



Lending Club Loan Status

2019 Data Analysis Report for Fontys University

Version number: 1

First published: May 2016

Prepared by: Borislav Pavlov

CONTENTS

| | | |
|----|-----------------------------|-----|
| 01 | FOREWORD | 4 |
| 02 | INTRODUCTION..... | 5 |
| 03 | DATA AND METHODOLOGY | 7 |
| | 3.1 Lending Club Data | 8 |
| | 3.2 Methods..... | 9 |
| | 3.3 Data analysis..... | 11 |
| | 3. 4 Results..... | 14 |
| 04 | CONCLUSION | 18 |
| 05 | APPENDICES | 19 |
| | Business Proposal..... | 200 |

01 FOREWORD

Lending Club is a US peer-to-peer lending company, headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. Lending Club is the world's largest peer-to-peer lending platform. The company claims that \$ 15.98 billion in loans had been originated through its platform up to December 31, 2015.

Lending Club enables borrowers to create unsecured personal loans between \$ 1,000 and \$ 40,000. The standard loan period is three years. Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee.

02 INTRODUCTION

If you're following the news in the world of online marketplace lending then you maybe heard about the internal scandal in Lending Club's company which resulted the resignation of its founder, chairman, and chief executive, Renaud Laplanche.

Let me tell you what happened in a few words. On Monday, after an internal review, Lending Club announced that \$22 million in subprime loans that had been sold in March and April of this year to a single investor, went against the investor's expressed terms.

Moreover, specific staff members, including Laplanche, were aware that the loans did not meet the investor's criteria, and during the review process there was less than full reveal by staff including Laplanche, which the board considered unacceptable.

The review also found out that, in an unrelated matter, Laplanche failed to include personal interests he held in a third-party fund while Lending Club was considering an investment in the same fund.

Lending Club led the review under the supervision of a sub-committee of its board of directors with the assistance of independent outside counsel and other advisors.

Laplanche and three other senior staff were fired.

The company's stock price went down dramatically, and investors Jeffries LLC and Goldman Sachs stopped their purchase of loans from Lending Club as they review the events.

Jeffries, for instance, had been looking to package roughly \$150 million of Lending Club loans into bonds for investor sales in early May.

This situation is a sign that frauds like this must be prevented and this is the purpose of this report too.

Could the disaster from 2016 for the company be prevented?

Is there any relation with the customer's data and the target question if they are going to return their loan?

The goal of this project will be to answer these questions.

If you want to read more about what happened with the company 2016 check the web pages.

<https://www.nytimes.com/2018/09/28/technology/lendingclub-renaud-laplanche-fraud.html>

<https://www.marketwatch.com/story/goldman-jefferies-stopped-buying-lendingclub-loans-2016-05-10>



03 DATA AND METHODOLOGY

3.1 Lending Club Data

There are available files containing complete loan data for all loans issued through the 2007-2019, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others. The file is a matrix of about 890 thousand observations and 75 variables.

Our main focus will be on the years from 2007 to 2018. Furthermore, a data dictionary is provided in a separate file.

3.2 Methods

All of the data is publicly available at [LendingClub.com](https://lendingclub.com).

We will use the data from 2007-2018 year which can be downloaded from [here](#).

The analysis was made in Python programming language supported by Anaconda environment and more precisely Jupyter Notebook, which is attached at bottom of the document.

First of all, all of the data had to be cleaned. Once we had the data cleaned and ready we started exploring it by plotting different distributions of the features.

After the whole exploratory data analysis, we were ready to choose which features to use for our prediction and they were the following:

- **loan_amount:** The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
- **Interest_rate:** Interest Rate on the loan.
- **annual_income:** The self-reported annual income provided by the borrower during registration.
- **delinq_2yrs:** The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
- **dti:** A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
- **emp_length:** Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
- **tot_cur_bal:** Total current balance of all accounts.
- **purpose:** A category provided by the borrower for the loan request.
- **home_ownership:** The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.

- **recoveries:** Post charge of gross recovery.
- **total_rec_prncp:** Principal received to date.
- **out_prncp:** Remaining outstanding principal for total amount funded.
- **last_pymnt_d:** Last month payment was received.

For making our predictive model a reality, we decided to use the following algorithms:

- Decision tree classifier,
- Random Forest (tree based model)
- Neural Network algorithms were used.

3.3 Data analysis

Through our Exploratory Data Analysis, we tried to find if there is any kind of relation between any of the features in the dataset. For the purpose, we used different libraries which helped us to better visualize the features that we are exploring and see if there were any trends and correlations.

The Figure below are the conclusions that we found by distributing some of the features which we think can give us some insights and help us with the prediction later.

LOANS ISSUED PER YEAR

1

- Most of the **loans issued** were in the range of 10,000 to 20,000 USD.
- The **year of 2018** was the year were most loans were issued.
- Loans were issued in an **incremental manner**. (Possible due to a recovery in the U.S economy)

DISTRIBUTION OF THE STATES

2

- **California, Texas, New York and Florida** are the states in which the highest amount of loans were issued.
- Interesting enough, all four states have an approximate **interest rate of 13%** which is at the same level of the average interest rate for all states (13.24%)
- California, Texas and New York are **all above the average annual income** (with the exclusion of Florida), this might give possible indication why most loans are issued in these states.

DISTRIBUTION OF THE LOAN STATUS

3

The conclusions that we can get from this chart are that we have a lot of loans which are fully paid with fair amount of current loans. Other categories (including) default have a really low number. This means the data is imbalanced and we might need to do something about this later in the analysis.

The approach that we are going to use is to replace all kind of "Late" loans, "Charged Off", "Does not meet the credit policy. Status: Charged Off", "In Grace Period" and "Default" status with "Bad Loan" status and the rest with "Good Loan" status. In this way we have only 2 options for the loan status and we increase the balance of the data.

DISTRIBUTION BY ICOME

4

- Borrowers that made part of the **high income category** took higher loan amounts than people from **low** and **medium income categories**. Of course, people with higher annual incomes are more likely to pay loans with a higher amount. (First row to the left of the subplots)
- Loans that were borrowed by the **Low income category** had a slightly higher change of becoming a bad loan. (First row to the right of the subplots)
- Borrowers with **High** and **Medium** annual incomes had a longer employment length than people with lower incomes (Second row to the left of the subplots)

DISTRIBUTION OF THE INTEREST RATE

5

- Loans that have a **high interest rate** (above 13.23%) are more likely to become a **bad loan**.
- Loans that have a longer **maturity date (60 months)** are more likely to be a bad loan.

DISTRIBUTION OF THE HOME OWNERSHIP

6

- The types of bad loans in the last year are having a tendency to decline, except for not meeting credit policy.
- Mortgage was the variable from the home ownership column that used the highest amount borrowed within loans that were considered to be bad.
- There is a slight **increase** on people who have mortgages that are applying for a loan.

DISTRIBUTION BY THE PURPOSE OF THE LOAN

7

- **Bad Loans Count:** People that apply for educational and small business purposed tend to have a higher risk of being a bad loan. (% wise)
- **Most frequent Purpose:** The reason that clients applied the most for a loan was to consolidate debt.
- **Less frequent purpose:** Clients applied less for educational purposes for all three income categories.
- **Interest Rates:** In all reasons for application except (medical, small business and credit card), the low income category has a higher interest rate. Something that could possibly explain this is the amount of capital that is needed from other income categories that might explain why the low income categories interest rate for these purposes are lower.

DISTRIBUTION BY THE 20 MOST COMMON EMPLOYEE TITLES

8

We can see that Director, Engineer, President, Vice President are the category's with highest incidence in Grade A.

Analyzing this we can get some insights about the profile of grades and professionals.

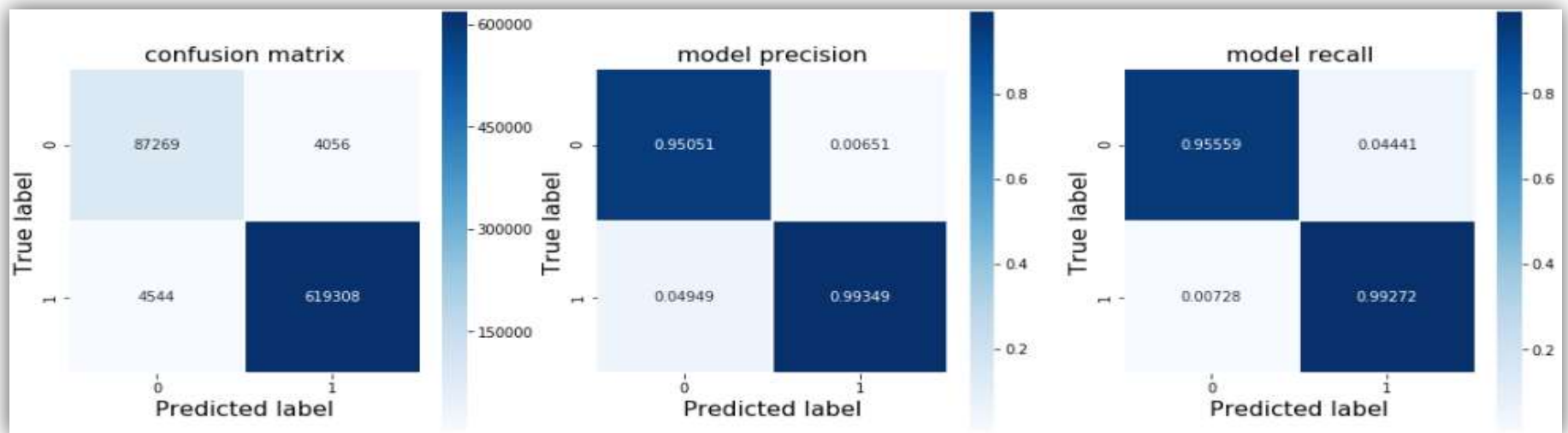
Fig.1: Features Analysis

3. 4 Results

After we applied the algorithms that we said we are going to us in the “Methods” section, we displayed in a graph and compared their results with each other.

The results are the following:

DECISION TREE CLASSIFIER RESULTS



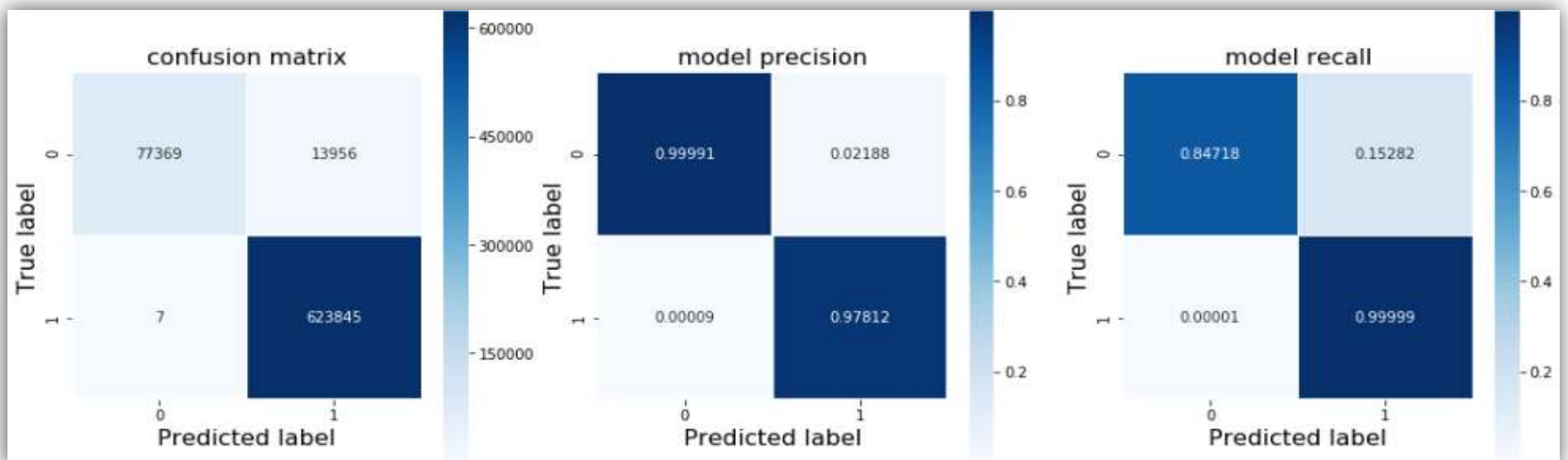
As you can see, this model had predicted right 87269 “Bad Loans” and 619308 “Good Loans”. However, it had wrong 4056 “Bad Loans” and 4544 “Good Loans”. The total model precision for the “Bad Loans” is 0.95 % and for the “Good Loans” is 0.99 %. The overall accuracy score for this model is 0.98% which is really good.

NEURAL NETWORKS RESULTS

For our neural networks model we used TensorFlow and we tried two not so different approaches:

- Using one of the build in TensorFlow estimators
- Using a Keras Model

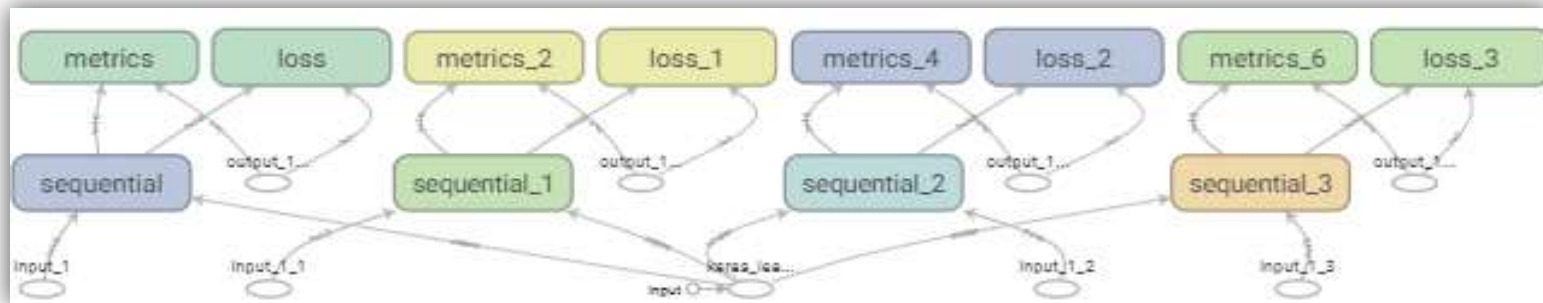
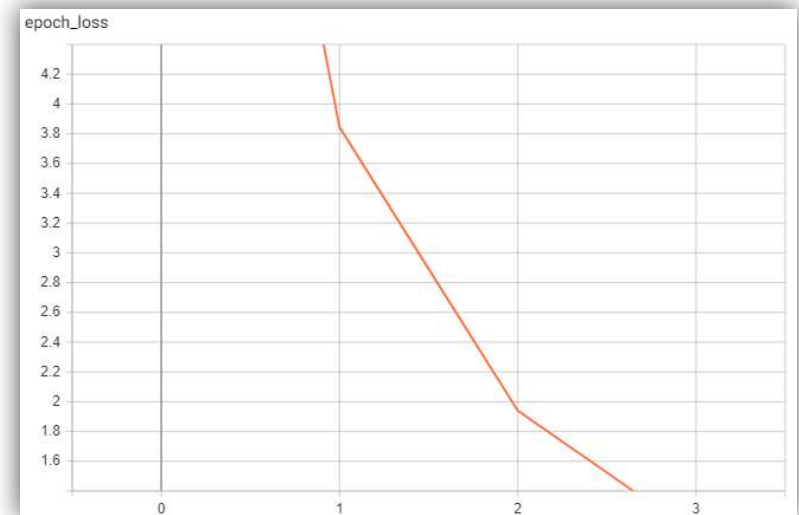
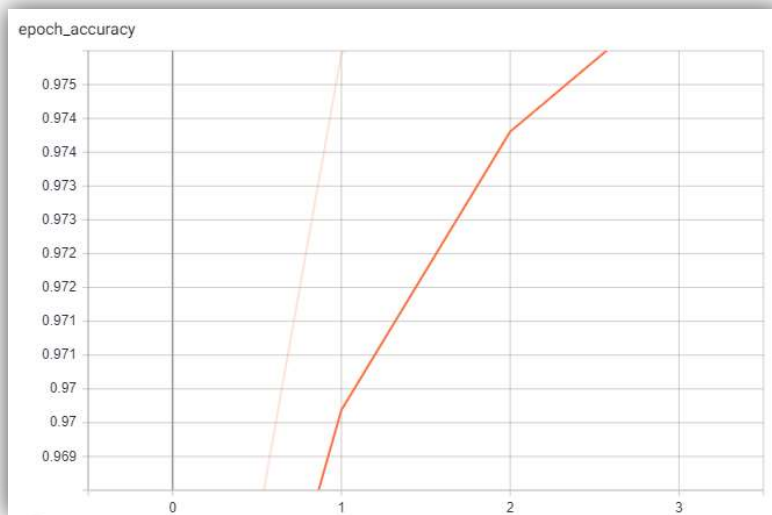
For our first approach, since we are trying to classify if the loan is going to be “Bad Loan” or “Good Loan”, we used a DNNClassifier estimator. Defining one hidden layer with 1500 neurons, we trained the model and the results that we got are the following.



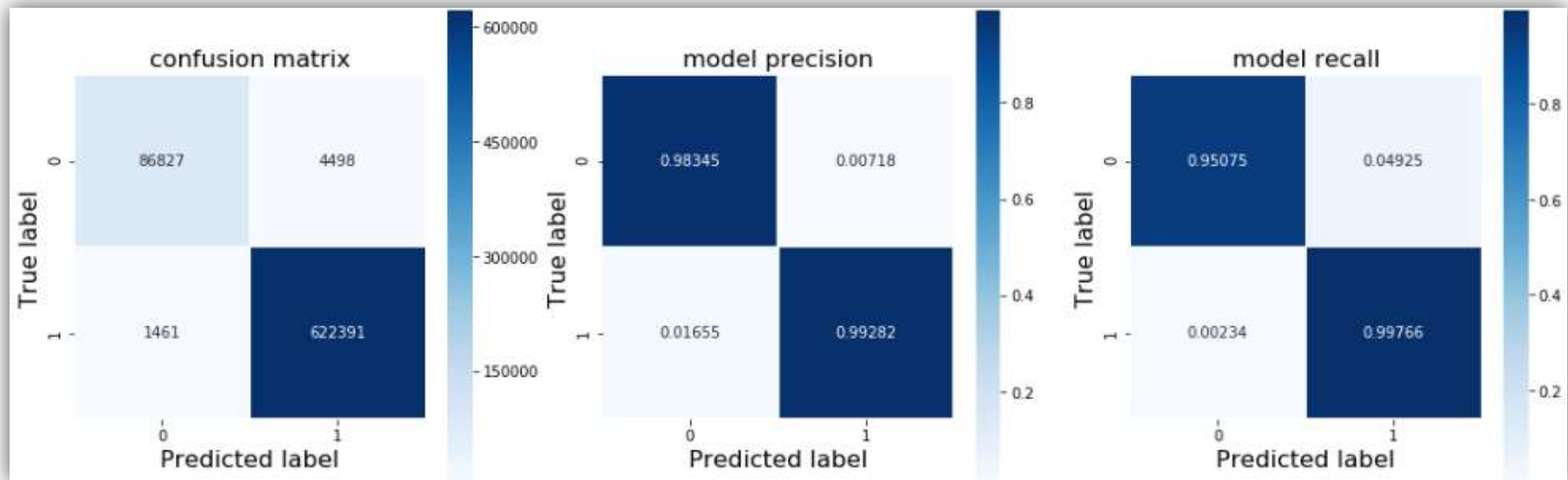
As you can see we have a 0.999 % accuracy in predicting whether the loan is going to be “Bad” and 0.97% whether it is going to be “Good”. The overall accuracy score for this model is again 0.98% which also a very good model.

The second approach, with defining a keras model, we have got an overall accuracy of 0.979% in 4 epochs.

We used the tensor board so we can see a visualization of the epoch accuracy, epoch loss and the main graph of the model.



RANDOM FOREST CLASSIFIER RESULTS



Our final model had overall accuracy score of 0.99% which make it the best. It is noticeable that only 1461 samples were got wrong for predicting if they are “Good Loans” which is 4 times better than the Decision Tree Classifier.

04 CONCLUSION

We can say that all of the models are very good with accuracy score 0.98% for the Decision Tree Classifier, Neural Networks algorithms and 0.99% for the Random Forest Classifier.

However, if we have to choose one of the models it will not be that easy. It really depends on which metrics you want to optimize. For instance, we can see that the precision of predicting “Bad Loans” using the Neural Networks is the highest – 0.99% but the recall is 0.84% and for the Random Forest Classifier the precision for predicting “Bad Loans” is 0.98% but the recall is 0.95% and that is 0.11% more than the Neural Networks.

In conclusion, I can say that we successfully build a machine learning predictive model with almost perfect accuracy which predicts whether a person will return his/her loan, which can be used further by Lending Club for their analysis.

05 APPENDICES

- BUSINESS PROPOSAL
- JUPYTER NOTEBOOK

Lending Club Loan Status

2019 Business Proposal

Version number: 3

Prepared by: Borislav Pavlov



CONTENTS

| | | |
|----|---------------------------------|-----|
| 01 | COMPANY INFORMATION | 222 |
| 02 | ABOUT THE DATASET | 244 |
| 03 | EXPLORATORY DATA ANALYSIS | 255 |
| 04 | RISK ANALYTICS | 40 |
| 05 | MODELING | 43 |
| | REFERENCES | 45 |

01 COMPANY INFORMATION

Lending Club is a US peer-to-peer lending company, headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. Lending Club is the world's largest peer-to-peer lending platform. The company claims that \$ 15.98 billion in loans had been originated through its platform up to December 31, 2015.

Lending Club enables borrowers to create unsecured personal loans between \$ 1,000 and \$ 40,000. The standard loan period is three years. Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee.

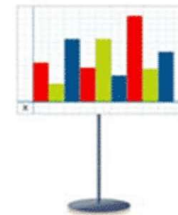
HOW LENDING CLUB WORKS



Borrowers apply for loans.
Investors open an account.



Borrowers get funded.
Investors build a portfolio.



Borrowers repay automatically.
Investors earn & reinvest.

02 ABOUT THE DATASET

There are available files containing complete loan data for all loans issued through the 2007-2019, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others. The file is a matrix of about 890 thousand observations and 75 variables.

Our main focus will be on the years from 2007 to 2018, because most of the data for 2019 year is still in the process and most of the loans are just issued.

Furthermore, a data dictionary is provided in a separate file.

03 EXPLORATORY DATA ANALYSIS

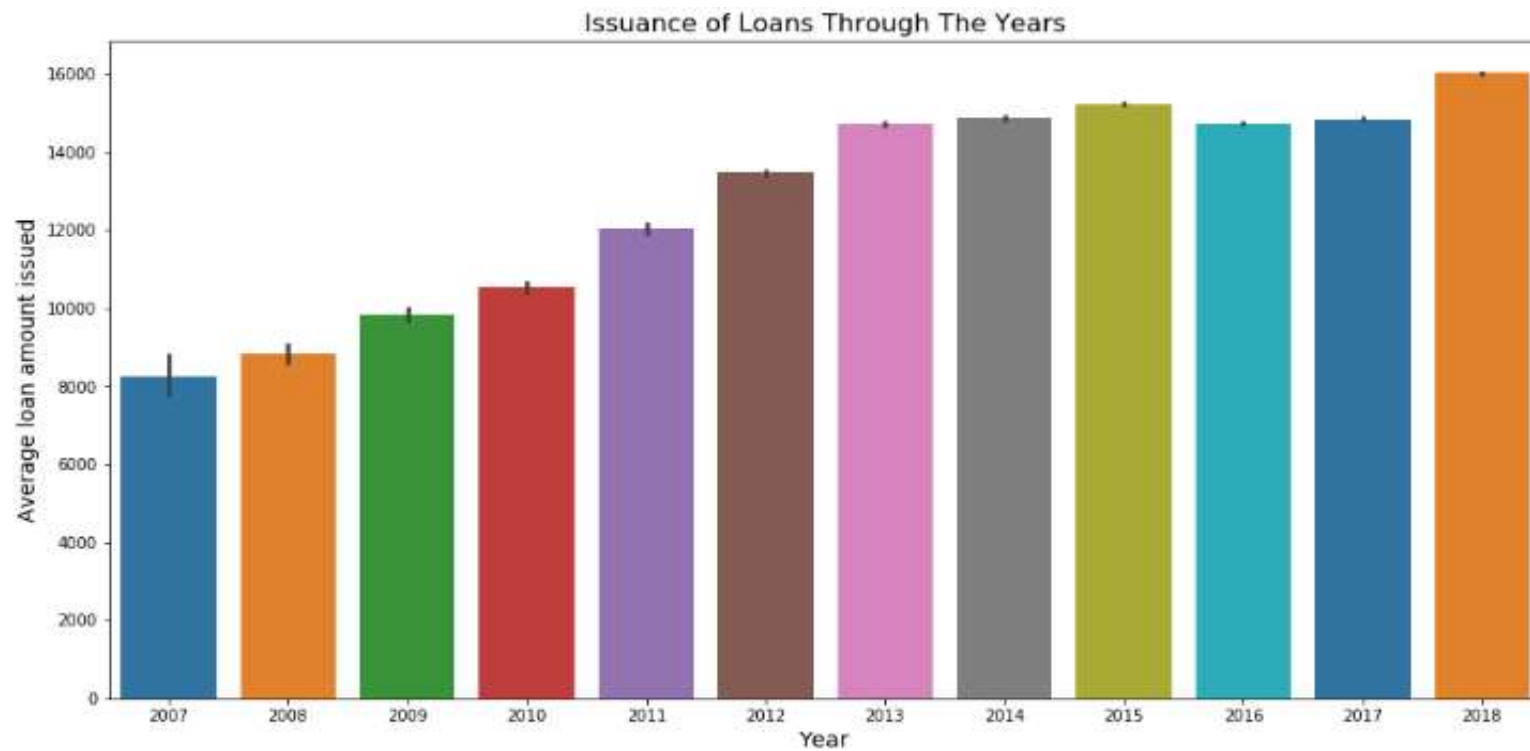
Our exploring of the data will begin by looking of the number and percentage of missing values.

| | | |
|--|---------|------------|
| id | 2280868 | 100.000000 |
| member_id | 2280868 | 100.000000 |
| url | 2280868 | 100.000000 |
| desc | 2134801 | 94.423482 |
| mths_since_last_delinq | 1158502 | 51.248003 |
| mths_since_last_record | 1901512 | 84.112837 |
| next_pymnt_d | 1303807 | 57.884881 |
| mths_since_last_major_derog | 1679893 | 74.309585 |
| annual_inc_joint | 2139958 | 94.880428 |
| dti_joint | 2139982 | 94.880805 |
| verification_status_joint | 2144938 | 94.880717 |
| mths_since_recent_bc_dlq | 1740967 | 77.011175 |
| mths_since_recent_revol_delinq | 1520309 | 87.250432 |
| revol_bal_joint | 2152848 | 95.221788 |
| sec_app_earliest_cr_line | 2152847 | 95.221722 |
| sec_app_inq_last_6mths | 2152847 | 95.221722 |
| sec_app_mort_acc | 2152847 | 95.221722 |
| sec_app_open_acc | 2152847 | 95.221722 |
| sec_app_revol_util | 2154484 | 95.302981 |
| sec_app_open_act_il | 2152847 | 95.221722 |
| sec_app_num_rev_accts | 2152847 | 95.221722 |
| sec_app_chargeoff_within_12_mths | 2152847 | 95.221722 |
| sec_app_collections_12_mths_ex_med | 2152847 | 95.221722 |
| sec_app_mths_since_last_major_derog | 2224726 | 98.410116 |
| hardship_type | 2250055 | 99.530537 |
| hardship_reason | 2250055 | 99.530537 |
| hardship_status | 2250055 | 99.530537 |
| deferral_term | 2250055 | 99.530537 |
| hardship_amount | 2250055 | 99.530537 |
| hardship_start_date | 2250055 | 99.530537 |
| hardship_end_date | 2250055 | 99.530537 |
| payment_plan_start_date | 2250055 | 99.530537 |
| hardship_length | 2250055 | 99.530537 |
| hardship_dpd | 2250055 | 99.530537 |
| hardship_loan_status | 2250055 | 99.530537 |
| orig_projected_additional_accrued_interest | 2252242 | 99.627278 |
| hardship_payoff_balance_amount | 2250055 | 99.530537 |

As you can see, there are a lot of columns which have huge chunk of data missing. These columns are not necessary for our analysis. Next step is to drop any columns where 50% or more data is missing. This will help us clean the Dataset a little bit.

We continue by displaying all of the loans issued per each year we understood from the chart that:

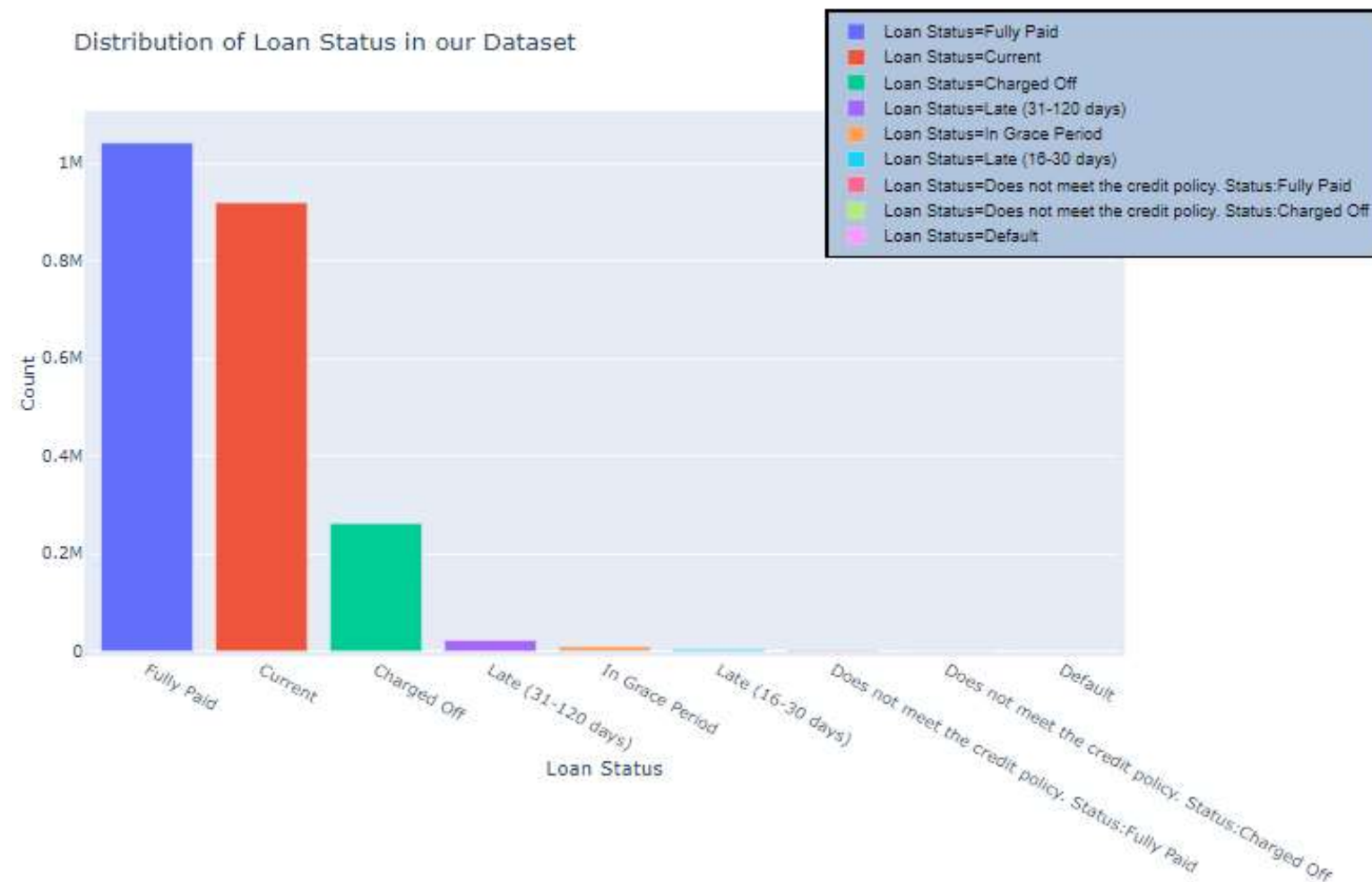
- Most of the **loans issued** were in the range of 10,000 to 20,000 USD.
- The **year of 2018** was the year were most loans were issued.
- Loans were issued in an **incremental manner**. (Possible due to a recovery in the U.S economy)



By using histogram, we can see the distribution of the loan status column.

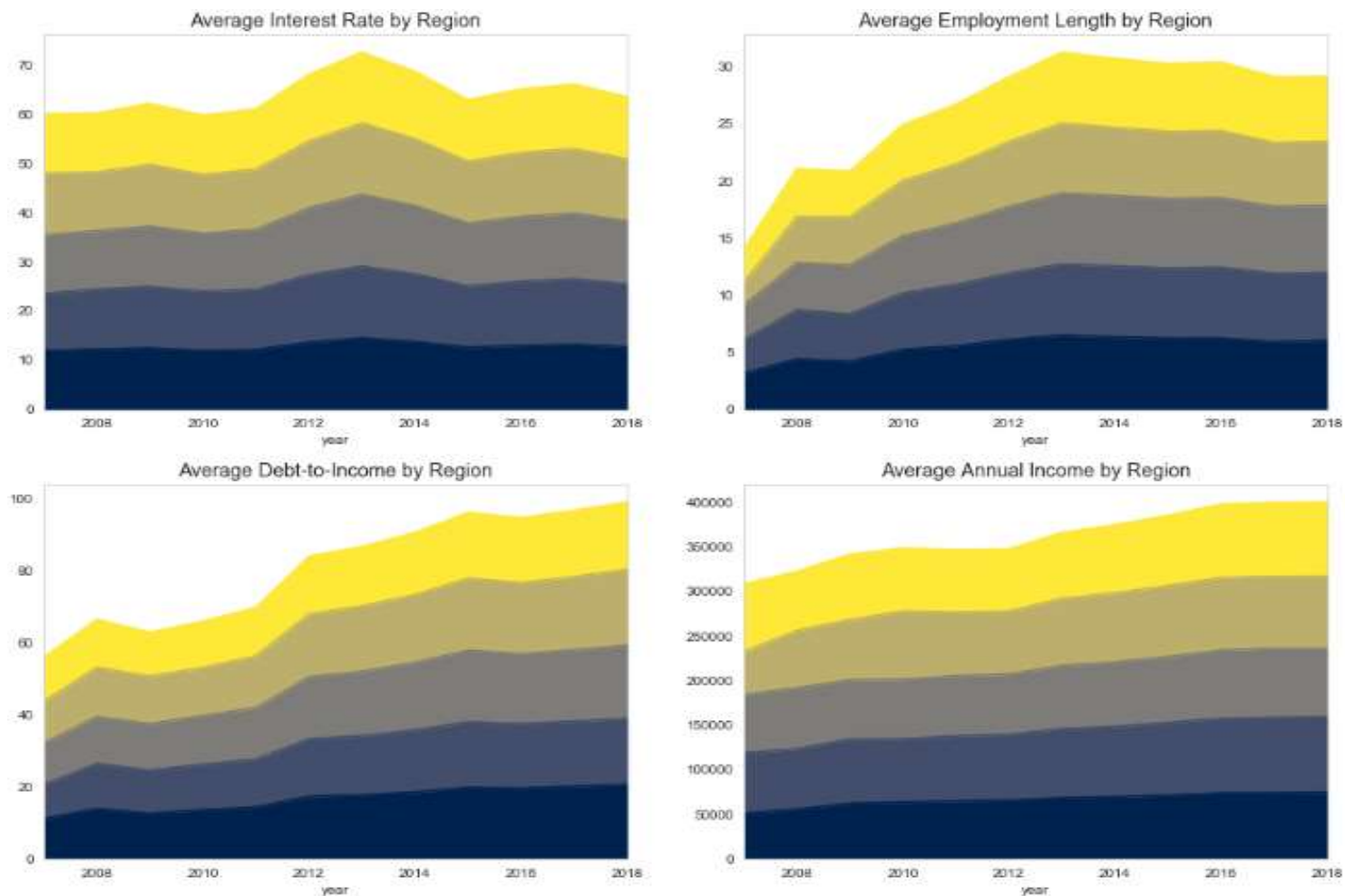
The conclusions that we can get from this chart are that we have a lot of loans which are fully paid with fair amount of current loans. Other categories (including) default have a really low number. This means the data is imbalanced and we might need to do something about this later in the analysis.

The approach that we are going to use is to replace all kind of “Late” loans,” Charged Off”, “Does not meet the credit policy. Status: Charged Off”, “In Grace Period” and “Default” status with “Bad Loan” status and the rest with “Good Loan” status. In this way we have only 2 options for the loan status and we increase the balance of the data.

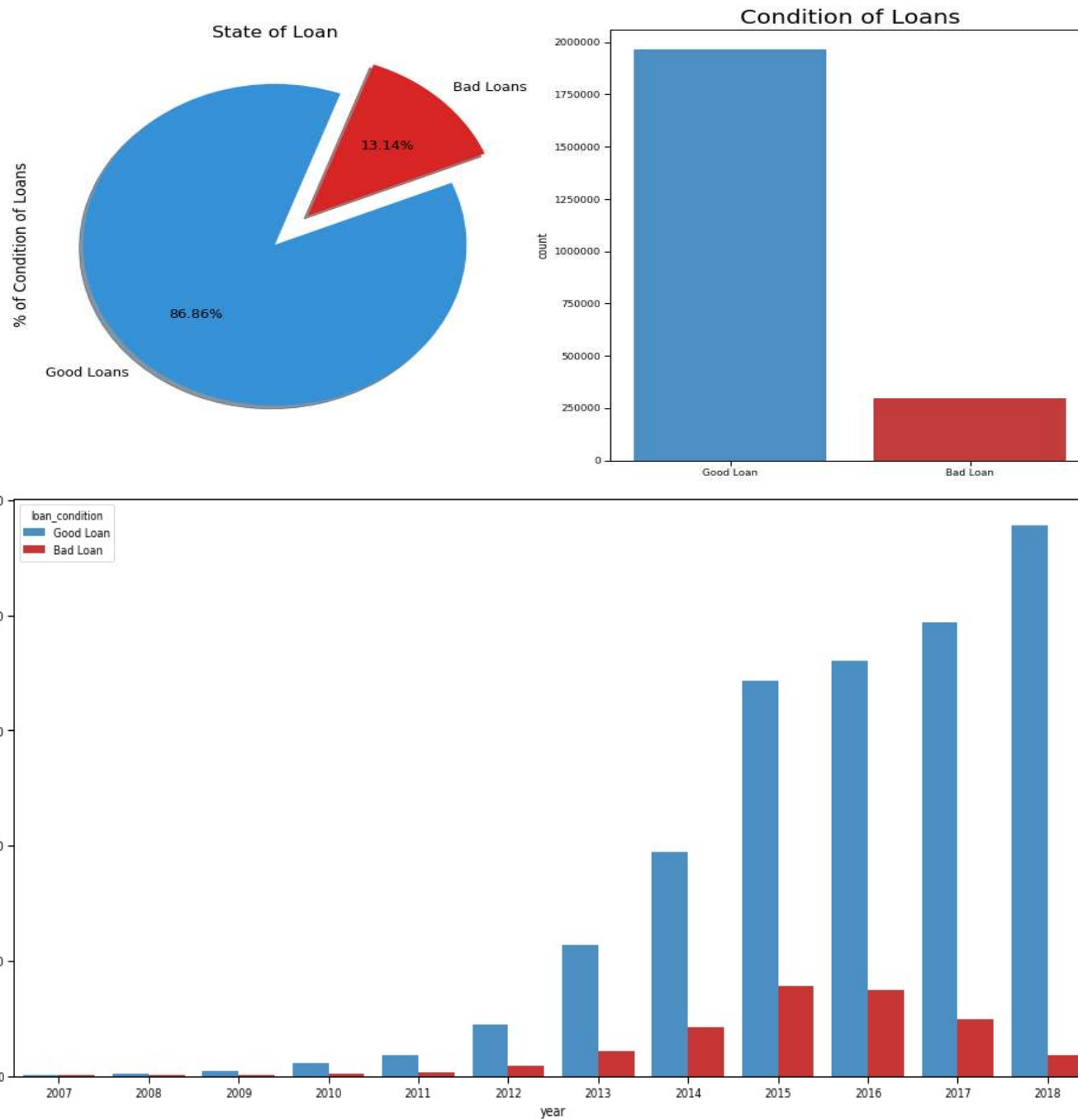


The summary from the charts below is as follow:

- Currently, **bad loans** consist 13.14% of total loans but remember that we still have **current loans** which have the risk of becoming bad loans. (So this percentage is subjected to possible changes.)
- The **South West** and **West** regions have experienced a slight increase in the "median income" in the past years.
- **Average interest** rates have declined since 2012 but this might explain the **increase in the volume** of loans.
- **Employment Length** tends to be greater in the regions of the **South West** and **West**
- Clients located in the regions of **Northeast** and **Midwest** have not experienced a drastic increase in debt-to-income(dti) as compared to the other regions.



Information on Loan Conditions

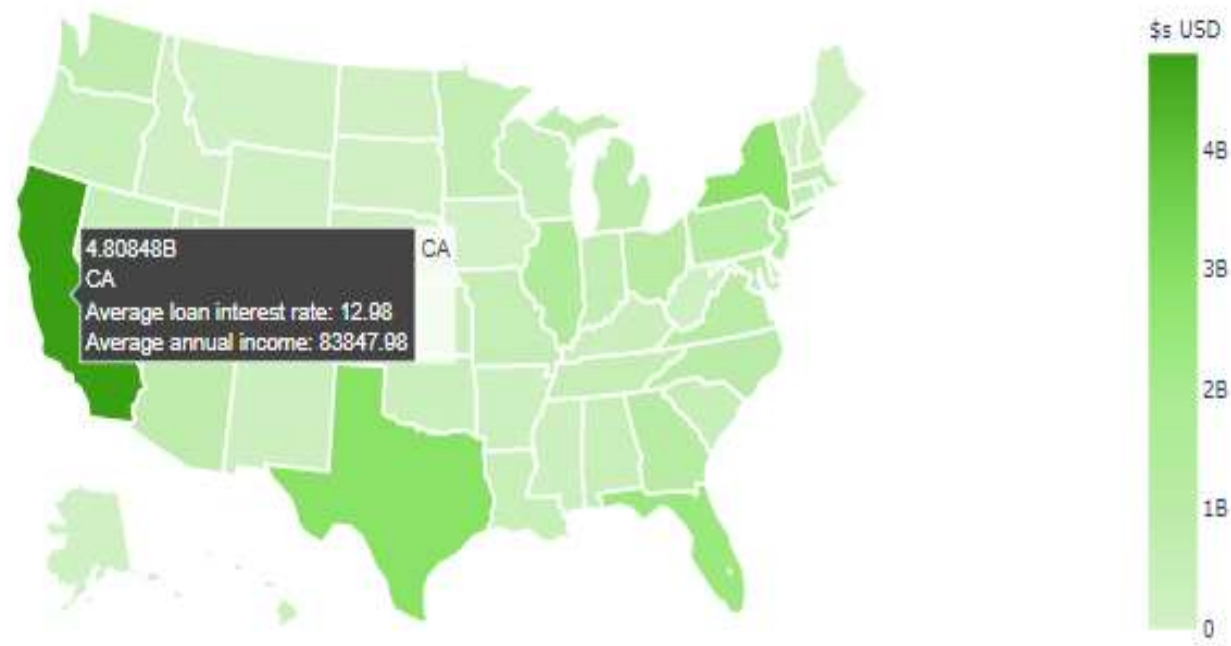


The next chart is distribution of the states by focusing on **three key metrics**: Loans issued by state (Total Sum), Average interest rates charged to customers and average annual income of all customers.

The conclusions that we can get from here are that:

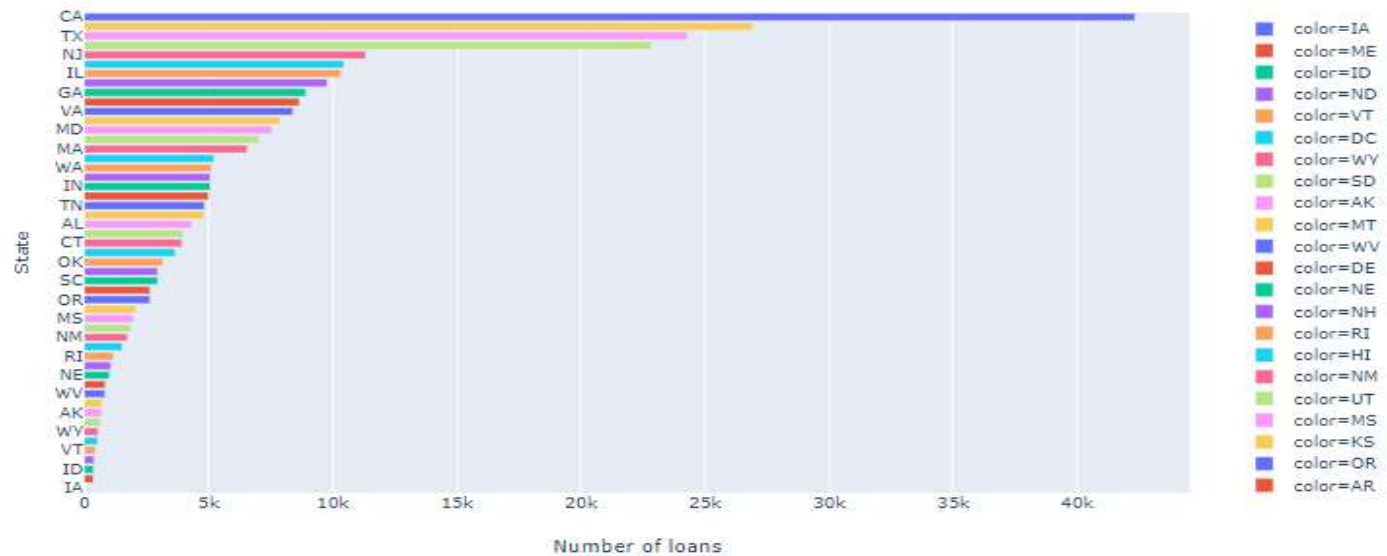
- **California, Texas, New York and Florida** are the states in which the highest amount of loans were issued.
- Interesting enough, all four states have an approximate **interest rate of 13%** which is at the same level of the average interest rate for all states (13.24%)
- California, Texas and New York are **all above the average annual income** (with the exclusion of Florida), this might give possible indication why most loans are issued in these states.

Lending Clubs Issued Loans (A Perspective for the Business Operations)

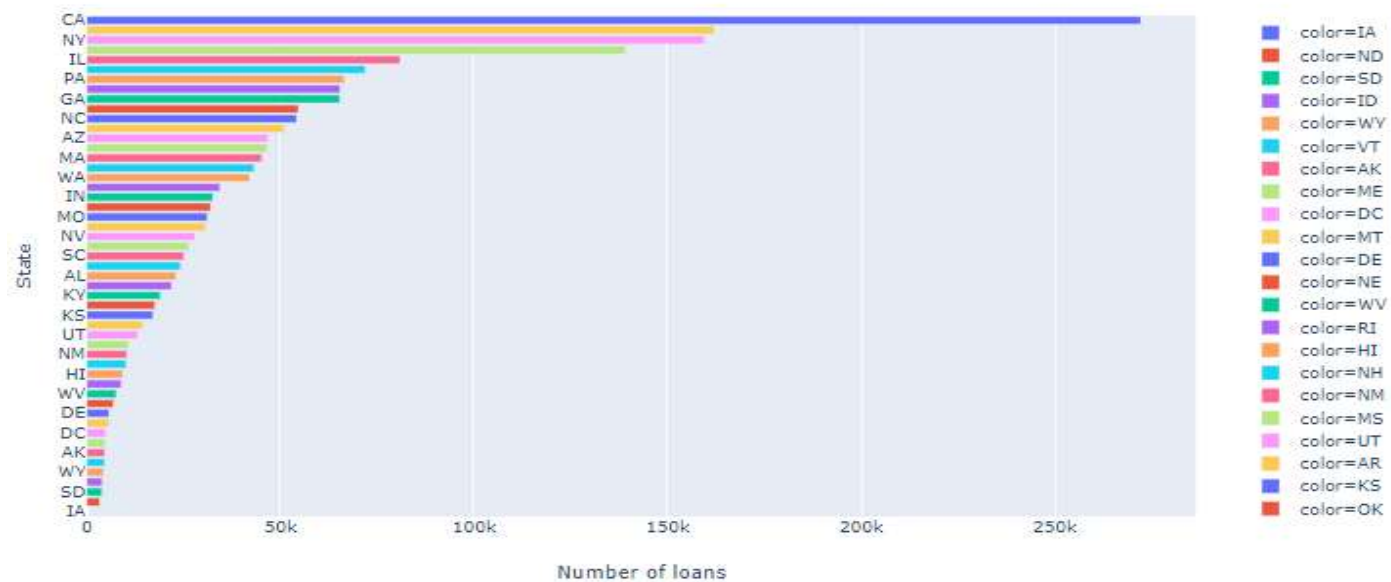


Using the states will not contribute much in our prediction, because as you can see from the graphs below, most of the states are symmetrical in the comparison between the good and the bad loans.

Number of bad loans per state



Number of good loans per state



By distributing the income category, this what we can get out of the charts:

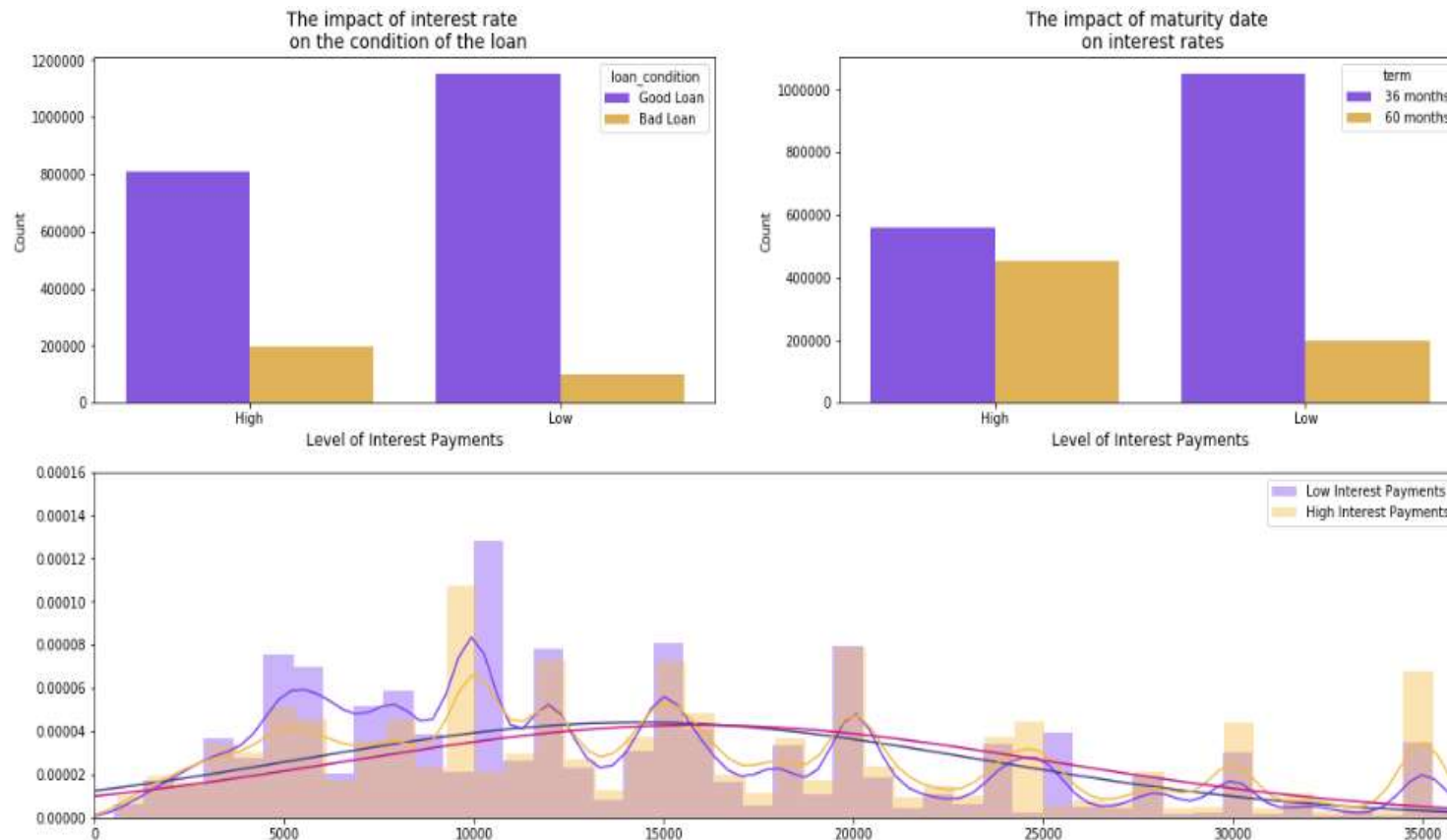
- Borrowers that made part of the **high income category** took higher loan amounts than people from **low** and **medium income categories**. Of course, people with higher annual incomes are more likely to pay loans with a higher amount. (First row to the left of the subplots)
- Loans that were borrowed by the **Low income category** had a slightly higher change of becoming a bad loan. (First row to the right of the subplots)
- Borrowers with **High** and **Medium** annual incomes had a longer employment length than people with lower incomes (Second row to the left of the subplots)
- Borrowers with a lower income had on average **higher interest rates** while people with a higher annual income had **lower interest rates** on their loans. (Second row to the right of the subplots)

Income Distribution



By distribution the interest rate combined with high and low level payments with the loan status and the impact of maturity date we can say that:

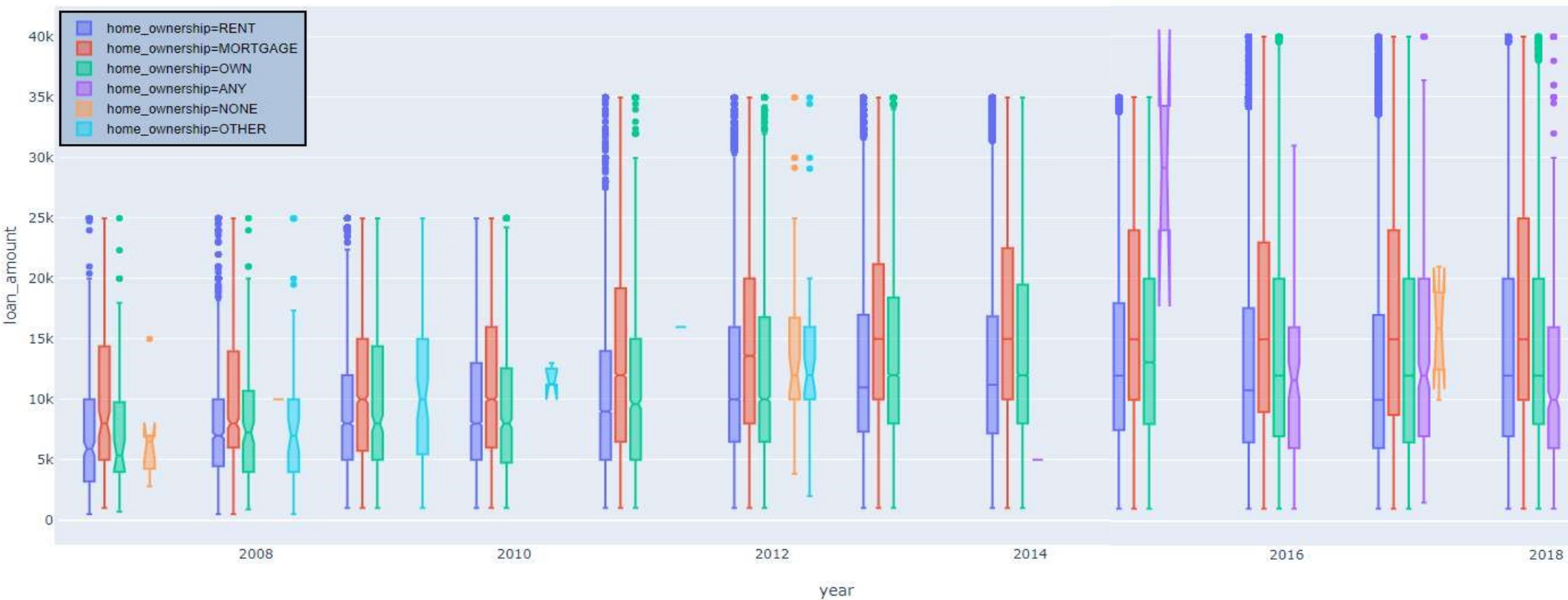
- Loans that have a **high interest rate** (above 13.23%) are more likely to become a **bad loan**.
- Loans that have a longer **maturity date (60 months)** are more likely to be a bad loan.



The main aim in here is to find the main factors that causes for a loan to be considered a "Bad Loan". Features like low credit grade or a high debt to income could be possible contributors in determining whether a loan is at a high risk of being bad.

From the charts below, we can see that:

- The types of bad loans in the last year are having a tendency to decline, except for not meeting credit policy.
- Mortgage was the variable from the home ownership column that used the highest amount borrowed within loans that were considered to be bad.
- There is a slight **increase** on people who have mortgages that are applying for a loan.
- People who have a mortgage (depending on other factors as well within the mortgage) are more likely to ask for



Types of Bad Loans (Amount Borrowed Throughout the Years)

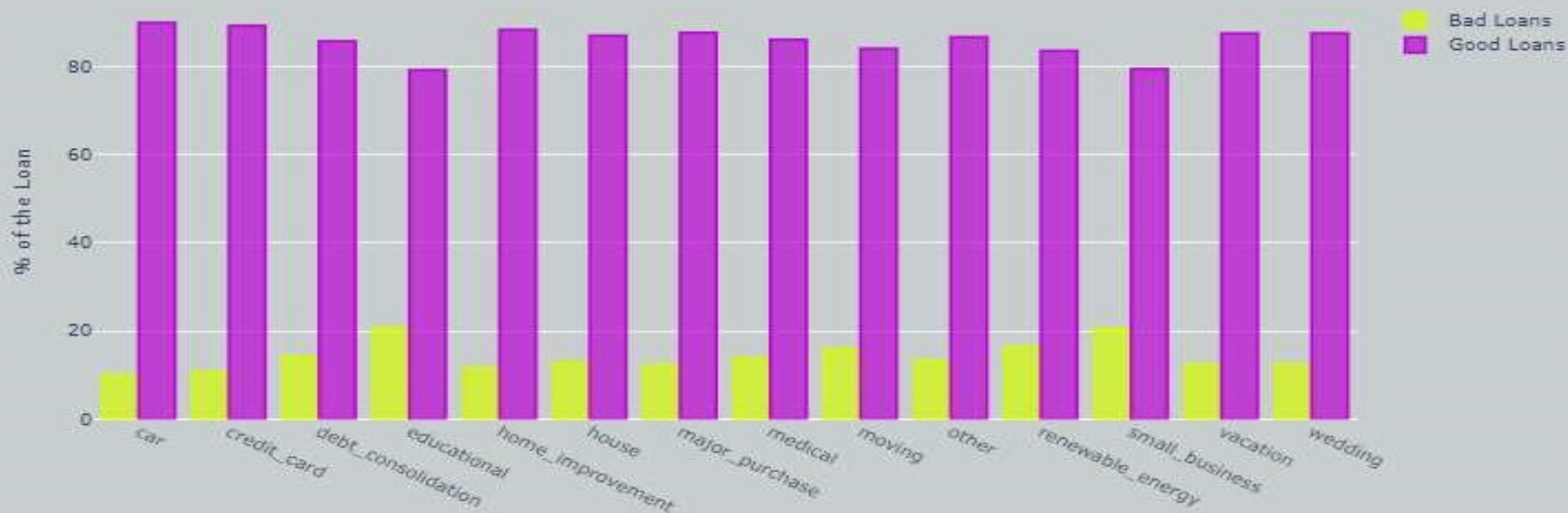


Our next step is to go into depth regarding the reasons for clients to apply for a loan. Our main aim is to see if there are purposes that contribute to a "higher" risk whether the loan will be repaid or not.

Summary:

- **Bad Loans Count:** People that apply for educational and small business purposed tend to have a higher risk of being a bad loan. (% wise)
- **Most frequent Purpose:** The reason that clients applied the most for a loan was to consolidate debt.
- **Less frequent purpose:** Clients applied less for educational purposes for all three income categories.
- **Interest Rates:** In all reasons for application except (medical, small business and credit card), the low income category has a higher interest rate. Something that could possibly explain this is the amount of capital that is needed from other income categories that might explain why the low income categories interest rate for these purposes are lower.
- **Bad/Good Ratio:** Except for educational purposes (we see a spike in high income this is due to the reasons that only two loans were issued and one was a bad loan which caused this ratio to spike to 50%), but we can see that in all other purposed the bad good ratio is lower the higher your income category.

Condition of Loan by Purpose



| | income_category | | purpose | interest_rate | total_loan_amount | good_loans_count | bad_loans_count | total_loans_issued | bad/good ratio (%) |
|----|-----------------|--|--------------------|---------------|-------------------|------------------|-----------------|--------------------|--------------------|
| 0 | High | | car | 10.7707 | 18261.8 | 371 | 27 | 398 | 6.78 |
| 6 | High | | major_purchase | 11.885 | 23518.8 | 1225 | 108 | 1333 | 8.1 |
| 1 | High | | credit_card | 10.483 | 26483.8 | 11502 | 814 | 12316 | 6.61 |
| 2 | High | | debt_consolidation | 11.9964 | 26586.3 | 23924 | 2392 | 26316 | 9.09 |
| 4 | High | | home_improvement | 11.4999 | 25283.3 | 6341 | 619 | 6960 | 8.89 |
| 5 | High | | house | 13.4026 | 27674.7 | 497 | 47 | 544 | 8.64 |
| 10 | High | | renewable_energy | 13.9004 | 23222.9 | 22 | 2 | 24 | 8.33 |
| 3 | High | | educational | 11.42 | 11000 | 1 | 1 | 2 | 50 |
| 8 | High | | moving | 13.4618 | 20310.1 | 297 | 26 | 323 | 8.05 |
| 9 | High | | other | 13.0708 | 22356.9 | 2824 | 293 | 3117 | 9.4 |
| 11 | High | | small_business | 14.047 | 26298.4 | 1086 | 202 | 1288 | 15.68 |
| 12 | High | | vacation | 12.1724 | 13601.4 | 161 | 19 | 180 | 10.56 |
| 13 | High | | wedding | 14.3942 | 20800.7 | 32 | 4 | 36 | 11.11 |
| 7 | High | | medical | 12.7023 | 18612 | 499 | 56 | 555 | 10.09 |
| 19 | Low | | house | 14.7226 | 13679 | 9208 | 1500 | 10708 | 14.01 |
| 14 | Low | | car | 12.3811 | 8621.24 | 18244 | 2147 | 20391 | 10.53 |
| 15 | Low | | credit_card | 11.8892 | 13575 | 368007 | 47160 | 413167 | 11.41 |
| 16 | Low | | debt_consolidation | 13.732 | 14366.5 | 880262 | 155070 | 1035332 | 14.98 |
| 17 | Low | | educational | 12.1935 | 6037.17 | 302 | 80 | 382 | 20.94 |
| 18 | Low | | home_improvement | 12.8863 | 12544.2 | 94178 | 13284 | 107462 | 12.36 |
| 20 | Low | | major_purchase | 12.9778 | 11151.7 | 34864 | 5104 | 39968 | 12.77 |
| 21 | Low | | medical | 13.7637 | 8542.1 | 19420 | 3275 | 22695 | 14.43 |
| 23 | Low | | other | 14.4278 | 9271.68 | 100342 | 16144 | 116486 | 13.86 |
| 24 | Low | | renewable_energy | 14.9076 | 9409.65 | 998 | 202 | 1200 | 16.83 |
| 25 | Low | | small_business | 15.5867 | 14115.7 | 13747 | 3808 | 17555 | 21.69 |
| 26 | Low | | vacation | 13.5781 | 5827.6 | 11522 | 1700 | 13222 | 12.86 |
| 27 | Low | | wedding | 14.0828 | 9727.41 | 1766 | 250 | 2016 | 12.4 |
| 22 | Low | | moving | 14.9422 | 7338.68 | 10738 | 2189 | 12927 | 16.93 |
| 28 | Medium | | car | 11.0979 | 13184.7 | 2991 | 233 | 3224 | 7.23 |
| 29 | Medium | | credit_card | 10.9895 | 21696.2 | 84226 | 7262 | 91488 | 7.94 |
| 30 | Medium | | debt_consolidation | 12.7093 | 22338.7 | 192183 | 24046 | 216229 | 11.12 |
| 40 | Medium | | vacation | 12.7658 | 9047.15 | 1911 | 212 | 2123 | 9.99 |
| 32 | Medium | | home_improvement | 12.0465 | 18943.5 | 32469 | 3566 | 36035 | 9.9 |
| 31 | Medium | | educational | 11.6883 | 11910 | 33 | 7 | 40 | 17.5 |
| 36 | Medium | | moving | 13.7073 | 12919.8 | 1917 | 236 | 2153 | 10.96 |
| 35 | Medium | | medical | 13.0493 | 13269 | 3762 | 476 | 4238 | 11.23 |
| 33 | Medium | | house | 13.372 | 20984.1 | 2602 | 282 | 2884 | 9.78 |
| 37 | Medium | | other | 13.2897 | 15716.8 | 17749 | 2084 | 19833 | 10.51 |
| 38 | Medium | | renewable_energy | 13.9434 | 16719.3 | 188 | 33 | 221 | 14.93 |
| 39 | Medium | | small_business | 14.5619 | 21259 | 4790 | 1056 | 5846 | 18.06 |
| 34 | Medium | | major_purchase | 11.9599 | 17793.7 | 8165 | 979 | 9144 | 10.71 |
| 41 | Medium | | wedding | 14.603 | 14252.7 | 265 | 38 | 303 | 12.54 |

Considering only the 20 most common Employer Titles we can see how the grades are distributed by different professionals.

Summary:

We can see that Director, Engineer, President, Vice President are the category's with highest incidence in Grade A;
Analyzing this table we can get some insights about the profile of grades and professionals.

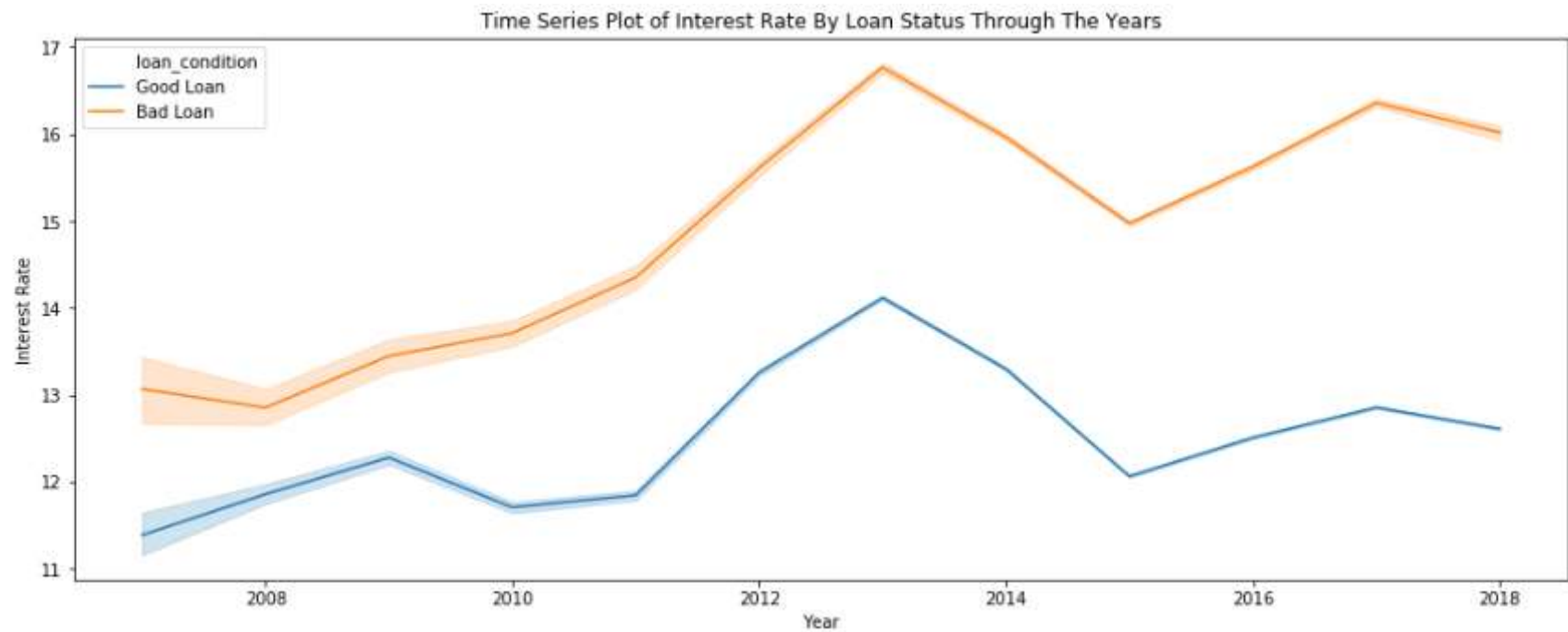
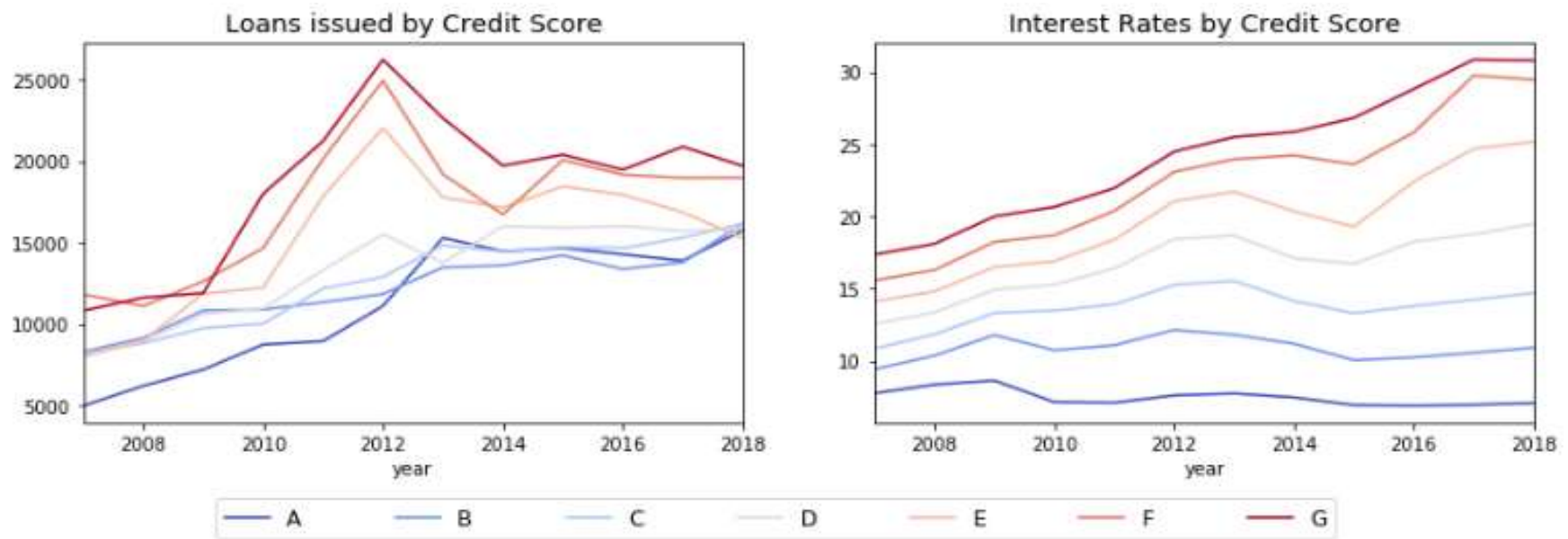
| sub_grade | A1 | A2 | A3 | A4 | A5 | B1 | B2 | B3 | B4 | B5 | C1 | C2 | C3 | C4 | C5 | D1 | D2 | D3 | D4 | D5 | E1 | E2 | E3 | E4 | E5 | F1 | F2 | F3 | F4 | F5 | G1 | G2 | G3 | G4 | G5 | | |
|--------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|--|--|
| emp_title | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Accountant | 3.43 | 3.41 | 3.45 | 4.81 | 5.22 | 6.34 | 6.18 | 5.88 | 6.18 | 6.49 | 6.59 | 5.65 | 5.56 | 5.4 | 4.84 | 3.13 | 2.97 | 2.97 | 2.11 | 1.98 | 1.41 | 1.16 | 0.98 | 0.79 | 1 | 0.48 | 0.29 | 0.27 | 0.23 | 0.21 | 0.3 | 0.11 | 0.07 | 0.04 | 0.07 | | |
| Director | 6.83 | 3.97 | 3.97 | 5.2 | 5.36 | 6.15 | 5.74 | 6.01 | 6.35 | 5.93 | 6.6 | 5.08 | 5.34 | 5.45 | 4.24 | 2.93 | 2.46 | 2.19 | 2.21 | 1.58 | 1.1 | 0.77 | 0.99 | 0.99 | 0.81 | 0.36 | 0.31 | 0.2 | 0.36 | 0.19 | 0.11 | 0.07 | 0.08 | 0.07 | 0 | | |
| Driver | 2.45 | 2.28 | 2.96 | 3.75 | 3.94 | 5.42 | 5.42 | 5.44 | 6.19 | 6.67 | 6.74 | 6.32 | 5.88 | 5.96 | 5.84 | 4.01 | 4.02 | 3.18 | 2.89 | 2.39 | 1.38 | 1.17 | 1.21 | 1.1 | 1.04 | 0.49 | 0.39 | 0.4 | 0.29 | 0.16 | 0.21 | 0.15 | 0.09 | 0.08 | 0.09 | | |
| Engineer | 6.43 | 4.44 | 4.49 | 5.26 | 5.63 | 5.98 | 5.89 | 6.12 | 5.89 | 5.5 | 6.02 | 5.6 | 4.81 | 4.83 | 4.65 | 2.88 | 2.52 | 2.62 | 2.07 | 1.49 | 1.44 | 0.96 | 0.88 | 0.71 | 0.86 | 0.63 | 0.37 | 0.33 | 0.16 | 0.1 | 0.16 | 0.08 | 0.05 | 0.08 | 0.07 | | |
| General Manager | 4.18 | 3.19 | 2.94 | 3.97 | 5.25 | 5.88 | 5.96 | 5.73 | 5.82 | 6.38 | 6.54 | 6.12 | 5.62 | 5.56 | 5.34 | 3.42 | 3.03 | 2.79 | 2.37 | 1.98 | 1.39 | 1.26 | 1.14 | 0.94 | 1.04 | 0.44 | 0.48 | 0.3 | 0.27 | 0.19 | 0.19 | 0.09 | 0.12 | 0.09 | 0.01 | | |
| Manager | 3.85 | 3.02 | 3.29 | 4.2 | 4.95 | 5.53 | 5.49 | 5.53 | 5.99 | 6.25 | 6.48 | 5.86 | 5.99 | 5.72 | 5.39 | 3.38 | 3.2 | 2.8 | 2.49 | 2.15 | 1.5 | 1.36 | 1.2 | 0.92 | 1.01 | 0.58 | 0.38 | 0.37 | 0.28 | 0.21 | 0.19 | 0.17 | 0.09 | 0.1 | 0.09 | | |
| Office Manager | 3.19 | 3.15 | 3.26 | 4.27 | 4.87 | 5.43 | 5.43 | 5.27 | 6 | 6.64 | 6.44 | 5.58 | 6.17 | 5.71 | 5.56 | 3.76 | 3.1 | 3.17 | 2.28 | 2.35 | 1.55 | 1.41 | 1.26 | 0.75 | 1.24 | 0.59 | 0.33 | 0.26 | 0.27 | 0.29 | 0.14 | 0.1 | 0.06 | 0.06 | 0.06 | | |
| Operations Manager | 3.85 | 3.08 | 3.38 | 4.11 | 5.12 | 5.89 | 5.79 | 5.32 | 5.76 | 6.02 | 6.58 | 6.2 | 5.83 | 5.74 | 4.99 | 3.46 | 3.21 | 2.33 | 2.63 | 2.25 | 1.57 | 1.24 | 1.32 | 0.86 | 1.17 | 0.6 | 0.26 | 0.34 | 0.33 | 0.18 | 0.24 | 0.08 | 0.08 | 0.1 | 0.05 | | |
| Owner | 4.56 | 3.84 | 3.58 | 4.46 | 4.81 | 6.77 | 6.24 | 5.97 | 6.14 | 6.04 | 6.55 | 5.95 | 5.2 | 5.28 | 4.81 | 3.56 | 3.02 | 2.53 | 2.08 | 1.88 | 1.18 | 1.13 | 0.99 | 0.78 | 0.77 | 0.37 | 0.29 | 0.28 | 0.25 | 0.22 | 0.19 | 0.1 | 0.07 | 0.06 | 0.08 | | |
| President | 7.06 | 4.56 | 4.67 | 5.85 | 5.47 | 6.48 | 6.2 | 5.59 | 5.74 | 5.84 | 6.76 | 5.18 | 4.84 | 4.8 | 4.2 | 2.78 | 2.42 | 2.21 | 1.79 | 1.45 | 0.99 | 0.9 | 0.95 | 1.01 | 0.69 | 0.35 | 0.34 | 0.21 | 0.09 | 0.1 | 0.14 | 0.13 | 0.08 | 0.07 | 0.05 | | |
| Project Manager | 5.66 | 4.01 | 3.5 | 4.88 | 5.7 | 6.22 | 5.84 | 5.69 | 5.84 | 6.44 | 6.26 | 5.65 | 5.34 | 5.3 | 4.78 | 2.76 | 2.54 | 2.43 | 2.26 | 1.97 | 1.19 | 1.14 | 1.11 | 0.79 | 0.84 | 0.5 | 0.36 | 0.22 | 0.19 | 0.15 | 0.15 | 0.07 | 0.06 | 0.05 | 0.07 | | |
| RN | 4.68 | 3.45 | 3.52 | 4.8 | 5.22 | 5.97 | 6 | 6.13 | 6.2 | 6.28 | 6.05 | 5.93 | 5.46 | 5.68 | 4.8 | 3.37 | 3.04 | 2.5 | 2.23 | 1.64 | 1.32 | 1.01 | 0.99 | 0.9 | 0.79 | 0.5 | 0.36 | 0.3 | 0.22 | 0.16 | 0.13 | 0.09 | 0.1 | 0.07 | 0.07 | | |
| Registered Nurse | 4.33 | 3.57 | 3.73 | 4.7 | 5.31 | 5.78 | 5.97 | 6.03 | 6.28 | 6.27 | 6.15 | 5.4 | 5.5 | 5.77 | 5.28 | 3.25 | 2.82 | 2.51 | 2.11 | 1.88 | 1.42 | 1.12 | 1.06 | 0.79 | 0.92 | 0.47 | 0.35 | 0.33 | 0.19 | 0.21 | 0.13 | 0.16 | 0.08 | 0.1 | 0.04 | | |
| Sales | 3.77 | 2.92 | 3.21 | 3.99 | 4.88 | 6.11 | 5.82 | 5.64 | 6.44 | 6.61 | 6.61 | 5.85 | 6 | 5.8 | 5.14 | 3.36 | 3.2 | 2.68 | 2.57 | 1.92 | 1.46 | 1.21 | 1.04 | 0.91 | 0.82 | 0.39 | 0.33 | 0.3 | 0.24 | 0.21 | 0.15 | 0.15 | 0.07 | 0.1 | 0.08 | | |
| Supervisor | 2.91 | 2.53 | 2.63 | 3.55 | 3.98 | 4.96 | 5.15 | 5.51 | 5.68 | 6.78 | 7.07 | 5.71 | 6.31 | 5.84 | 5.77 | 3.99 | 3.64 | 2.83 | 2.85 | 2.38 | 1.57 | 1.61 | 1.42 | 1.27 | 1.25 | 0.71 | 0.45 | 0.41 | 0.28 | 0.32 | 0.27 | 0.12 | 0.1 | 0.06 | 0.07 | | |
| Teacher | 3.97 | 3.28 | 3.52 | 4.42 | 4.99 | 6.1 | 5.91 | 5.97 | 6.25 | 6.39 | 6.46 | 5.77 | 5.53 | 5.61 | 5.21 | 3.45 | 3.15 | 2.61 | 2.23 | 1.99 | 1.39 | 1.14 | 1.03 | 0.84 | 0.89 | 0.48 | 0.3 | 0.3 | 0.22 | 0.18 | 0.15 | 0.07 | 0.06 | 0.07 | 0.07 | | |
| Vice President | 7.41 | 4.72 | 4.22 | 5.41 | 5.75 | 5.99 | 5.36 | 5.79 | 5.64 | 5.75 | 6.15 | 5.46 | 5.58 | 4.72 | 4.19 | 2.88 | 2.52 | 2.37 | 2.08 | 1.58 | 1.17 | 0.99 | 0.87 | 0.87 | 0.65 | 0.49 | 0.29 | 0.29 | 0.2 | 0.19 | 0.09 | 0.12 | 0.17 | 0.02 | 0.03 | | |
| manager | 2.54 | 2.32 | 2.9 | 3.46 | 3.8 | 5.57 | 5.04 | 5.08 | 6.29 | 6.63 | 6.71 | 6.36 | 6.27 | 5.58 | 5.48 | 4.39 | 3.92 | 2.9 | 2.75 | 2.27 | 1.7 | 1.42 | 1.46 | 1.1 | 1.05 | 0.76 | 0.47 | 0.47 | 0.33 | 0.17 | 0.3 | 0.14 | 0.16 | 0.14 | 0.07 | | |
| owner | 4.23 | 3.36 | 3.74 | 4.11 | 4.81 | 6.62 | 6.24 | 5.8 | 6.32 | 5.91 | 7.04 | 5.78 | 5.55 | 5.17 | 5.27 | 3.55 | 2.8 | 2.8 | 2.43 | 1.83 | 1.26 | 1.32 | 0.98 | 0.63 | 0.67 | 0.39 | 0.32 | 0.32 | 0.21 | 0.13 | 0.06 | 0.12 | 0.11 | 0.08 | 0.05 | | |
| teacher | 3.65 | 3.33 | 2.97 | 4.21 | 5.33 | 6.2 | 6.22 | 6.14 | 6.23 | 6.69 | 6.68 | 5.53 | 5.9 | 5.38 | 5.23 | 3.51 | 2.9 | 2.6 | 2.17 | 1.91 | 1.42 | 1.36 | 1.03 | 0.88 | 0.7 | 0.49 | 0.33 | 0.36 | 0.16 | 0.12 | 0.09 | 0.09 | 0.09 | 0.04 | 0.03 | | |

04 RISK ANALYTICS

Credit scores are important metrics for assessing the overall level of risk.

By plotting the loans issued by credit score, the interest rates by credit score and the interest rate through the years the conclusions that we made are that:

- The scores that has a lower grade received a larger amounts of loans (which might had contributed to a higher level of risk).
- Logically, the **lower the grade the higher the interest** the customer had to pay back to investors.
- Remember also, most loan statuses are “**Current**” so there is a risk that at the end of maturity some of these loans might become bad loans.



In the figure below, the risk analyses of the project have been made.

The factors mentioned could have a severe impact on the project, if no action would be taken in case they would occur.

| ID | Risks | Chance | Impact | Rate | Actions |
|----|----------------------------|--------|--------|------|---|
| 1 | Co-worker is sick | M | M | 14 | Contact with a doctor |
| 2 | Algorithm not trained well | M | H | 12 | Use more data to train the algorithm or select different algorithm. |
| 3 | Internet down unexpected | L | M | 4 | Call the maintenance team and ask for help. |
| 4 | Data breach | L | H | 6 | Call customers and inform them. |

05 MODELING

As it can be seen from the graph, approximately 14% of the loans are “Bad Loans”.

Also, it was expected that the company will take some kind of actions after the scandal and the fraud that happened on 2016 and the “Bad Loans” will decrease but still it is noticeable that there is again slight increase after 2017.

The goal of this project will be to create a model which predict whether a loan is most likely to be return. A binary classification system is used, in which the values for the loan status field are classified into two categories:

1: "Fully Paid" or "Current" – Good Loan

0: "Late" (for any time period) or "Charged Off" – Bad Loan

The data has 145 features but the following have the most relation with our may goal and they will be used to predict the loan status category (descriptions are from the "LCDataDictionary.xlsx" file):

- **loan_amount:** The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
- **Interest_rate:** Interest Rate on the loan.
- **annual_income:** The self-reported annual income provided by the borrower during registration.
- **delinq_2yrs:** The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
- **dti:** A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
- **emp_length:** Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
- **tot_cur_bal:** Total current balance of all accounts.
- **purpose:** A category provided by the borrower for the loan request.
- **home_ownership:** The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
- **recoveries:** Post charge of gross recovery.
- **total_rec_prncp:** Principal received to date.
- **out_prncp:** Remaining outstanding principal for total amount funded.
- **last_pymnt_d:** Last month payment was received.

Finally, for the purpose, we are going to classify if the loan status will be a "Good" one or "Bad" one by using and comparing the following algorithms:

- Decision tree classifier
- Random Forest (tree based model)
- Neural Network

REFERENCES

Lending Club. (n.d.). Retrieved from Wikipedia: <https://en.wikipedia.org/wiki/LendingClub>