**LENDING CLUB LOAN DEFAULT PREDICTION**

MACHINE LEARNING ALGORITHM

**CAPSTONE PROJECT – MILESTONE 1**

|  |  |  |  |
| --- | --- | --- | --- |
| VERSION HISTORY | DATE | PREPARER | APPROVER |
| ISSUED FOR INFORMATION | 04/14/2020 | JUGAL SHAH | KEVIN GLYNN |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Table of Contents

[1. SUMMARY 4](#_Toc41831142)

[2. ABOUT LENDING CLUB 4](#_Toc41831143)

[3. INFORMATion about datasets 6](#_Toc41831144)

[4. DATA CLEANING METHODOLOGY 7](#_Toc41831145)

[5. EXPLORATORY DATA ANALYSIS 11](#_Toc41831146)

[6. Train and test data set conversion 18](#_Toc41831147)

[7. STATISTICAL ANALYSIS OF LENDING CLUB LOAN DATA 19](#_Toc41831148)

[8. MACHINE LEARNING PROCESS 25](#_Toc41831149)

[9. conclusion 29](#_Toc41831150)

[10. FUTURE WORK 29](#_Toc41831151)

[APPENDIX I - LENDING CLUB LOAN DATA DICTIONARY 30](#_Toc41831152)

[APPENDIX II - DATA CLEANING TABLES 36](#_Toc41831153)

List of Figures

[Figure 3‑1: Lending Club Loan Data Size 6](#_Toc41831154)

[Figure 4‑1: Columns data where data is missing between 10% and 60% 8](#_Toc41831155)

[Figure 4‑2: Lending Club Loan Data – After Initial Data Cleaning 9](#_Toc41831156)

[Figure 5‑1: Interest rate distribution for different loan grades and loan terms 11](#_Toc41831157)

[Figure 5‑2: Proportion of ‘Good Loans’ vs ‘Bad Loans’ – Overall and By years 12](#_Toc41831158)

[Figure 5‑3: Loan Amount and Funded Amount distribution 12](#_Toc41831159)

[Figure 5‑4: Loan Installment distribution range 13](#_Toc41831160)

[Figure 5‑5: Total Number of Loans and Total Amount of Loans (in Billions) over the years. 13](#_Toc41831161)

[Figure 5‑6: Loan Default rate per year for different 36 months and 60 months loan terms 14](#_Toc41831162)

[Figure 5‑7: Loan Default rate per Loan Grade and Loan terms 14](#_Toc41831163)

[Figure 5‑8: Loan Default rate based on Employment Length for different Loan terms 15](#_Toc41831164)

[Figure 5‑9: Interest rate distribution over “Good Loans” vs “Bad Loans” 16](#_Toc41831165)

List of Tables

[Table 4‑1: Statistical Description of Columns between 10% and 60% data missing 8](#_Toc41831166)

[Table 5‑1: Number of “Good” and “Bad” Loans for Grade Level 16](#_Toc41831167)

[Table 5‑2: Number of “Good” and “Bad” Loans for Homeownership type 16](#_Toc41831168)

[Table 5‑3: Interest rate distribution over Loan Grades 17](#_Toc41831169)

[Table 5‑4: Proportion of Loan Status over Income verification 17](#_Toc41831170)

[Table 8‑1 TRAINING SET RESULT 26](#_Toc41831171)

[Table 8‑1 TRAINING SET RESULT 27](#_Toc41831172)

[Table 8‑2 TRAINING SET RESULT 27](#_Toc41831173)

[Table 8‑3 TRAINING SET RESULT 28](#_Toc41831174)

[Table 8‑4 TRAINING SET RESULT 28](#_Toc41831175)

[Table 8‑2: Columns with more than 60% of missing data 36](#_Toc41831176)

[Table 8‑3: Columns with missing data between 10% and 60% 37](#_Toc41831177)

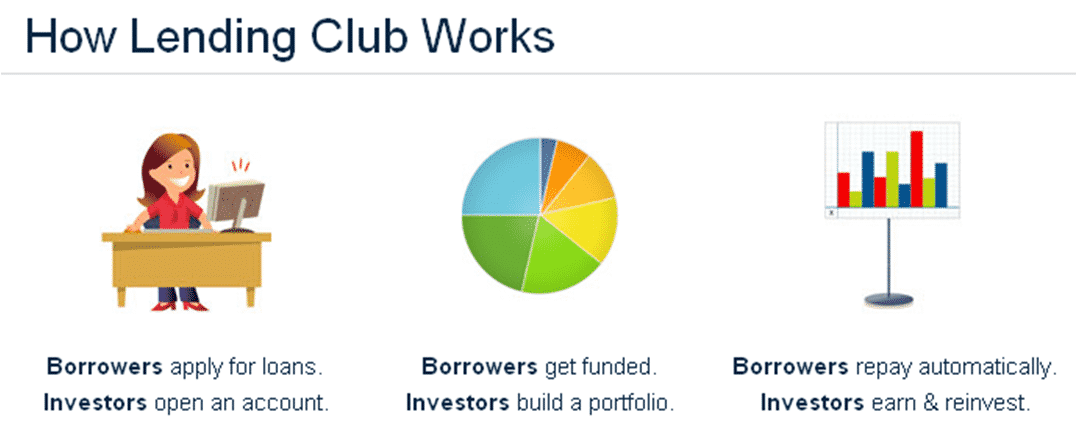
# SUMMARY

Credit and default risks have been at the forefront of financial news since the subprime mortgage crisis that began in 2008. Indeed, people realized that one of the main causes of that crisis was that loans were granted to people whose risk prole was too high. That is why, in order to restore trust in the finance system and to prevent this from happening again, banks and other credit companies have recently tried to develop new models to assess the credit risk of individuals even more accurately. Besides, the financialization of our economies implies that more and more stakeholders are involved, however, it can still be very difficult for some people - either because of their banking history or of their atypical situations - to get a loan. This imbalance has led to the development of new alternatives to the bank system. The number of peers to peer lending websites, MicroFinance Institutions (MFI) and companies that back their development, is currently growing quickly, and the quite recent stock market listing of LendingClub is adding evidence of that. It is precisely in that dynamic that this project fits, its main goal is to predict if a consumer will experience delinquency during the period of the loan.

# ABOUT LENDING CLUB

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 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 3-year or 5-year. Investors can search and browse the loan listings on the 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.



Lending Club provides the "bridge" between investors and borrowers. While Lending Club does not directly lend money to borrowers, it provides an important platform to select borrowers based on their employment length, income, mortgage, and other parameters. Based on these criteria Lending Club put borrowers into one of seven loan grades. Typical Loan Grades are A Thru G where interest rate typically increases as it moves from A to G.

# INFORMATion about datasets

Dataset for this project is obtained from ‘Kaggle.com’. These files contain complete loan data for all loans issued through 2011-2018, 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 2.2 Million observations and 145 variables. A data dictionary is provided in a separate file. Refer to Appendix 1 for the data dictionary.

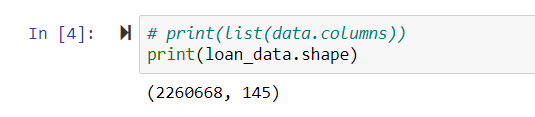


Figure ‑: Lending Club Loan Data Size

# DATA CLEANING METHODOLOGY

Current datasets contain more than 2.2 million rows and 145 columns. While initially examine the data, it was realized that not all columns of the dataset are useful. The following steps were taken in order to develop a clean dataset.

1. **Dropping columns with more than 60% data:** There were several columns within datasets that have more than 60% of data is missing. It is possible to predict these missing values based on statistical methods and other criteria, however filling more than 60% missing data is not a good strategy as it may contaminate dataset and may result in biased datasets. Hence, the best method is to drop all the columns that have more than 60% missing datasets. There are a total of 42 columns with more than 60% of the data missing. Refer Table 10‑1 for details of the column list with missing values more than 60%.
2. **Analyzing columns with more than 10% missing data:**  After dropping that, we have 2.2 million rows and a total of 103 columns. Now, it is also possible to drop all the rows that have missing data, but it is possible that we may waste our data. So, we will predict missing values for columns where the missing value is more than 10% and 60%. We will use statistical mode, medians, and means in order to predict these missing values. Table 5‑2 for details of columns list with missing values more than 10%.
3. **Predicting missing values for columns with more than 10% missing data:** There were 17 columns with more than 10% of data missing. Most of the columns that have more than 10% of the data missing were numeric. Refer Table 4 for details of columns' statistical description. By looking at the name of columns, it makes more sense to use the statistical “mode” parameter for filling out those missing values.

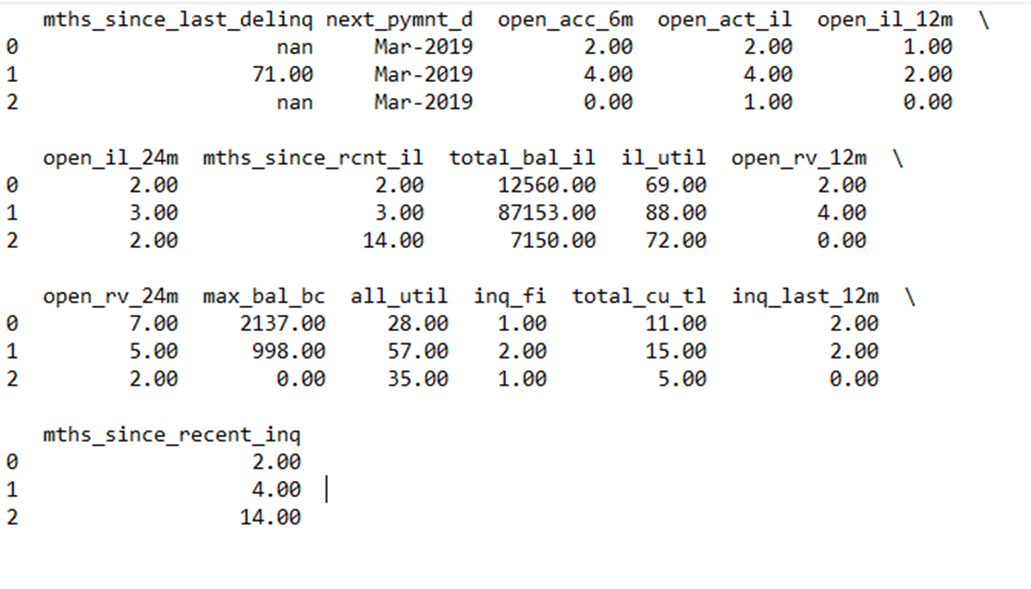


Figure ‑: Columns data where data is missing between 10% and 60%

Table ‑: Statistical Description of Columns between 10% and 60% data missing

|  | **Count** | **Percent** | **mode** | **median** | **mean** |
| --- | --- | --- | --- | --- | --- |
| mths\_since\_last\_delinq | 1158502 | 51.2 | 12 | 31 | 34.5 |
| open\_acc\_6m | 866130 | 38.3 | 0 | 1 | 0.9 |
| open\_act\_il | 866129 | 38.3 | 1 | 2 | 2.8 |
| open\_il\_12m | 866129 | 38.3 | 0 | 0 | 0.7 |
| open\_il\_24m | 866129 | 38.3 | 1 | 1 | 1.6 |
| mths\_since\_rcnt\_il | 909924 | 40.3 | 7 | 13 | 21.2 |
| total\_bal\_il | 866129 | 38.3 | 0 | 23127 | 35507 |
| il\_util | 1068850 | 47.3 | 78 | 72 | 69.1 |
| open\_rv\_12m | 866129 | 38.3 | 0 | 1 | 1.3 |
| open\_rv\_24m | 866129 | 38.3 | 1 | 2 | 2.7 |
| max\_bal\_bc | 866129 | 38.3 | 0 | 4413 | 5806.4 |
| all\_util | 866348 | 38.3 | 59 | 58 | 57 |
| inq\_fi | 866129 | 38.3 | 0 | 1 | 1 |
| total\_cu\_tl | 866130 | 38.3 | 0 | 0 | 1.5 |
| inq\_last\_12m | 866130 | 38.3 | 0 | 1 | 2 |
| mths\_since\_recent\_inq | 295435 | 13.1 | 1 | 5 | 7 |

1. **Dropping rows with missing data:** Dataset had columns where less than 10% of data is missing. These rows would be simply dropped as datasets is very large and still, remaining datasets contains enough information to develop a machine learning algorithm.

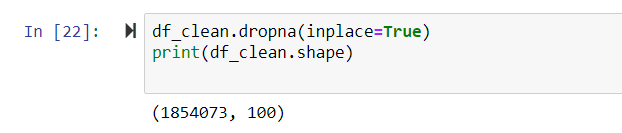


Figure ‑: Lending Club Loan Data – After Initial Data Cleaning

1. **Dropping columns with unnecessary information:** There were several columns in datasets, which will not contribute to the prediction algorithm. These columns can be dropped in order to reduce the size of datasets. These columns can be divided into two categories.
   * 1. Non-numeric columns such as payment date, next payment date, etc.
     2. Columns where most of the rows contain the same data. Columns where more than 80% of data is redundant, it can be dropped.

There were a total of 66 columns that have information does not require for developing the machine learning algorithm. Some of the columns have over 99% of rows contain the same data. These columns would not contribute anything to our prediction algorithm. Several columns contain non-numeric descriptive information. These columns are dropped.

1. **Eliminating unnecessary strings or words:** Employment length column has keywords like *“<1 year*”, *“5 years”*, *“10+ years”* etc. This information can be converted to numeric data by eliminating string *“years”*. Also, *“<1 year”* is converted to 0 and *“10+ years”* is converted to 10.
2. **Creating a categorical variable:** Three of the columns are converted to Categorical variables. These three columns are “*max\_bal\_bc*”, “*num\_sats*” and “*pct\_tl\_nvr\_dlq*”. The three columns contain data into continuous numbers and in order to make meaningful use of it, it is converted to a categorical variable using Pandas “*pandas.cut*” command.
3. **Creating “Loan Status Column”:**  Currently “Loan Status” column has the following categories.
   * 1. Paid off
     2. Current
     3. Late by 30 days
     4. Late by 120 days
     5. In the Grace period
     6. Default

For all loan status from no.3 to no.6 can be considered as “Bad Loan” as there is no guaranteed chance loan can be paid off fully. Loans that are completely paid off are considered as “Good Loan”. For Loan Status “Current”, we are not sure if the loan will be completely paid off or go to the grace period. So, we will be considered that loan status as “Unknown”. These unknown categories will form as test dataset. Refer to section 6 for more details.

The results of the “dataframe“are saved in the ‘CSV’ format to use it for the next section. The next section will include exploratory data visualization to get graphical information from data.

# EXPLORATORY DATA ANALYSIS

A clean dataframe is used to perform exploratory data analysis. This analysis gives an idea regarding the relationship between each parameter and the overall distribution of data.

Parameters are analyzed based on loan status which is ‘good loan’ and ‘bad loan’.

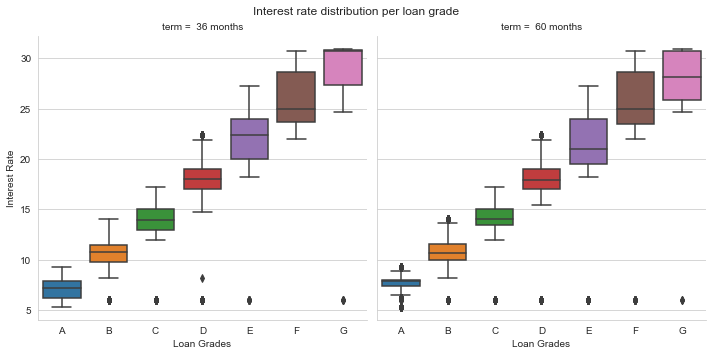


Figure ‑: Interest rate distribution for different loan grades and loan terms

As shown in Figure 5‑1, the interest rate is increasing as loan grade changes from Grade A to Grade G. There are very few outliers and those can be ignored. Also, for the same loan grade, there is a slightly higher interest rate for 36 months terms as compared to 60 months terms. This indicates interest rates are higher for the shorter loan term. Further, in Section 7.1, a statistical hypothesis test is performed to analyze interest rate distribution over loan grades.

As shown in Figure 5 2, a pie chart indicates overall 21.5% of loans are ‘Bad Loans’. It is important to mention that, this pie chart deliberately excludes loans with loan status as ‘Current’. Since outcomes of these loans are unknown, it is not included in exploratory data analysis.

Also as shown in Figure 5 2, there is consistency between proportions of “Good Loans” and “Bad Loans” over the number of years. In a particular year, where numbers of “Good Loans” increase, also the numbers of “Bad Loans” increases.

Also as shown in Figure 5‑3, Distribution of Loan amount and Funded Amount is presented. By looking at the chart, the distribution of loan amounts looks normally distributed with the right-skewed tail. The majority of Loans are ranging from $5000 to $15000.

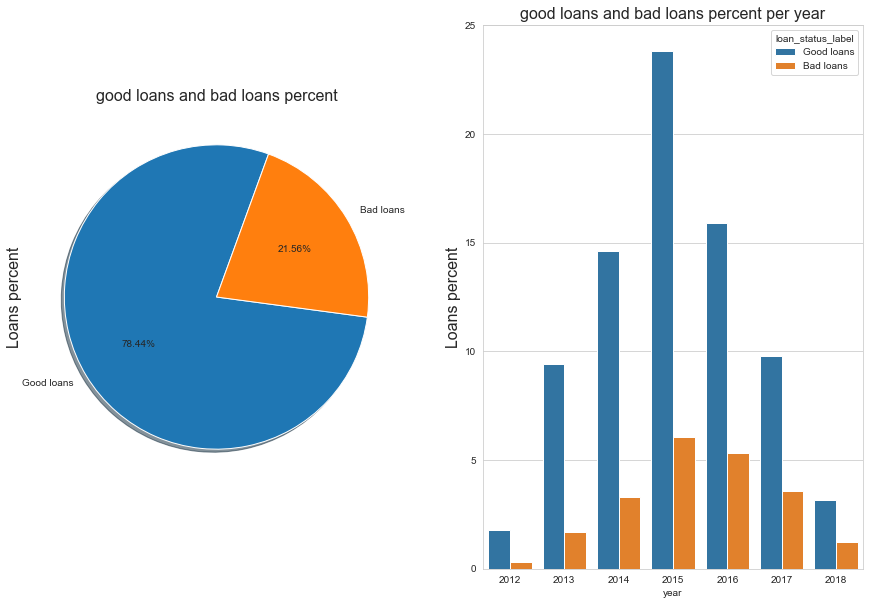


Figure ‑: Proportion of ‘Good Loans’ vs ‘Bad Loans’ – Overall and By years

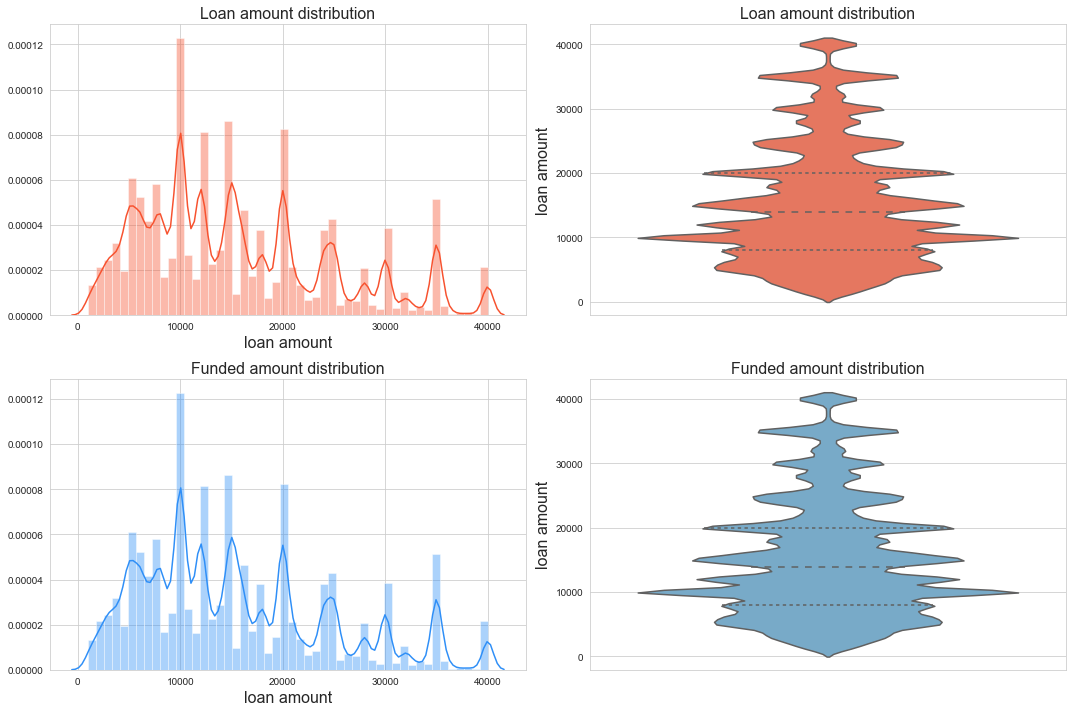


Figure ‑: Loan Amount and Funded Amount distribution

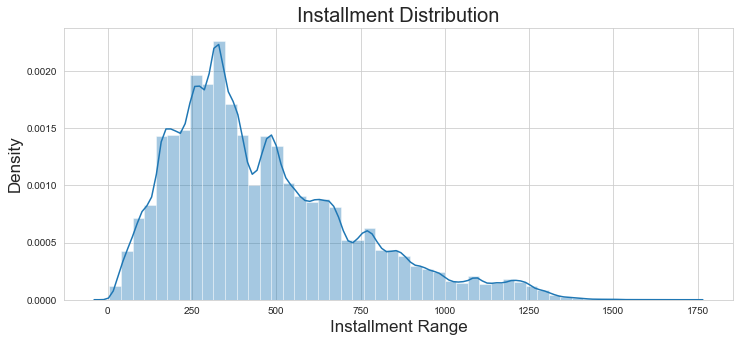


Figure ‑: Loan Installment distribution range

Loan Installment distribution range follows the normal distribution curve with the right tail skewed. It is obvious that the majority of installment ranges between $250 to $750 with very few installments are above $1000.

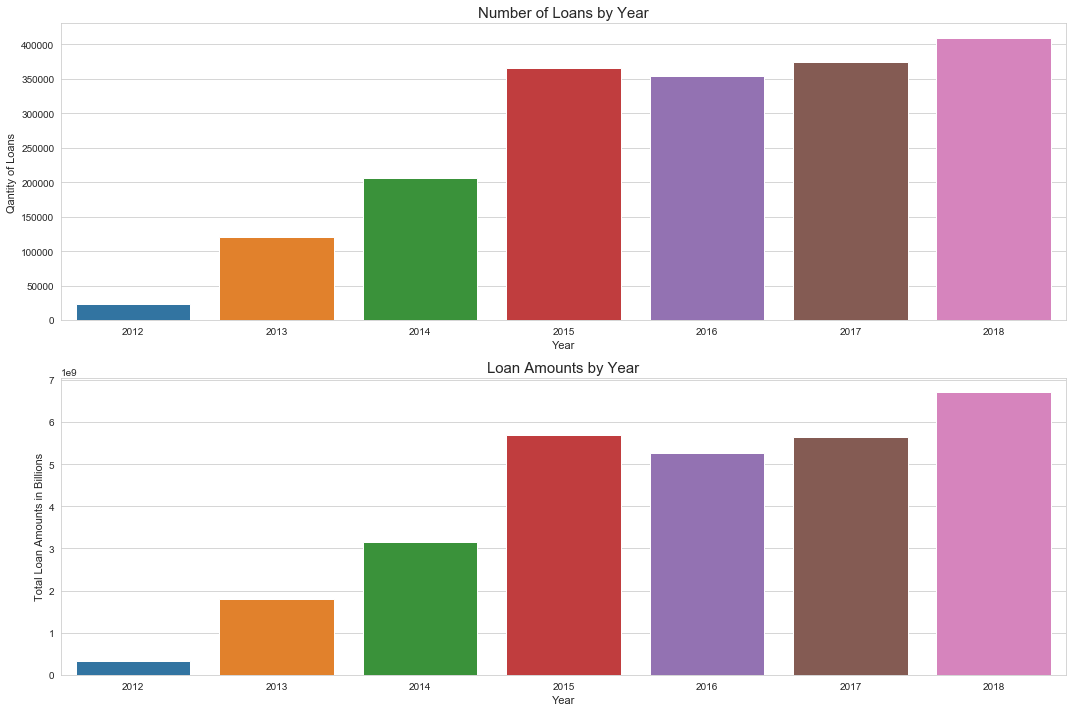


Figure ‑: Total Number of Loans and Total Amount of Loans (in Billions) over the years.

Lending Club is constantly growing as it is seen from the above figure, there are more loans issued in recent years compared to earlier years.

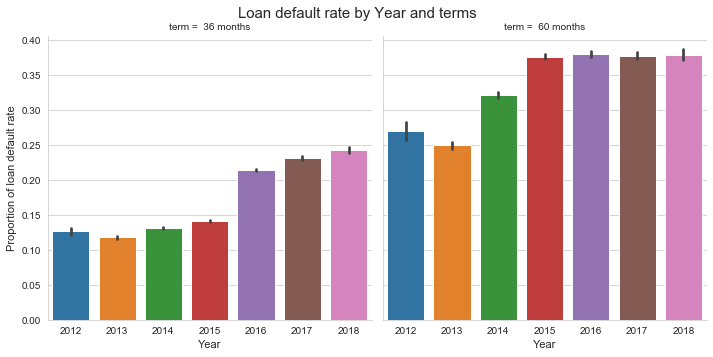


Figure ‑: Loan Default rate per year for different 36 months and 60 months loan terms

As shown in Figure 5‑6, the loan default rate over the years increases. As an example, for a 3-year loan period in 2012 around 12.5% of loans were resulted in default, whereas in 2018 around 25% of loans were resulted in default. A number of bad loans for the 5-year term are even higher. This is a concern as the company is issuing more loans every year and hence the percentage of bad loans is also increasing.

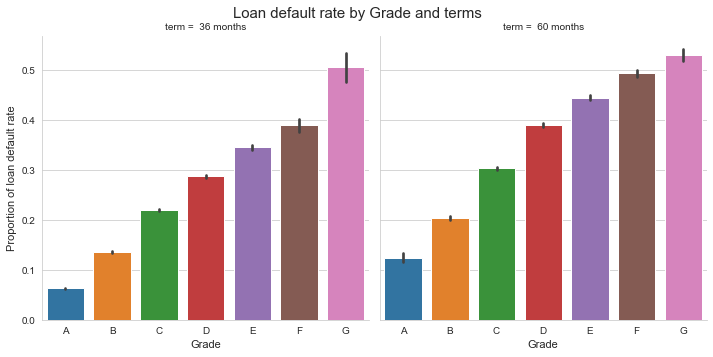


Figure ‑: Loan Default rate per Loan Grade and Loan terms

Figure 5‑7 shows the Proportion of bad loans over each grade. This is correlated with interest rate distribution for each grade. As loan grades move from A to G, interest rates increase as Grade A is least risky loan vs grade G is the high-risk loan. Hence, there is a higher chance of default as well higher interest rate. It is concluded that as risk on loan increases, interest rate and proportion of “Bad Loans” both increase.

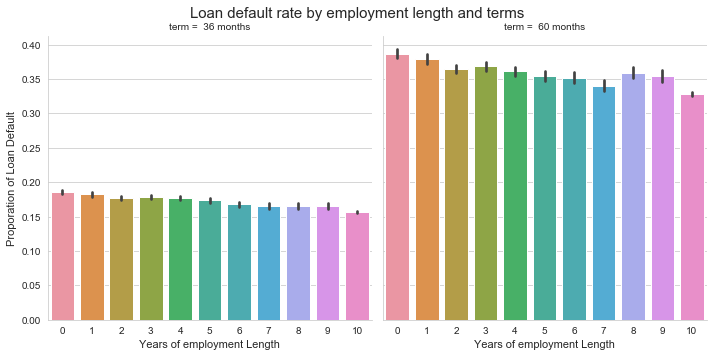


Figure ‑: Loan Default rate based on Employment Length for different Loan terms

Another important parameter is the length of employment. Longer employment is considered as stable and hence risk on loan is lower. Figure 5‑8, shows the proportion of bad loans for the range of employment length. The lowest employment length indicated by “0” is less than a year with a current employer. The highest employment length indicated by “10” is 10 or more than 10 years of employment with current employer. As shown in Figure 5‑8, there is a slight decrease in the proportion of bad loans as employment length increases. However, it is not conclusive if there is a correlation between employment length and “Bad loans”. It may require performing additional statistical hypotheses tests to conclude a correlation between Years of employment length and “Bad Loan” status.

Figure 5‑9 shows interest rate distribution for “Good Loans” and “Bad Loans” for 3-year and 5-year loan terms. For both terms, the overall median and 95% confidence interval is higher for “Bad Loans”. It clearly indicates a strong relation between loan status and 3 parameters as mentioned below.

* Loan Terms
* Loan Grade
* Interest rate

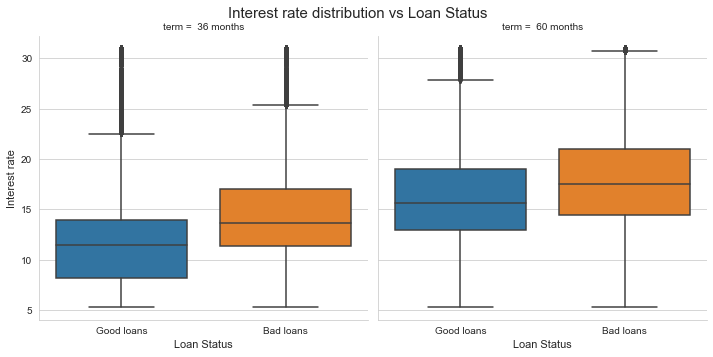


Figure ‑: Interest rate distribution over “Good Loans” vs “Bad Loans”

Table ‑: Number of “Good” and “Bad” Loans for Grade Level

| **GRADE** | **A** | **B** | **C** | **D** | **E** | **F** | **G** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **LOAN\_STATUS\_LABEL** |  |  |  |  |  |  |  |
| **BAD LOANS** | 12252 | 45249 | 76074 | 52894 | 31147 | 12378 | 4052 |
| **GOOD LOANS** | 176387 | 269211 | 235119 | 108381 | 44993 | 13829 | 3646 |

Table ‑: Number of “Good” and “Bad” Loans for Homeownership type

| **HOME\_OWNERSHIP** | **ANY** | **MORTGAGE** | **OWN** | **RENT** |
| --- | --- | --- | --- | --- |
| **LOAN\_STATUS\_LABEL** |  |  |  |  |
| **BAD LOANS** | 0 | 0.44 | 0.11 | 0.46 |
| **GOOD LOANS** | 0 | 0.52 | 0.1 | 0.38 |

Table ‑: Interest rate distribution over Loan Grades

| **GRADE** | **A** | **B** | **C** | **D** | **E** | **F** | **G** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **LOAN\_STATUS\_LABEL** |  |  |  |  |  |  |  |
| **BAD LOANS** | 7.36 | 10.76 | 14.11 | 17.9 | 21.36 | 25.34 | 28.26 |
| **GOOD LOANS** | 7.08 | 10.63 | 14.01 | 17.73 | 21.22 | 25.07 | 27.86 |

Table ‑: Proportion of Loan Status over Income verification

| **VERIFICATION\_STATUS** | **NOT VERIFIED** | **SOURCE VERIFIED** | **VERIFIED** |
| --- | --- | --- | --- |
| **LOAN\_STATUS\_LABEL** |  |  |  |
| **BAD LOANS** | 0.22 | 0.45 | 0.33 |
| **GOOD LOANS** | 0.32 | 0.41 | 0.27 |

# Train and test data set conversion

A clean lending club loan dataset contains more than 1.8 million rows and 37 columns. Our Target Variable is “Loan Status” which is categorized into Good Loan, Bad Loan and Unknown.

Dataset excluding “Unknown” loan status is created to develop a prediction algorithm for lending club loan datasets. Further, this dataset is separated as Train and test dataset as ration of 70% and 30% respectively.

Once a satisfactory prediction score is achieved, this algorithm will be used to predict unknown loan status.

Hence, the following datasets will be generated.

1. Lending Club Loan – known Dataset
   1. It contains 70% of dataset as train dataset
   2. It contains randomly selected 30% dataset as the test dataset
2. Lending Club Loan – Unknown Dataset

# STATISTICAL ANALYSIS OF LENDING CLUB LOAN DATA

Statistical analysis is performed on loan datasets in order to determine the correlation between various parameters as well as verifying confidence interval for different terms. This is an important step in order to identify outliers as outlier can skew prediction result significantly.

Statistical analysis performed in this report as shown below.

1. Correlation between the interest rate and Loan Grade
2. Effect of Loan terms (3-year and 5-year) on Loan Default rate
3. Distribution of loan interest rate over “Good Loans” and “Bad Loans”

## Loan grade and interest rate

A statistical hypothesis test will be performed to determine the correlation between loan grade and interest rate. The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of an interest rate for different loan grades.

The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other. Specifically, it tests the null hypothesis:

One-way ANOVA Null Hypothesis

where µ = group mean and k = number of groups. If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis (Ha), which is that there are at least two group means that are statistically significantly different from each other.

There are three main assumptions for this test.

1. The sample of the dependent variable is normally distributed in each group that is being compared in the one-way ANOVA.
2. Independence of observations.
3. There is the homogeneity of variances.

The solution to this problem takes four steps:

1. state the hypotheses

2. formulate an analysis plan

3. analyze sample data

4. interpret results.

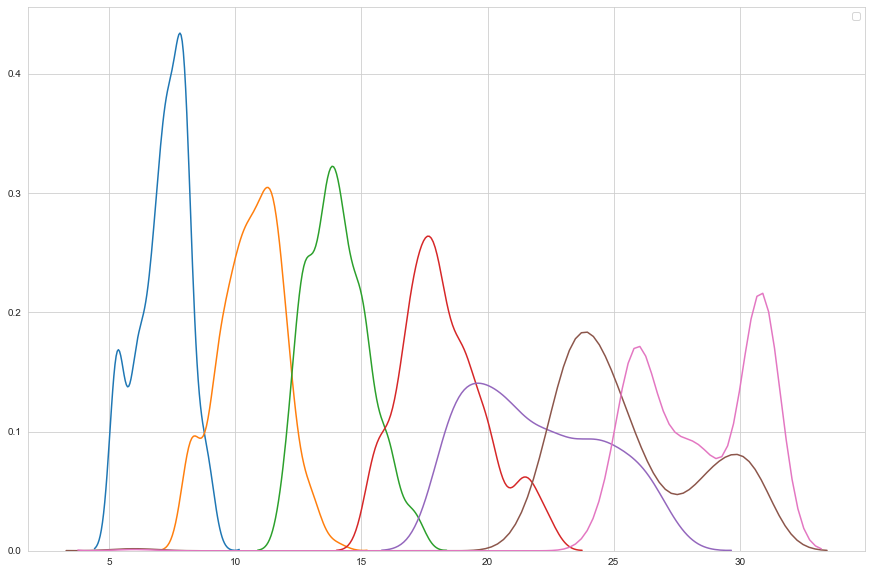
**STATE THE HYPOTHESES**

**Null Hypotheses or H0**: The interest rate is evenly distributed among loan grades. Which means the mean of an interest rate for all loan grade is same.

**Alternative Hypotheses or Ha**: The interest rate is not evenly distributed among loan grades. The mean of the interest rate for each loan grade is different.

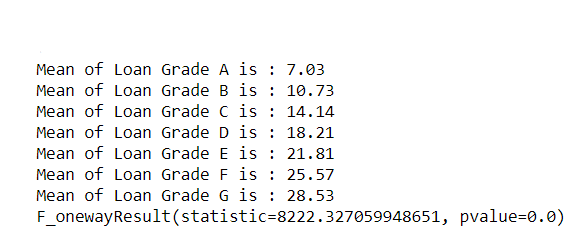
**FORMULATE AN ANALYSIS PLAN**

The sample size of 500 for the interest rate is collected from each loan grade. As shown in the figure, sample distribution is normally distributed.

****

**ANALYZE SAMPLE DATA**

Using Python’s “Statsmodel” module, the One Way ANOVA test is performed. The results are as shown below.



As seen above, the p-value is almost 0.0 which is less than specified significant level 0.05. So Null Hypotheses is rejected. Hence, it is concluded that there is a significant correlation between loan grades and interest rates.

## EFFECT OF LOAN TERMS ON DEFAULT RATE

There is an equal amount of loan default rate between 36 months and 60 months loan terms. In other words, loan terms of 36 months or 60 months have no effect on the loan default rate.

To test this claim, a random sample of 500 loan status is chosen from every 36 months and 60 months from a population of more than 1.8 million.

At the end of the sample study, **19.2% of loans in 36 months** results in default rate while **37.8% of loans in 60 months** results in default rate. Based on these findings, the hypothesis test is conducted to verify a claim that the loan default rate is equal for 36 months and 60 months term. A significance level of 0.05 is used.

The test procedure, called the **two-proportion z-test**, is appropriate when the following conditions are met:

* The sampling method for each population is [simple random sampling](https://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The samples are [independent](https://stattrek.com/Help/Glossary.aspx?Target=Independent).
* Each sample includes at least 10 successes and 10 failures.
* Each population is at least 20 times as big as its sample.

The solution to this problem takes four steps:

1. state the hypotheses

2. formulate an analysis plan

3. analyze sample data

4. interpret results.

We work through those steps below:

**State the hypotheses.**

The first step is to state the null hypothesis and an alternative hypothesis.

Null Hypotheses or H0: P1 = P2

Alternative Hypotheses or Ha: P1 ≠ P2

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the proportion from population 1 is too big or if it is too small.

**Formulate an analysis plan.**

For this analysis, the significance level is 0.05. The test method is two-proportion z-test.

**Analyze sample data**

Using sample data, the pooled sample proportion (p) and the standard error (SE) is calculated. Using those measures, the z-score test statistic (z) is computed.

p1 = 0.192

p2 = 0.378

n1 = 500

n2 = 500

p = (p1\*n1 + p2\*n2)/(n1+n2) = 0.285

standard error or se = sqrt[p\*(1-p)\*(1/n1 + 1/n2)] = 0.0272

z\_score or z = (p1-p2) / se = -6.24

From Normal Distribution CDF p-val = 2.17\*10^-10 ~= 0

Since p-val is less than the significance level 0.05, the null hypotheses is rejected. Hence, there is a significant effect of loan terms on the loan default rate.

## LOAN iNTEREST RATE VS LOAN STATUS

Within lending club loan data set, categorical variable 'loan\_status\_label' which is categorized into 'Good Loans' and 'Bad Loans'.

The mean interest rate for 500 samples of 'Good Loans' is 12.68, whereas for 500 samples of 'Bad Loans' is 15.48. Standard Deviation for 'Good Loans' is 4.68 whereas for 'Bad Loans' is 4.73, which is almost close and two-sample t-test assuming equal variance is conducted. A hypothesis test conducted to verify if interest rates are higher for ‘Bad Loans’ is just by chance. Hence, the difference in the mean interest rate between “Good Loan” and “Bad Loan” is 0.

We will use a 0.05 level of significance.

To test this claim, random samples of size 500 loan status from each 'Good Loans' and 'Bad Loans' is selected from a population of more than 2 million.

The solution to this problem takes four steps:

1. state the hypotheses
2. formulate an analysis plan
3. analyze sample data
4. interpret results.

**State the hypotheses.**

The first step is to state the null hypothesis and an alternative hypothesis.

**Null Hypotheses or**𝐻0**:**𝜇1**=**𝜇2

**Alternative Hypotheses or**𝐻𝑎**:**𝜇1≠𝜇2

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the proportion from population 1 is too big or if it is too small.

**Formulate an analysis plan.**

 For this analysis, the significance level is 0.05. Using sample data, we will conduct a two-sample t-test of the null hypothesis.

**Analyze sample data.**

Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t statistic test statistic (t).

* 𝜇1 = 12.68
* 𝜇2= 15.48
* s1 = 4.68
* s2 = 4.72
* n1 = 500
* n2 = 500
* SE = sqrt[(s12/n1) + (s22/n2)]
* standard error or se = sqrt[(s12/n1 + s22/n2)] = 0.298
* DF = [(s12/n1 + s22/n2)]2 / ((s12/n1)2 / (n1-1) + \

(s22/n2)2 / (n2-1)) = 998

t\_score or t = (𝜇1 - ) / se = 9.40

From Normal Distribution CDF p-val = 0

Since p-val is less than our significance level 0.05, we will reject our null hypotheses. Hence, Interest rates are not equally distributed among good loans and bad loans.

# MACHINE LEARNING PROCESS

There have been many studies on classification models predicting LendingClub loan default. Our classification goal is to predict which class the loan belongs to either **Default** or **Fully Paid**. In the following sections, we will share and discuss our experiments using Logistic Regression, Neutral Networks and Random Forest for classification problem.

We also measure precision, recall, f1-score (the harmonic mean of precision and recall) and weighted average as defined below.

Support = the number of true instances for each label Weighted-avg metric = metric weighted by support

## LOGISTIC REGRESSION

Logistic Regression takes in a list of features as input and outputs the Sigmoid of a linear combination of features weighted by learned parameters θ,

To derive optimal parameters, the model iteratively updates weights by minimizing the negative log likelihood with L2 regularization

To tackle the class imbalance problem (only 19% of our dataset are negative examples), we used balanced weight for class labels, which is inversely proportional to class frequencies in the input data:

After running Logistic Regression with the above setting for a maximum of 1000 iterations, we arrived at the following results:

Table ‑ TRAINING SET RESULT

|  |  |  |
| --- | --- | --- |
|  | **Predicted Default** | **Predicted Paid Off** |
| **Actual Default** | 131982 | 100932 |
| **Actual Paid Off** | 91681 | 141113 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Default** | 0.59 | 0.57 | 0.58 | 232914 |
| **Paid Off** | 0.58 | 0.61 | 0.59 | 232794 |
| **Weighted Avg** | 0.59 | 0.59 | 0.59 | 465708 |

As we can see, Logistic Regression is doing well compared to naive models that blindly predict positive for all examples, or randomly guess positive and negative with 50% chance. Thanks to L2 regularization, we did not observe overfitting issues.

Table ‑ TESTING SET RESULT

|  |  |  |
| --- | --- | --- |
|  | **Predicted Default** | **Predicted Paid Off** |
| **Actual Default** | 56439 | 43296 |
| **Actual Paid Off** | 39256 | 60599 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Default** | 0.59 | 0.57 | 0.58 | 99735 |
| **Paid Off** | 0.58 | 0.61 | 0.59 | 99855 |
| **Weighted Avg** | 0.59 | 0.59 | 0.59 | 199590 |

## Gradient Boosting

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets and have recently been used to win many Kaggle data science competitions.

The Python machine learning library Scikit-Learn, supports different implementations of gradient boosting classifiers, including XGBoost.

Classification algorithms frequently use logarithmic loss, while regression algorithms can use squared errors. Gradient boosting systems don't have to derive a new loss function every time the boosting algorithm is added, rather any differentiable loss function can be applied to the system.

Table ‑ TRAINING SET RESULT

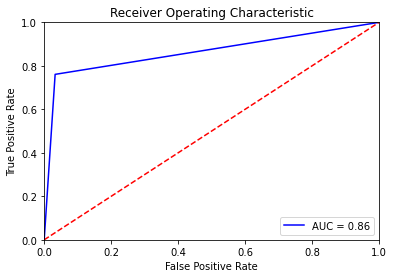
|  |  |  |
| --- | --- | --- |
|  | **Predicted Default** | **Predicted Paid Off** |
| **Actual Default** | 225897 | 7017 |
| **Actual Paid Off** | 55645 | 177149 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Default** | 0.8 | 0.97 | 0.88 | 232914 |
| **Paid Off** | 0.96 | 0.76 | 0.85 | 232794 |
| **Weighted Avg** | 0.88 | 0.87 | 0.86 | 465708 |

Table ‑ TESTING SET RESULT

|  |  |  |
| --- | --- | --- |
|  | **Predicted Default** | **Predicted Paid Off** |
| **Actual Default** | 96478 | 3257 |
| **Actual Paid Off** | 23877 | 75978 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Default** | 0.8 | 0.97 | 0.88 | 99735 |
| **Paid Off** | 0.96 | 0.76 | 0.85 | 99855 |
| **Weighted Avg** | 0.88 | 0.86 | 0.86 | 199590 |



Compared to Logistic regression mode, Gradient Boosting model achieves higher accuracy as well as higher precision.

## RANDOM FOREST TREE CLASSIFIER

Random Forest classifier is one of the tree ensemble methods that make decision splits using a random subset of features and combine the output of multiple weak classifiers to derive a strong classifier of lower variance at the cost of higher bias.

We started off our venture into Random Forest with 200 trees using Gini loss



Decision splits are based on at most 50 features to reduce variance. After training, we reached the following result:

Table ‑ TRAINING SET RESULT

|  |  |  |
| --- | --- | --- |
|  | **Predicted Default** | **Predicted Paid Off** |
| **Actual Default** | 232914 | 0 |
| **Actual Paid Off** | 0 | 232794 |

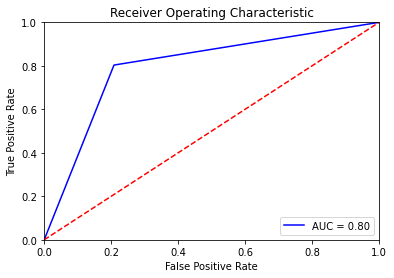
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Default** | 1 | 1 | 1 | 232914 |
| **Paid Off** | 1 | 1 | 1 | 232794 |
| **Weighted Avg** | 1 | 1 | 1 | 465708 |

Table ‑ TESTING SET RESULT

|  |  |  |
| --- | --- | --- |
|  | **Predicted Default** | **Predicted Paid Off** |
| **Actual Default** | 78946 | 20789 |
| **Actual Paid Off** | 19610 | 80245 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Default** | 0.8 | 0.79 | 0.8 | 99735 |
| **Paid Off** | 0.79 | 0.8 | 0.8 | 99855 |
| **Weighted Avg** | 0.8 | 0.8 | 0.8 | 199590 |

Although the performance is on par with Gradient Boosting, Random Forest’s overfitting problem is much more prominent than any other models even after restricting the maximum number of features considered for decision splits to 50.



# conclusion

Based on our explorations with Logistic Regression, Neural Network and Random Forest, we can achieve weighted average of 0.95 for both precision and recall. More specifically, our classification results appear to be more logically reasonable and practical. However, classification models can only predict the probability of loan defaults. This does not offer us a very fine-grained view in terms of how much return each loan can generate, which is essential for investors. Therefore, we would also like to predict the expected return rate, which naturally leads to our experiments with regression models in future.

# FUTURE WORK

We obtained a classification prediction based on the training set and simulated the strategy on the test set.

Both sets comprise loans initiated within the same periods (2012-2015). We can check to see if the strategy generalizes to future loans by testing it on 2016-2018 loans that have finalized. Practically speaking, this would be a much more useful metric for investors.

We worked with a 70% training and 30% test split for simplicity in this project. The absence of a development

set did not afford us much opportunity to tune the hyperparameters of our models, such as the number of decision trees to use in random forest models, and the number of hidden layers and neurons of each layer in neural network models. Having a small development set would enable us to tune some hyper-parameters quickly to help improve model performance metrics.

We can also make better use of existing features in the Lending Club dataset. One example is loan description which the borrower enters at the time of loan application. Instead of dropping such freeform features, we can try applying some statistical natural language processing techniques.

1. LENDING CLUB LOAN DATA DICTIONARY

|  |  |
| --- | --- |
| **LoanStatNew** | **Description** |
| acc\_now\_delinq | The number of accounts on which the borrower is now delinquent. |
| acc\_open\_past\_24mths | The number of trades opened in the past 24 months. |
| addr\_state | The state provided by the borrower in the loan application |
| all\_util | Balance to the credit limit on all trades |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers during registration |
| application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| avg\_cur\_bal | The average current balance of all accounts |
| bc\_open\_to\_buy | Total open to buy on revolving bankcards. |
| bc\_util | The ratio of total current balance to high credit/credit limit for all bank card accounts. |
| chargeoff\_within\_12\_mths | Number of charge-offs within 12 months |
| collection\_recovery\_fee | post charge off collection fee |
| collections\_12\_mths\_ex\_med | Number of collections in 12 months excluding medical collections |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| delinq\_amnt | The past-due amount owed for the accounts on which the borrower is now delinquent. |
| desc | Loan description provided by the borrower |
| 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. |
| dti\_joint | A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income |
| earliest\_cr\_line | The month the borrower's earliest reported credit line was opened |
| 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. |
| emp\_title | The job title supplied by the Borrower when applying for the loan.\* |
| fico\_range\_high | The upper boundary range the borrower’s FICO at loan origination belongs to. |
| fico\_range\_low | The lower boundary range the borrower’s FICO at loan origination belongs to. |
| funded\_amnt | The total amount committed to that loan at that point in time. |
| funded\_amnt\_inv | The total amount committed by investors for that loan at that point in time. |
| grade | LC assigned loan grade |
| home\_ownership | The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER |
| id | A unique LC assigned ID for the loan listing. |
| il\_util | Ratio of total current balance to high credit/credit limit on all install acct |
| initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| inq\_fi | Number of personal finance inquiries |
| inq\_last\_12m | Number of credit inquiries in past 12 months |
| inq\_last\_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| installment | The monthly payment owed by the borrower if the loan originates. |
| int\_rate | Interest Rate on the loan |
| issue\_d | The month which the loan was funded |
| last\_credit\_pull\_d | The most recent month LC pulled credit for this loan |
| last\_fico\_range\_high | The upper boundary range the borrower’s last FICO pulled belongs to. |
| last\_fico\_range\_low | The lower boundary range the borrower’s last FICO pulled belongs to. |
| last\_pymnt\_amnt | Last total payment amount received |
| last\_pymnt\_d | Last month payment was received |
| loan\_amnt | 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. |
| loan\_status | Current status of the loan |
| max\_bal\_bc | Maximum current balance owed on all revolving accounts |
| member\_id | A unique LC assigned Id for the borrower member. |
| mo\_sin\_old\_il\_acct | Months since oldest bank installment account opened |
| mo\_sin\_old\_rev\_tl\_op | Months since oldest revolving account opened |
| mo\_sin\_rcnt\_rev\_tl\_op | Months since most recent revolving account opened |
| mo\_sin\_rcnt\_tl | Months since most recent account opened |
| mort\_acc | Number of mortgage accounts. |
| mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating |
| mths\_since\_last\_record | The number of months since the last public record. |
| mths\_since\_rcnt\_il | Months since most recent installment accounts opened |
| mths\_since\_recent\_bc | Months since most recent bankcard account opened. |
| mths\_since\_recent\_bc\_dlq | Months since most recent bankcard delinquency |
| mths\_since\_recent\_inq | Months since most recent inquiry. |
| mths\_since\_recent\_revol\_delinq | Months since most recent revolving delinquency. |
| next\_pymnt\_d | Next scheduled payment date |
| num\_accts\_ever\_120\_pd | Number of accounts ever 120 or more days past due |
| num\_actv\_bc\_tl | Number of currently active bankcard accounts |
| num\_actv\_rev\_tl | Number of currently active revolving trades |
| num\_bc\_sats | Number of satisfactory bankcard accounts |
| num\_bc\_tl | Number of bankcard accounts |
| num\_il\_tl | Number of installment accounts |
| num\_op\_rev\_tl | Number of open revolving accounts |
| num\_rev\_accts | Number of revolving accounts |
| num\_rev\_tl\_bal\_gt\_0 | Number of revolving trades with balance >0 |
| num\_sats | Number of satisfactory accounts |
| num\_tl\_120dpd\_2m | Number of accounts currently 120 days past due (updated in past 2 months) |
| num\_tl\_30dpd | Number of accounts currently 30 days past due (updated in past 2 months) |
| num\_tl\_90g\_dpd\_24m | Number of accounts 90 or more days past due in last 24 months |
| num\_tl\_op\_past\_12m | Number of accounts opened in past 12 months |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| open\_acc\_6m | Number of open trades in last 6 months |
| open\_il\_12m | Number of installment accounts opened in past 12 months |
| open\_il\_24m | Number of installment accounts opened in past 24 months |
| open\_act\_il | Number of currently active installment trades |
| open\_rv\_12m | Number of revolving trades opened in past 12 months |
| open\_rv\_24m | Number of revolving trades opened in past 24 months |
| out\_prncp | Remaining outstanding principal for total amount funded |
| out\_prncp\_inv | Remaining outstanding principal for portion of total amount funded by investors |
| pct\_tl\_nvr\_dlq | Percent of trades never delinquent |
| percent\_bc\_gt\_75 | Percentage of all bankcard accounts > 75% of limit. |
| policy\_code | publicly available policy\_code=1 new products not publicly available policy\_code=2 |
| pub\_rec | Number of derogatory public records |
| pub\_rec\_bankruptcies | Number of public record bankruptcies |
| purpose | A category provided by the borrower for the loan request. |
| pymnt\_plan | Indicates if a payment plan has been put in place for the loan |
| recoveries | post charge off gross recovery |
| revol\_bal | Total credit revolving balance |
| revol\_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| sub\_grade | LC assigned loan subgrade |
| tax\_liens | Number of tax liens |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| title | The loan title provided by the borrower |
| tot\_coll\_amt | Total collection amounts ever owed |
| tot\_cur\_bal | Total current balance of all accounts |
| tot\_hi\_cred\_lim | Total high credit/credit limit |
| total\_acc | The total number of credit lines currently in the borrower's credit file |
| total\_bal\_ex\_mort | Total credit balance excluding mortgage |
| total\_bal\_il | Total current balance of all installment accounts |
| total\_bc\_limit | Total bankcard high credit/credit limit |
| total\_cu\_tl | Number of finance trades |
| total\_il\_high\_credit\_limit | Total installment high credit/credit limit |
| total\_pymnt | Payments received to date for total amount funded |
| total\_pymnt\_inv | Payments received to date for the portion of total amount funded by investors |
| total\_rec\_int | Interest received to date |
| total\_rec\_late\_fee | Late fees received to date |
| total\_rec\_prncp | The principal received to date |
| total\_rev\_hi\_lim | Total revolving high credit/credit limit |
| URL | URL for the LC page with listing data. |
| verification\_status | Indicates if income was verified by LC, not verified, or if the income source was verified |
| verified\_status\_joint | Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified |
| zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |
| revol\_bal\_joint | Sum of revolving credit balance of the co-borrowers, net of duplicate balances |
| sec\_app\_fico\_range\_low | FICO range (high) for the secondary applicant |
| sec\_app\_fico\_range\_high | FICO range (low) for the secondary applicant |
| sec\_app\_earliest\_cr\_line | The earliest credit line at the time of application for the secondary applicant |
| sec\_app\_inq\_last\_6mths | Credit inquiries in the last 6 months at The sample time of application for the secondary applicant |
| sec\_app\_mort\_acc | Number of mortgage accounts at the time of application for the secondary applicant |
| sec\_app\_open\_acc | Number of open trades at the time of application for the secondary applicant |
| sec\_app\_revol\_util | The ratio of total current balance to high credit/credit limit for all revolving accounts |
| sec\_app\_open\_act\_il | Number of currently active installment trades at the time of application for the secondary applicant |
| sec\_app\_num\_rev\_accts | Number of revolving accounts at the time of application for the secondary applicant |
| sec\_app\_chargeoff\_within\_12\_mths | Number of charge-offs within the last 12 months at the time of application for the secondary applicant |
| sec\_app\_collections\_12\_mths\_ex\_med | Number of collections within the last 12 months excluding medical collections at the time of application for the secondary applicant |
| sec\_app\_mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating at the time of application for the secondary applicant |
| hardship\_flag | Flags whether or not the borrower is on a hardship plan |
| hardship\_type | Describes the hardship plan offering |
| hardship\_reason | Describes the reason the hardship plan was offered |
| hardship\_status | Describes if the hardship plan is active, pending, canceled, completed, or broken |
| deferral\_term | Amount of months that the borrower is expected to pay less than the contractual monthly payment amount due to a hardship plan |
| hardship\_amount | The interest payment that the borrower has committed to making each month while they are on a hardship plan |
| hardship\_start\_date | The start date of the hardship plan period |
| hardship\_end\_date | The end date of the hardship plan period |
| payment\_plan\_start\_date | The day the first hardship plan payment is due. For example, if a borrower has a hardship plan period of 3 months, the start date is the start of the three-month period in which the borrower is allowed to make interest-only payments. |
| hardship\_length | The number of months the borrower will make smaller payments than normally obligated due to a hardship plan |
| hardship\_dpd | Account days past due as of the hardship plan start date |
| hardship\_loan\_status | Loan Status as of the hardship plan start date |
| orig\_projected\_additional\_accrued\_interest | The original projected additional interest amount that will accrue for the given hardship payment plan as of the Hardship Start Date. This field will be null if the borrower has broken their hardship payment plan. |
| hardship\_payoff\_balance\_amount | The payoff balance amount as of the hardship plan start date |
| hardship\_last\_payment\_amount | The last payment amount as of the hardship plan start date |
| disbursement\_method | The method by which the borrower receives their loan. Possible values are: CASH, DIRECT\_PAY |
| debt\_settlement\_flag | Flags whether or not the borrower, who has charged-off, is working with a debt settlement company. |
| debt\_settlement\_flag\_date | The most recent date that the Debt\_Settlement\_Flag has been set |
| settlement\_status | The status of the borrower’s settlement plan. Possible values are: COMPLETE, ACTIVE, BROKEN, CANCELLED, DENIED, DRAFT |
| settlement\_date | The date that the borrower agrees to the settlement plan |
| settlement\_amount | The loan amount that the borrower has agreed to settle for |
| settlement\_percentage | The settlement amount as a percentage of the payoff balance amount on the loan |
| settlement\_term | The number of months that the borrower will be on the settlement plan |

1. DATA CLEANING TABLES

Table ‑: Columns with more than 60% of missing data

|  |  |  |
| --- | --- | --- |
| **COLUMN NAME VARIABLE** | **Count** | **Percent** |
| id | 2260668 | 100.0 |
| url | 2260668 | 100.0 |
| member\_id | 2260668 | 100.0 |
| orig\_projected\_additional\_accrued\_interest | 2252242 | 99.6 |
| hardship\_dpd | 2250055 | 99.5 |
| hardship\_length | 2250055 | 99.5 |
| hardship\_reason | 2250055 | 99.5 |
| hardship\_status | 2250055 | 99.5 |
| deferral\_term | 2250055 | 99.5 |
| hardship\_amount | 2250055 | 99.5 |
| hardship\_start\_date | 2250055 | 99.5 |
| hardship\_end\_date | 2250055 | 99.5 |
| payment\_plan\_start\_date | 2250055 | 99.5 |
| hardship\_loan\_status | 2250055 | 99.5 |
| hardship\_type | 2250055 | 99.5 |
| hardship\_payoff\_balance\_amount | 2250055 | 99.5 |
| hardship\_last\_payment\_amount | 2250055 | 99.5 |
| settlement\_amount | 2227612 | 98.5 |
| debt\_settlement\_flag\_date | 2227612 | 98.5 |
| settlement\_status | 2227612 | 98.5 |
| settlement\_date | 2227612 | 98.5 |
| settlement\_percentage | 2227612 | 98.5 |
| settlement\_term | 2227612 | 98.5 |
| sec\_app\_mths\_since\_last\_major\_derog | 2224726 | 98.4 |
| sec\_app\_revol\_util | 2154484 | 95.3 |
| revol\_bal\_joint | 2152648 | 95.2 |
| sec\_app\_open\_act\_il | 2152647 | 95.2 |
| sec\_app\_num\_rev\_accts | 2152647 | 95.2 |
| sec\_app\_open\_acc | 2152647 | 95.2 |
| sec\_app\_mort\_acc | 2152647 | 95.2 |
| sec\_app\_inq\_last\_6mths | 2152647 | 95.2 |
| sec\_app\_earliest\_cr\_line | 2152647 | 95.2 |
| sec\_app\_chargeoff\_within\_12\_mths | 2152647 | 95.2 |
| sec\_app\_collections\_12\_mths\_ex\_med | 2152647 | 95.2 |
| verification\_status\_joint | 2144938 | 94.9 |
| dti\_joint | 2139962 | 94.7 |
| annual\_inc\_joint | 2139958 | 94.7 |
| desc | 2134601 | 94.4 |
| mths\_since\_last\_record | 1901512 | 84.1 |
| mths\_since\_recent\_bc\_dlq | 1740967 | 77.0 |
| mths\_since\_last\_major\_derog | 1679893 | 74.3 |
| mths\_since\_recent\_revol\_delinq | 1520309 | 67.3 |

Table ‑: Columns with missing data between 10% and 60%

|  |  |  |
| --- | --- | --- |
| **COLUMNS** | **Count** | **Percent** |
| mths\_since\_last\_delinq | 1158502 | 51.2 |
| next\_pymnt\_d | 1303607 | 57.7 |
| open\_acc\_6m | 866130 | 38.3 |
| open\_act\_il | 866129 | 38.3 |
| open\_il\_12m | 866129 | 38.3 |
| open\_il\_24m | 866129 | 38.3 |
| mths\_since\_rcnt\_il | 909924 | 40.3 |
| total\_bal\_il | 866129 | 38.3 |
| il\_util | 1068850 | 47.3 |
| open\_rv\_12m | 866129 | 38.3 |
| open\_rv\_24m | 866129 | 38.3 |
| max\_bal\_bc | 866129 | 38.3 |
| all\_util | 866348 | 38.3 |
| inq\_fi | 866129 | 38.3 |
| total\_cu\_tl | 866130 | 38.3 |
| inq\_last\_12m | 866130 | 38.3 |
| mths\_since\_recent\_inq | 295435 | 13.1 |
|  |  |  |