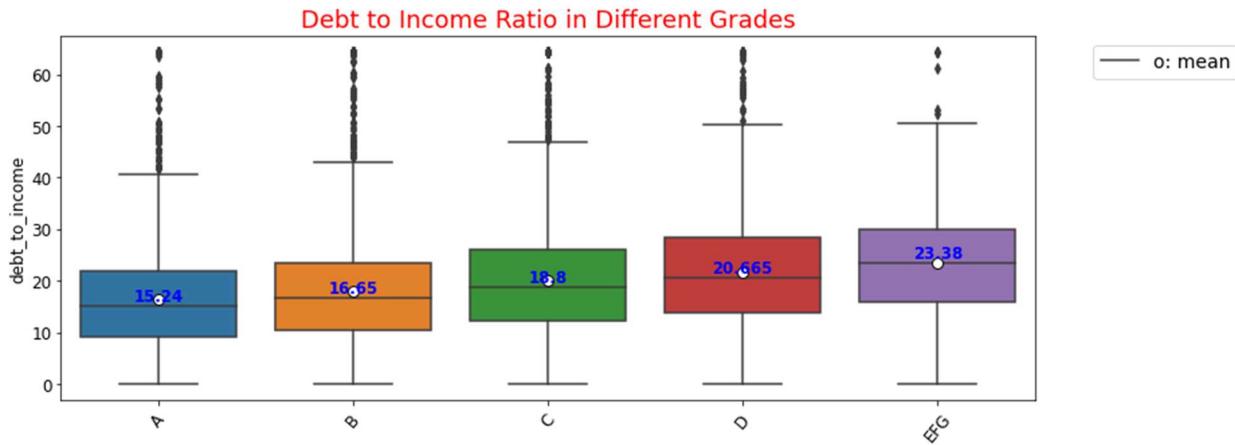
There were 10 states with market share higher than 3%. I considered those states as top markets.

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

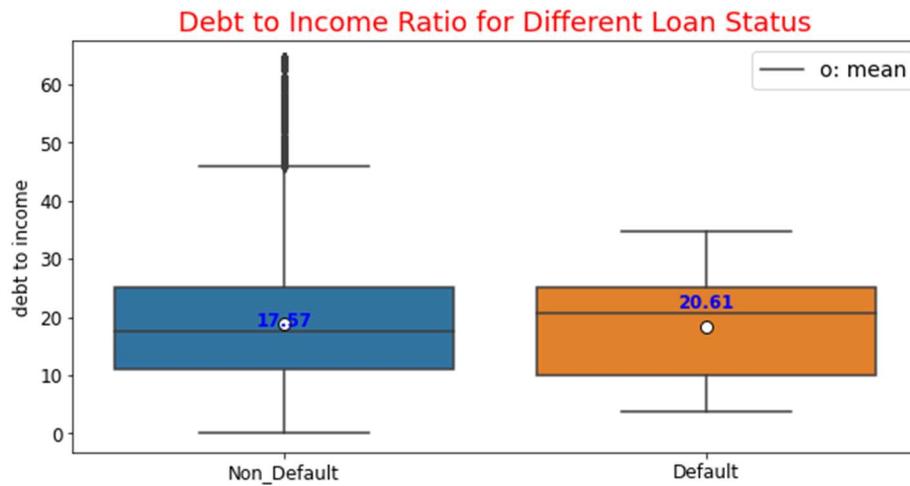


11. Default, Grade, and Debt to Income Ratio

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 than median for non-default.

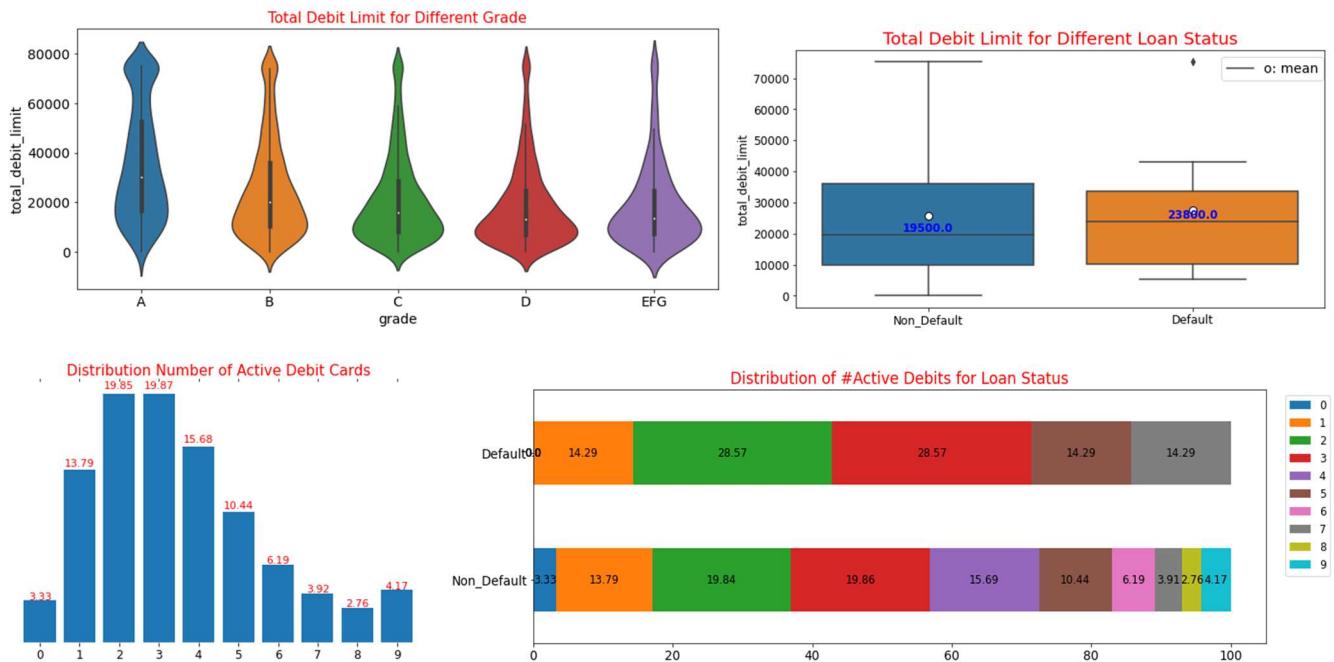


12. Grade and Total Debit Limit

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.

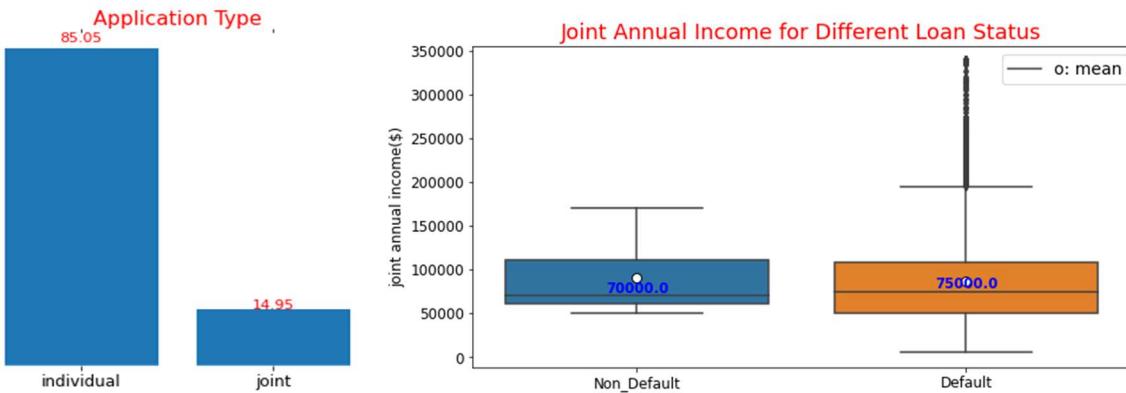
Median of Total Debit limit for borrowers with default loans was 4.3k dollars higher than median for borrowers with non-default.

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



13. Default and Joint

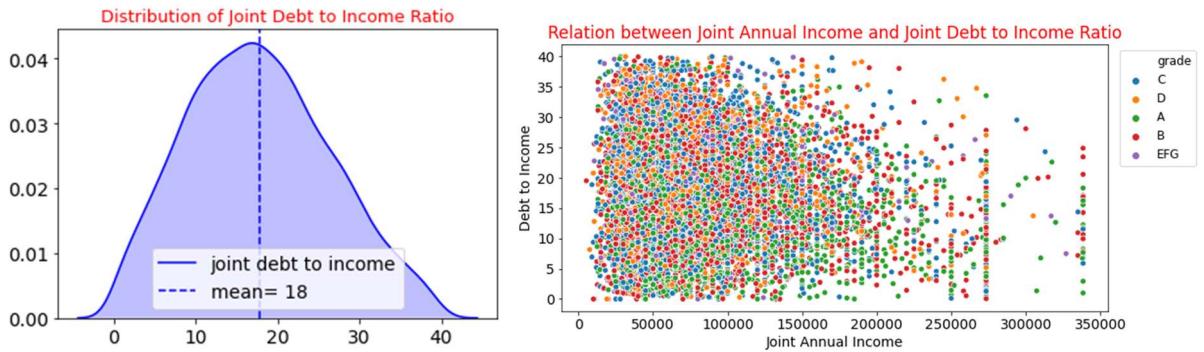
About 15% of borrowers had co-signer. The median of joint annual income for Borrowers with default loans was by 5k dollars higher than borrowers with non-default loans.



14. Debt to Income Joint

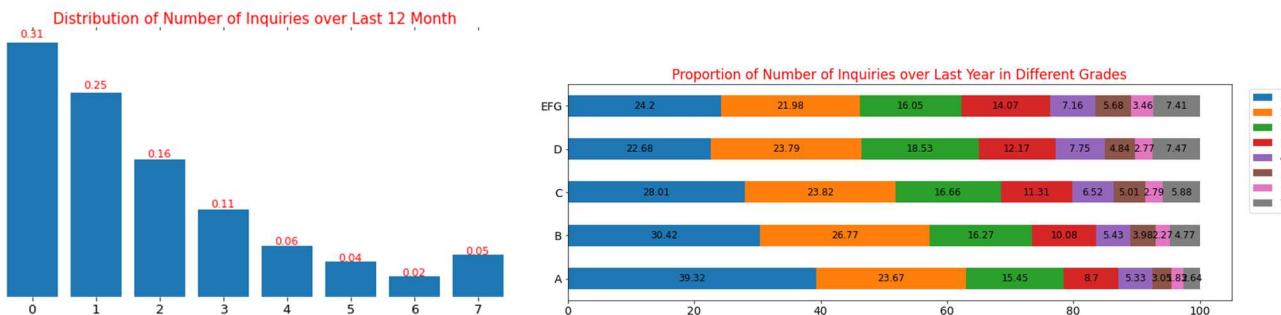
Those 85% of single borrowers had Null value in column of debt_to_income_joint, which makes sense to replace null values with the borrowers' debt_to_income'.

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.

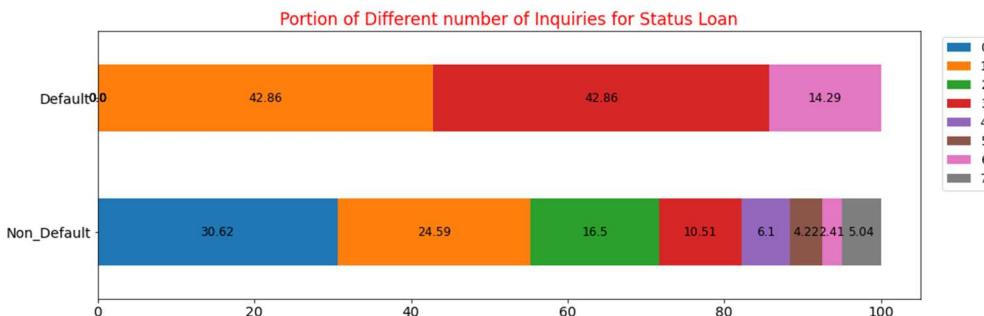


15. Grade and Inquiries over Last 12 Month

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

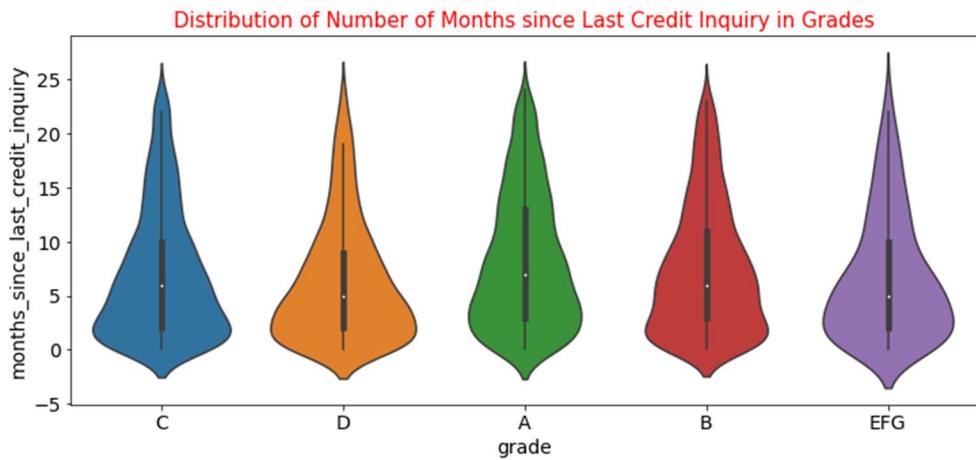


Majority of borrowers with default loans had one or three inquiries over last 2 years.



16. Grade and Number of Months since Last Credit Inquiry

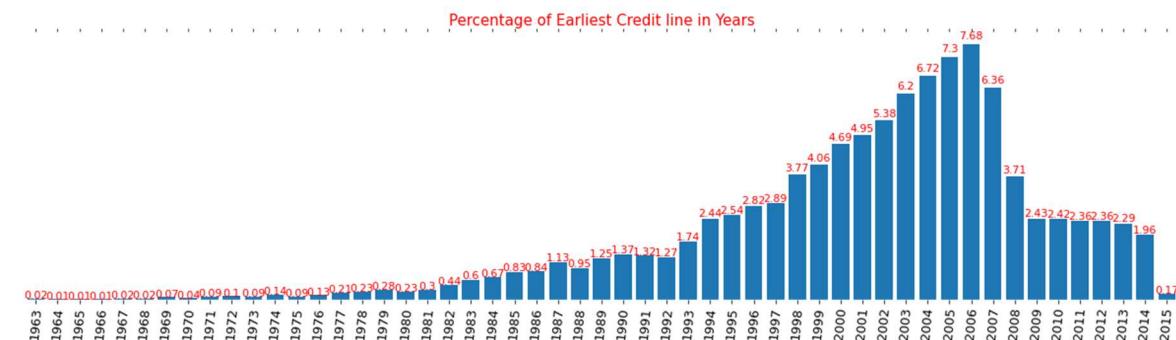
Fewer number of borrowers with positive number of months since last credit inquiry got grade A loans.



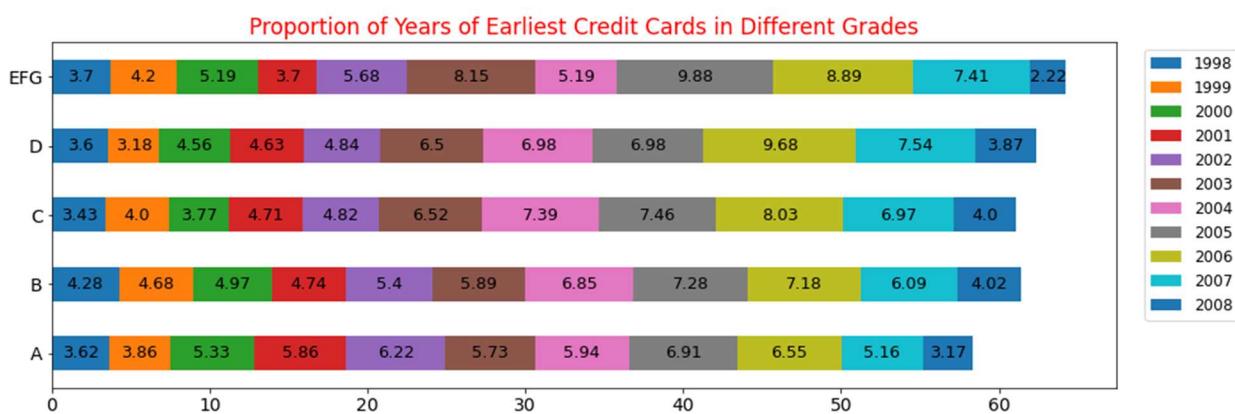
17. Earliest Credit Line

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.



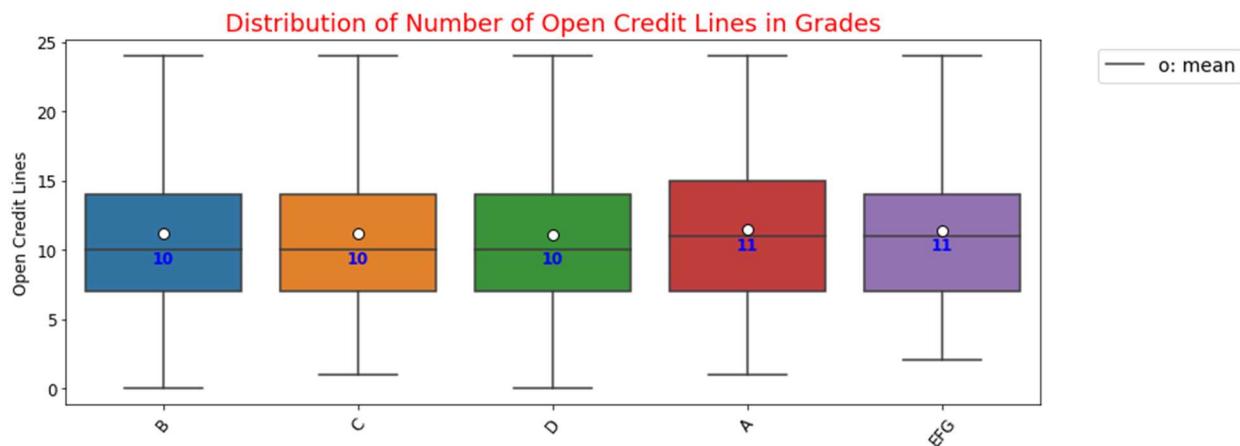
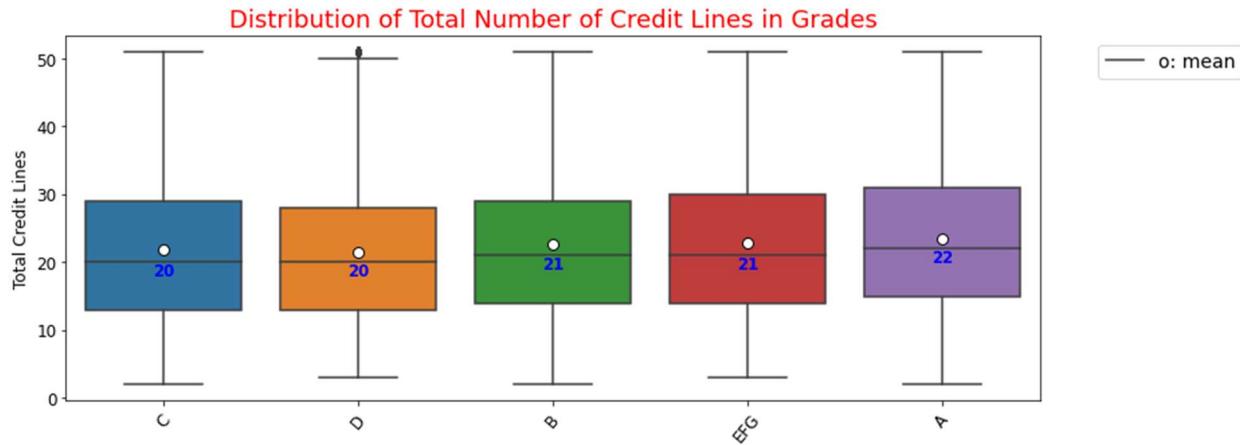
Borrowers who their earliest credit cards issued in 2003, 2006 and 2007 had bigger portion in high-risk loans than low risk ones.



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

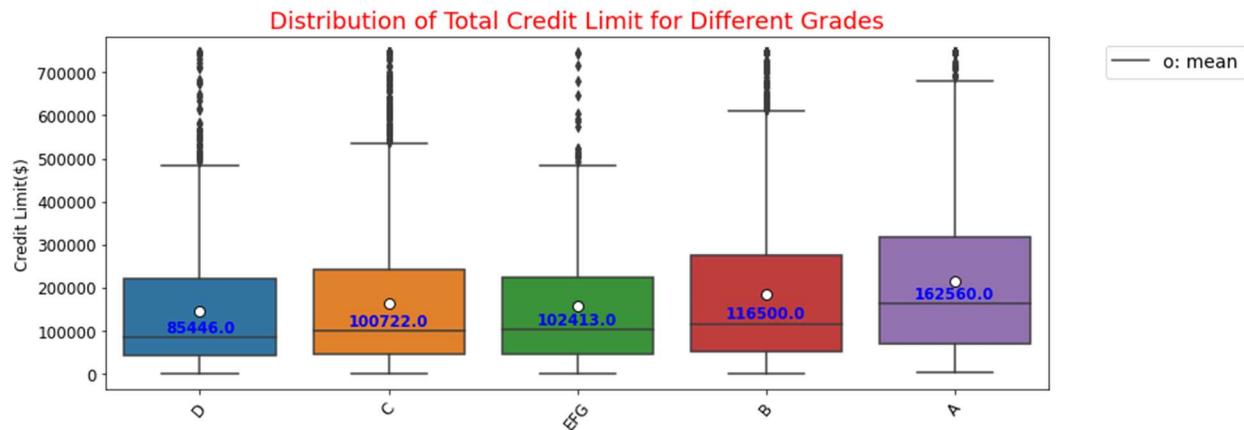
Distribution of total number of credit lines for different grades are similar. Only among grade A loans, the median is greater by one compared with other grades.

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

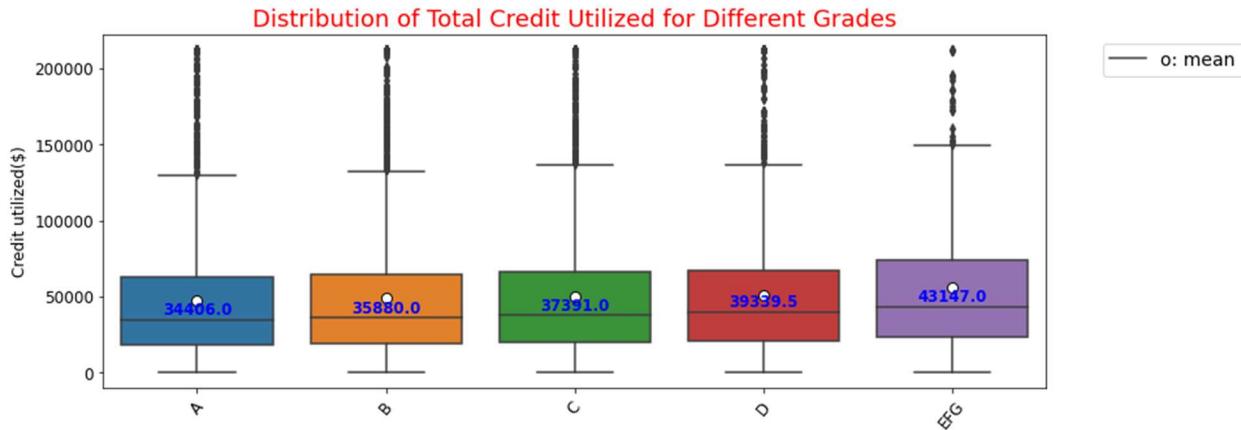


19. Grade and Total Credit Limit, and Total Credit Utilized

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



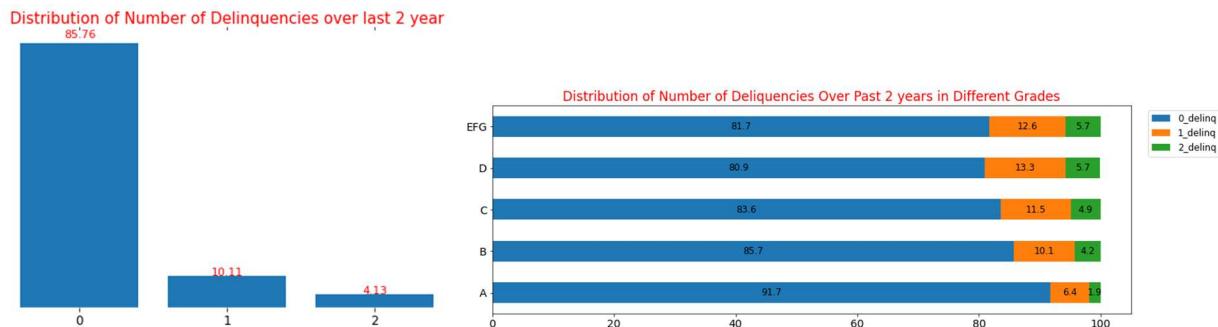
The lowest median of total credit utilized belonged to borrowers of loans with grade A. Borrowers of loans with grade EFG had median of credit utilized equaled to 43,147 \$ which was at least 3800 \$ greater than other groups.



20. Grade and Number of Delinquencies over Past 2 Years

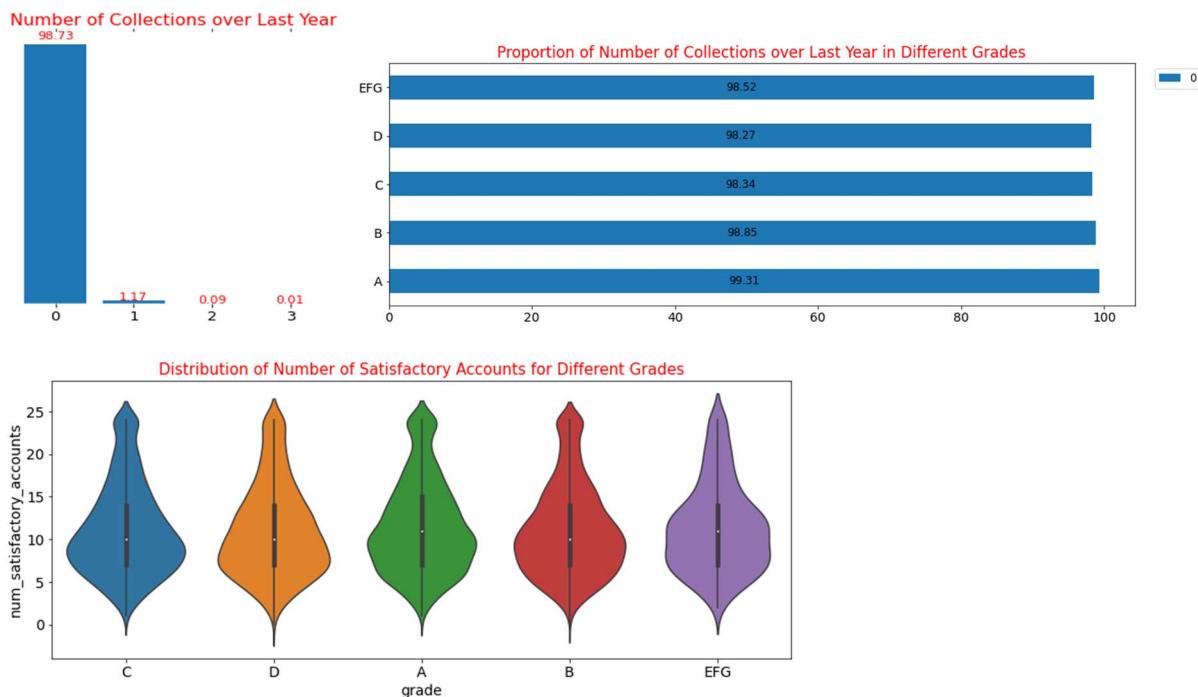
Only 4.13% of borrowers had delinquencies for 2 to 13 times.

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.

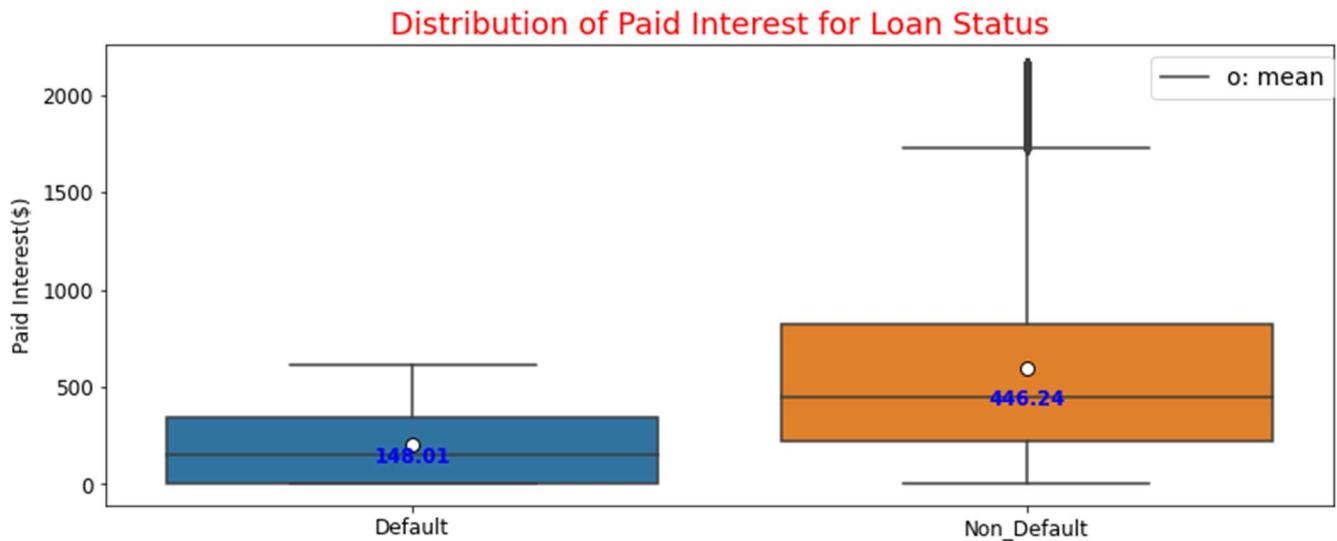


21. Grade and Number of Collections over Last 12 month

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.

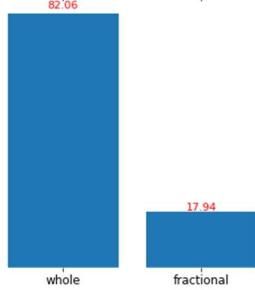


22. Default and Paid Interest

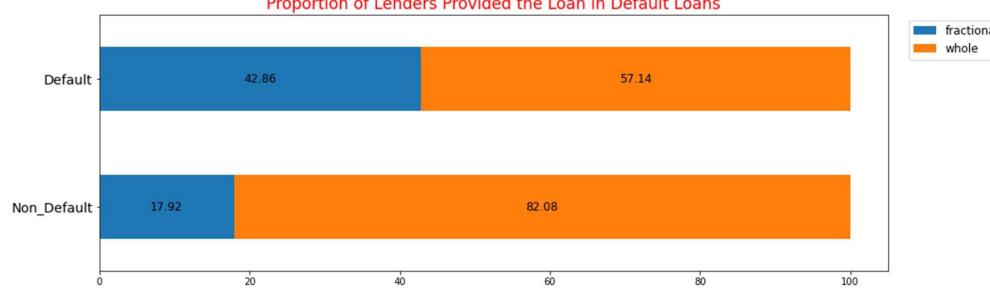


23. Default and initial_listing_status

Distribution of Lenders Provided the Loan



Proportion of Lenders Provided the Loan in Default Loans



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 emp_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
	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

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 and subgrades were set as X1.
- 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.

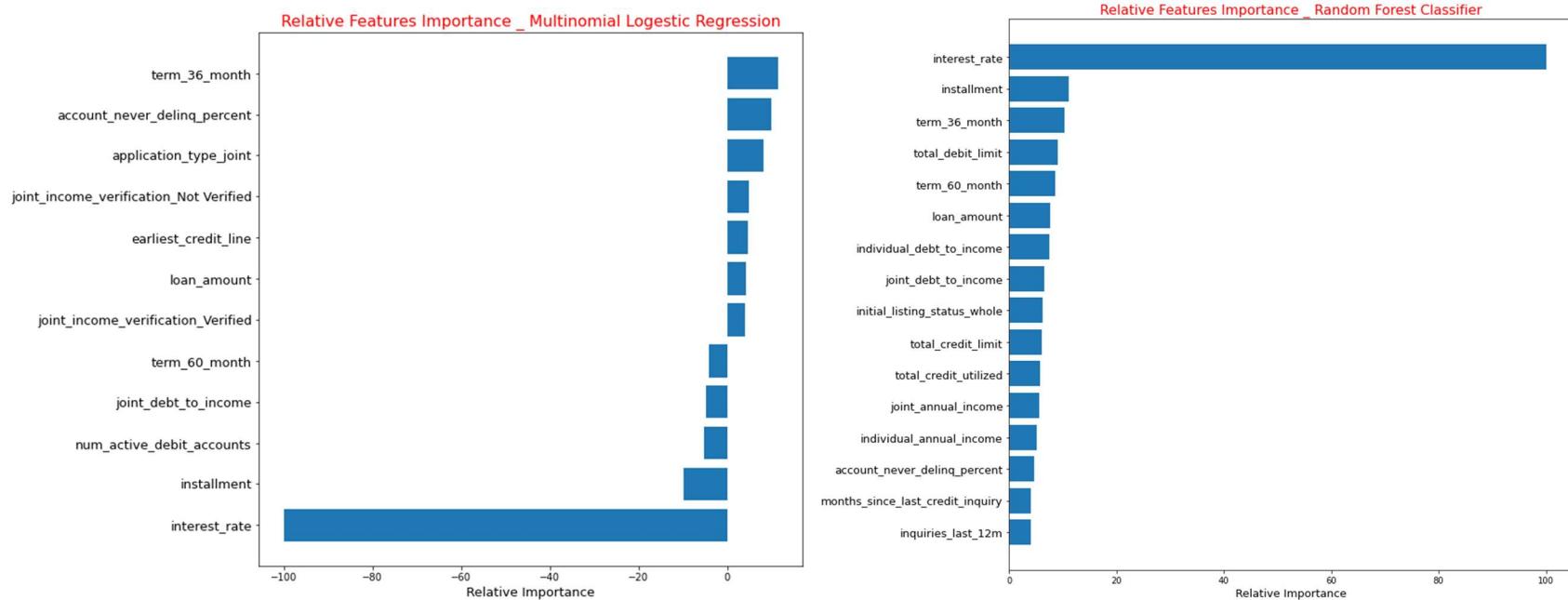
- Multinomial Logistic Regression

Multinomial had 95.2% accuracy on train set and 92.23% accuracy on test set.																																																																																																																				
Accuracy Score on Train set: 95.2				Accuracy Score on Train set: 92.23																																																																																																																
Classification Report on Train Set <table> <thead> <tr> <th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr> </thead> <tbody> <tr><td>A</td><td>0.98</td><td>0.99</td><td>0.99</td><td>2126</td></tr> <tr><td>B</td><td>0.95</td><td>0.94</td><td>0.95</td><td>2126</td></tr> <tr><td>C</td><td>0.93</td><td>0.92</td><td>0.92</td><td>2126</td></tr> <tr><td>D</td><td>0.94</td><td>0.94</td><td>0.94</td><td>2126</td></tr> <tr><td>E</td><td>0.96</td><td>0.93</td><td>0.95</td><td>2126</td></tr> <tr><td>F</td><td>0.95</td><td>0.96</td><td>0.96</td><td>2126</td></tr> <tr><td>G</td><td>0.95</td><td>0.98</td><td>0.96</td><td>2126</td></tr> <tr><td>accuracy</td><td></td><td></td><td>0.95</td><td>14882</td></tr> <tr><td>macro avg</td><td>0.95</td><td>0.95</td><td>0.95</td><td>14882</td></tr> <tr><td>weighted avg</td><td>0.95</td><td>0.95</td><td>0.95</td><td>14882</td></tr> </tbody> </table>					precision	recall	f1-score	support	A	0.98	0.99	0.99	2126	B	0.95	0.94	0.95	2126	C	0.93	0.92	0.92	2126	D	0.94	0.94	0.94	2126	E	0.96	0.93	0.95	2126	F	0.95	0.96	0.96	2126	G	0.95	0.98	0.96	2126	accuracy			0.95	14882	macro avg	0.95	0.95	0.95	14882	weighted avg	0.95	0.95	0.95	14882	Classification Report on Test Set <table> <thead> <tr> <th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr> </thead> <tbody> <tr><td>A</td><td>0.98</td><td>0.99</td><td>0.98</td><td>738</td></tr> <tr><td>B</td><td>0.94</td><td>0.93</td><td>0.94</td><td>911</td></tr> <tr><td>C</td><td>0.88</td><td>0.92</td><td>0.90</td><td>796</td></tr> <tr><td>D</td><td>0.90</td><td>0.86</td><td>0.88</td><td>434</td></tr> <tr><td>E</td><td>0.78</td><td>0.75</td><td>0.77</td><td>100</td></tr> <tr><td>F</td><td>0.78</td><td>0.41</td><td>0.54</td><td>17</td></tr> <tr><td>G</td><td>0.00</td><td>0.00</td><td>0.00</td><td>4</td></tr> <tr><td>accuracy</td><td></td><td></td><td>0.92</td><td>3000</td></tr> <tr><td>macro avg</td><td>0.75</td><td>0.70</td><td>0.71</td><td>3000</td></tr> <tr><td>weighted avg</td><td>0.92</td><td>0.92</td><td>0.92</td><td>3000</td></tr> </tbody> </table>				precision	recall	f1-score	support	A	0.98	0.99	0.98	738	B	0.94	0.93	0.94	911	C	0.88	0.92	0.90	796	D	0.90	0.86	0.88	434	E	0.78	0.75	0.77	100	F	0.78	0.41	0.54	17	G	0.00	0.00	0.00	4	accuracy			0.92	3000	macro avg	0.75	0.70	0.71	3000	weighted avg	0.92	0.92	0.92	3000
	precision	recall	f1-score	support																																																																																																																
A	0.98	0.99	0.99	2126																																																																																																																
B	0.95	0.94	0.95	2126																																																																																																																
C	0.93	0.92	0.92	2126																																																																																																																
D	0.94	0.94	0.94	2126																																																																																																																
E	0.96	0.93	0.95	2126																																																																																																																
F	0.95	0.96	0.96	2126																																																																																																																
G	0.95	0.98	0.96	2126																																																																																																																
accuracy			0.95	14882																																																																																																																
macro avg	0.95	0.95	0.95	14882																																																																																																																
weighted avg	0.95	0.95	0.95	14882																																																																																																																
	precision	recall	f1-score	support																																																																																																																
A	0.98	0.99	0.98	738																																																																																																																
B	0.94	0.93	0.94	911																																																																																																																
C	0.88	0.92	0.90	796																																																																																																																
D	0.90	0.86	0.88	434																																																																																																																
E	0.78	0.75	0.77	100																																																																																																																
F	0.78	0.41	0.54	17																																																																																																																
G	0.00	0.00	0.00	4																																																																																																																
accuracy			0.92	3000																																																																																																																
macro avg	0.75	0.70	0.71	3000																																																																																																																
weighted avg	0.92	0.92	0.92	3000																																																																																																																
Confusion Matrix on Train Set																																																																																																																				
<table> <thead> <tr> <th></th><th>predicted_A</th><th>predicted_B</th><th>predicted_C</th><th>predicted_D</th><th>predicted_E</th><th>predicted_F</th><th>predicted_G</th></tr> </thead> <tbody> <tr><td>A</td><td>2108</td><td>16</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>B</td><td>37</td><td>2009</td><td>80</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>C</td><td>0</td><td>85</td><td>1957</td><td>83</td><td>0</td><td>0</td><td>1</td></tr> <tr><td>D</td><td>3</td><td>0</td><td>73</td><td>1999</td><td>47</td><td>0</td><td>4</td></tr> <tr><td>E</td><td>0</td><td>0</td><td>0</td><td>50</td><td>1979</td><td>52</td><td>45</td></tr> <tr><td>F</td><td>0</td><td>0</td><td>0</td><td>0</td><td>33</td><td>2039</td><td>54</td></tr> <tr><td>G</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>48</td><td>2077</td></tr> </tbody> </table>								predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G	A	2108	16	2	0	0	0	0	B	37	2009	80	0	0	0	0	C	0	85	1957	83	0	0	1	D	3	0	73	1999	47	0	4	E	0	0	0	50	1979	52	45	F	0	0	0	0	33	2039	54	G	0	0	0	0	1	48	2077																																														
	predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G																																																																																																													
A	2108	16	2	0	0	0	0																																																																																																													
B	37	2009	80	0	0	0	0																																																																																																													
C	0	85	1957	83	0	0	1																																																																																																													
D	3	0	73	1999	47	0	4																																																																																																													
E	0	0	0	50	1979	52	45																																																																																																													
F	0	0	0	0	33	2039	54																																																																																																													
G	0	0	0	0	1	48	2077																																																																																																													
Confusion Matrix on Test Set																																																																																																																				
<table> <thead> <tr> <th></th><th>predicted_A</th><th>predicted_B</th><th>predicted_C</th><th>predicted_D</th><th>predicted_E</th><th>predicted_F</th><th>predicted_G</th></tr> </thead> <tbody> <tr><td>A</td><td>729</td><td>9</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>B</td><td>17</td><td>845</td><td>49</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>C</td><td>0</td><td>42</td><td>736</td><td>18</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>D</td><td>0</td><td>0</td><td>51</td><td>375</td><td>8</td><td>0</td><td>0</td></tr> <tr><td>E</td><td>0</td><td>0</td><td>0</td><td>24</td><td>75</td><td>1</td><td>0</td></tr> <tr><td>F</td><td>0</td><td>0</td><td>0</td><td>0</td><td>10</td><td>7</td><td>0</td></tr> <tr><td>G</td><td>0</td><td>0</td><td>0</td><td>0</td><td>3</td><td>1</td><td>0</td></tr> </tbody> </table>								predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G	A	729	9	0	0	0	0	0	B	17	845	49	0	0	0	0	C	0	42	736	18	0	0	0	D	0	0	51	375	8	0	0	E	0	0	0	24	75	1	0	F	0	0	0	0	10	7	0	G	0	0	0	0	3	1	0																																														
	predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G																																																																																																													
A	729	9	0	0	0	0	0																																																																																																													
B	17	845	49	0	0	0	0																																																																																																													
C	0	42	736	18	0	0	0																																																																																																													
D	0	0	51	375	8	0	0																																																																																																													
E	0	0	0	24	75	1	0																																																																																																													
F	0	0	0	0	10	7	0																																																																																																													
G	0	0	0	0	3	1	0																																																																																																													

- Random Forest Classifier

Random Forest Classifier had 100% accuracy on train set and 97.33% accuracy on test set.																																																																																																																				
Accuracy Score on Train set: 100.0				Accuracy Score on Train set: 97.33																																																																																																																
Classification Report on Train Set <table> <thead> <tr> <th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr> </thead> <tbody> <tr><td>A</td><td>1.00</td><td>1.00</td><td>1.00</td><td>2126</td></tr> <tr><td>B</td><td>1.00</td><td>1.00</td><td>1.00</td><td>2126</td></tr> <tr><td>C</td><td>1.00</td><td>1.00</td><td>1.00</td><td>2126</td></tr> <tr><td>D</td><td>1.00</td><td>1.00</td><td>1.00</td><td>2126</td></tr> <tr><td>E</td><td>1.00</td><td>1.00</td><td>1.00</td><td>2126</td></tr> <tr><td>F</td><td>1.00</td><td>1.00</td><td>1.00</td><td>2126</td></tr> <tr><td>G</td><td>1.00</td><td>1.00</td><td>1.00</td><td>2126</td></tr> <tr><td>accuracy</td><td></td><td></td><td>1.00</td><td>14882</td></tr> <tr><td>macro avg</td><td>1.00</td><td>1.00</td><td>1.00</td><td>14882</td></tr> <tr><td>weighted avg</td><td>1.00</td><td>1.00</td><td>1.00</td><td>14882</td></tr> </tbody> </table>					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	Classification Report on Test Set <table> <thead> <tr> <th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr> </thead> <tbody> <tr><td>A</td><td>1.00</td><td>0.97</td><td>0.99</td><td>738</td></tr> <tr><td>B</td><td>0.98</td><td>1.00</td><td>0.99</td><td>911</td></tr> <tr><td>C</td><td>0.95</td><td>0.99</td><td>0.97</td><td>796</td></tr> <tr><td>D</td><td>0.96</td><td>0.93</td><td>0.95</td><td>434</td></tr> <tr><td>E</td><td>0.99</td><td>0.82</td><td>0.90</td><td>100</td></tr> <tr><td>F</td><td>0.94</td><td>0.88</td><td>0.91</td><td>17</td></tr> <tr><td>G</td><td>1.00</td><td>0.50</td><td>0.67</td><td>4</td></tr> <tr><td>accuracy</td><td></td><td></td><td>0.97</td><td>3000</td></tr> <tr><td>macro avg</td><td>0.97</td><td>0.87</td><td>0.91</td><td>3000</td></tr> <tr><td>weighted avg</td><td>0.97</td><td>0.97</td><td>0.97</td><td>3000</td></tr> </tbody> </table>				precision	recall	f1-score	support	A	1.00	0.97	0.99	738	B	0.98	1.00	0.99	911	C	0.95	0.99	0.97	796	D	0.96	0.93	0.95	434	E	0.99	0.82	0.90	100	F	0.94	0.88	0.91	17	G	1.00	0.50	0.67	4	accuracy			0.97	3000	macro avg	0.97	0.87	0.91	3000	weighted avg	0.97	0.97	0.97	3000
	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																																																																																																																
	precision	recall	f1-score	support																																																																																																																
A	1.00	0.97	0.99	738																																																																																																																
B	0.98	1.00	0.99	911																																																																																																																
C	0.95	0.99	0.97	796																																																																																																																
D	0.96	0.93	0.95	434																																																																																																																
E	0.99	0.82	0.90	100																																																																																																																
F	0.94	0.88	0.91	17																																																																																																																
G	1.00	0.50	0.67	4																																																																																																																
accuracy			0.97	3000																																																																																																																
macro avg	0.97	0.87	0.91	3000																																																																																																																
weighted avg	0.97	0.97	0.97	3000																																																																																																																
Confusion Matrix on Train Set																																																																																																																				
<table> <thead> <tr> <th></th><th>predicted_A</th><th>predicted_B</th><th>predicted_C</th><th>predicted_D</th><th>predicted_E</th><th>predicted_F</th><th>predicted_G</th></tr> </thead> <tbody> <tr><td>A</td><td>2126</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>B</td><td>0</td><td>2126</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>C</td><td>0</td><td>0</td><td>2126</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>D</td><td>0</td><td>0</td><td>0</td><td>2126</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>E</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2126</td><td>0</td><td>0</td></tr> <tr><td>F</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2126</td><td>0</td></tr> <tr><td>G</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2126</td></tr> </tbody> </table>								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																																														
	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																																																																																																													
Confusion Matrix on Test Set																																																																																																																				
<table> <thead> <tr> <th></th><th>predicted_A</th><th>predicted_B</th><th>predicted_C</th><th>predicted_D</th><th>predicted_E</th><th>predicted_F</th><th>predicted_G</th></tr> </thead> <tbody> <tr><td>A</td><td>718</td><td>20</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>B</td><td>0</td><td>907</td><td>4</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>C</td><td>0</td><td>2</td><td>792</td><td>2</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>D</td><td>0</td><td>0</td><td>30</td><td>404</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>E</td><td>0</td><td>0</td><td>4</td><td>14</td><td>82</td><td>0</td><td>0</td></tr> <tr><td>F</td><td>0</td><td>0</td><td>0</td><td>1</td><td>1</td><td>15</td><td>0</td></tr> <tr><td>G</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>1</td><td>2</td></tr> </tbody> </table>								predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G	A	718	20	0	0	0	0	0	B	0	907	4	0	0	0	0	C	0	2	792	2	0	0	0	D	0	0	30	404	0	0	0	E	0	0	4	14	82	0	0	F	0	0	0	1	1	15	0	G	0	1	0	0	0	1	2																																														
	predicted_A	predicted_B	predicted_C	predicted_D	predicted_E	predicted_F	predicted_G																																																																																																													
A	718	20	0	0	0	0	0																																																																																																													
B	0	907	4	0	0	0	0																																																																																																													
C	0	2	792	2	0	0	0																																																																																																													
D	0	0	30	404	0	0	0																																																																																																													
E	0	0	4	14	82	0	0																																																																																																													
F	0	0	0	1	1	15	0																																																																																																													
G	0	1	0	0	0	1	2																																																																																																													

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:



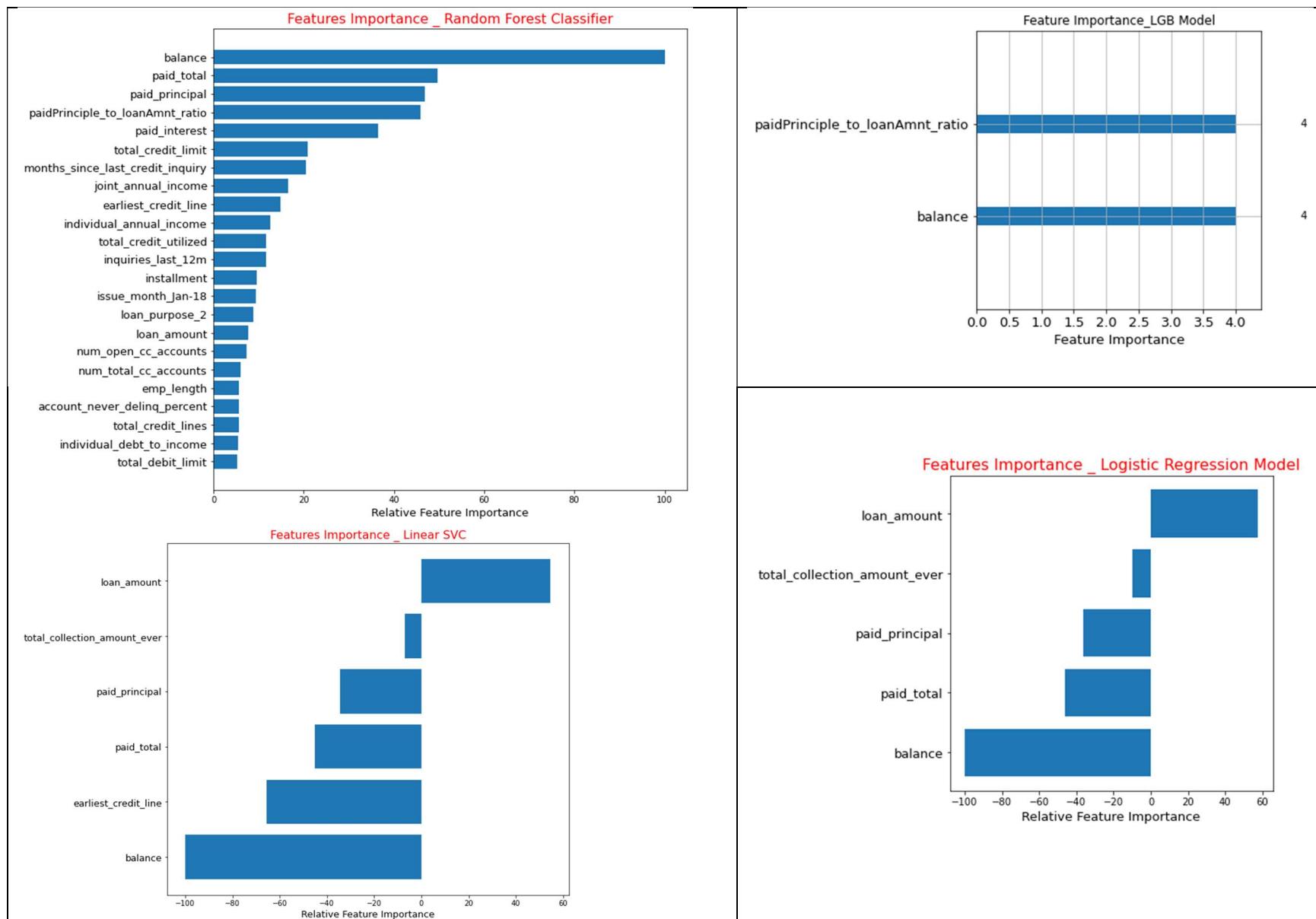
Three most important features which contribute in grade classification were:

- Interest Rate
- Installment
- term 36-month

Default Loan Detection

The hyper parameters tunned for Random Forest Classifier, Logistic Regression, and Light Gradient Boosting model through Grid Search and Bayesian Optimization. Except Ridge Classifier the other models perform well in detecting default loans.

Random Forest Classifier					Light Gradient Boosting Model					Logistic Regression				
Random Forest Clf on Train					lightgbm on Train					Logistic Regression on Train				
precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support	
0	1.00	1.00	1.00	6995	0	1.00	1.00	6995		0	1.00	1.00	1.00	6995
1	1.00	1.00	1.00	6505	1	1.00	1.00	6505		1	1.00	1.00	1.00	6505
accuracy			1.00	13500	accuracy			13500		accuracy			1.00	13500
macro avg	1.00	1.00	1.00	13500	macro avg	1.00	1.00	13500		macro avg	1.00	1.00	1.00	13500
weighted avg	1.00	1.00	1.00	13500	weighted avg	1.00	1.00	13500		weighted avg	1.00	1.00	1.00	13500
Random Forest Clf on Test					lightgbm on Test					Logistic Regression				
precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support	
0	1.00	1.00	1.00	2998	0	1.00	1.00	2998		0	1.00	1.00	1.00	2998
1	1.00	1.00	1.00	2	1	1.00	1.00	2		1	1.00	1.00	1.00	2
accuracy			1.00	3000	accuracy			3000		accuracy			1.00	3000
macro avg	1.00	1.00	1.00	3000	macro avg	1.00	1.00	3000		macro avg	1.00	1.00	1.00	3000
weighted avg	1.00	1.00	1.00	3000	weighted avg	1.00	1.00	3000		weighted avg	1.00	1.00	1.00	3000
Linear SVC					Ridge Classifier									
LinearSVC on Train					Ridge Classifier on Train									
precision	recall	f1-score	support		precision	recall	f1-score	support						
0	1.00	1.00	1.00	6995	0	1.00	1.00	6995						
1	1.00	1.00	1.00	6505	1	1.00	1.00	6505						
accuracy			1.00	13500	accuracy			13500						
macro avg	1.00	1.00	1.00	13500	macro avg	1.00	1.00	13500						
weighted avg	1.00	1.00	1.00	13500	weighted avg	1.00	1.00	13500						
LinearSVC on Test					Ridge Classifier on Test									
precision	recall	f1-score	support		precision	recall	f1-score	support						
0	1.00	1.00	1.00	2998	0	1.00	1.00	2998						
1	1.00	1.00	1.00	2	1	0.29	1.00	0.44	2					
accuracy			1.00	3000	accuracy			3000						
macro avg	1.00	1.00	1.00	3000	macro avg	0.64	1.00	0.72	3000					
weighted avg	1.00	1.00	1.00	3000	weighted avg	1.00	1.00	1.00	3000					



Based on the four classifiers with well performance, the most important features in detecting default loans were:

- Balance
- Paid total
- Paid principle to loan amount ratio
- Paid interest

