

# **INFO6105 Data Science Engineering Methods and Tools**

# *Lending Club -**Interest Rate Analysis and Prediction*

**Group Number** – 11

Rishabh Singh (002743830),

Shubham Singh (002762502)

# Introduction

# Background: **Lending Club** is a financial services company headquartered in [San Francisco, California](https://en.wikipedia.org/wiki/San_Francisco,_California). It was the first peer-to-peer lender to register its offerings as [securities](https://en.wikipedia.org/wiki/Security_(finance)) with the [Securities and Exchange Commission](https://en.wikipedia.org/wiki/U.S._Securities_and_Exchange_Commission) (SEC) and to offer loan trading on a secondary market. At its height, Lending Club was the world's largest peer-to-peer lending platform. Interest income benefits investors. Each loan application is given a grade ranging from A1 to E5 and a matching interest rate based on their credit history. Lending Club's business strategy is established by the fact that investors profit from the interest rate and that both investors and borrowers pay a fee to the company. Lending Club probably isn’t the best option for borrowers with bad credit. That would bring a high interest rate and steep origination fee, meaning you could do better with a different type of loan. Lending Club requirements get high marks, but they might not be for everyone. Here are some pros and cons that might help clarify the advantages and disadvantages:

Pros:

1. Repayment terms of the loan can be stretched up to three years and five years.
2. Lending Club credit score has a minimum acceptance of 600.
3. No hard credit inquiry is needed to check rates which comes handy for the customers

Cons:

1. Lending Club money requires a seven-day period to become available whereas there are other places that provide money around in a day.
2. Lending Club will provide an interest rate, but part of that is an origination fee, which will cut into your loan.
3. You are charged $7 if you pay by check. Late payment fee is 5% or the unpaid instalment amount or $15, whichever is greater.

# Motivation: With the rising popularity of peer-to-peer lending platforms in recent years, investors now have easy access to this alternative investment asset class by lending money to individual borrowers through platforms such as Lending Club, Prosper Marketplace, and Upstart, or to small businesses through Funding Circle. The process starts with borrowers submitting loan applications to the platform, which performs credit reviews and either approves or denies each application. Investors should be able to promptly and independently evaluate the credit risk of many listed loans to invest in loans with lower perceived risks. The platform also uses a proprietary model to determine the interest rate of approved loans based on the creditworthiness of borrowers. Approved loans are then listed on the platform for investor funding. Investors usually want to diversify their portfolio by only investing a small amount, e.g. $25, in each loan. Hence, it is desirable for investors to be able to independently evaluate the credit risk of many listed loans quickly and invest in those with lower perceived risks. This motivates us to develop machine-learned classification and regression models, from which we will choose the model with the highest accuracy and lowest MAPE, which can be used to assess and estimate the interest rate using a historical loan dataset from Lending Club.

# Goal: This project aims to create a machine learning model that can predict the best interest rate for a given set of parameters using transactional loan data from Lending Club.

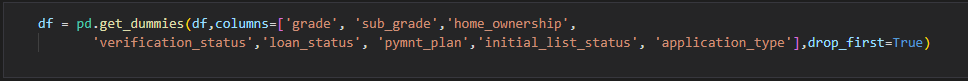
# Methodology

To develop the interest rate classification and regression model we need to follow the below-mentioned machine learning process which includes:

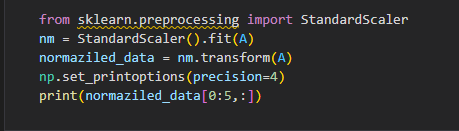
1. Data Collection: In this step, we looked through many online dataset sources to find the most pertinent dataset that matched the project's goal and had a large enough data set to assist the learning and development of the machine learning model. We have used the Kaggle dataset of Lending Club Loan data. The file is a matrix that includes a total of 151 variables and about 2260701 observations. Also, The Total Missing Values in Dataset is 108486249.The Description of the Dataset will be further explained in details in the next section.
2. Data Exploration and Preparation: While exploring the dataset we came to an understanding that we have a large number of records and features with many discrepancies and to fix the noise we have performed data cleaning and pre-processing using python libraries such as NumPy, pandas.

We have followed below steps for Data Cleaning and Pre-processing:

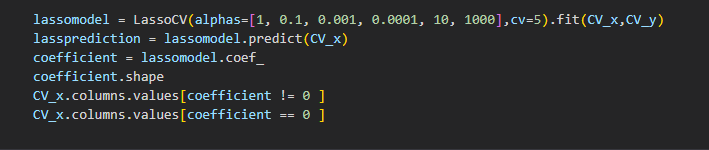
* Initially we have read the csv file into a DataFrame and post understanding the DataFrame we have selected a set of 65 features which were relevant to our Model.
* Since then, we've eliminated features with null/NaN values greater than 50%. In order to do this, we made two empty lists and added the columns with more than 50% of null values to columns\_with\_more\_null and the remaining columns to coulmns\_with\_data, after which we removed the columns with more null values from the DataFrame.
* We have then categorized the DataFrame into Numeric Columns and Categorical Columns by using select\_dtypes function of a DataFrame. If the dtype is either ‘int64’ or ‘float’ then we will store it in numeric\_columns and if dtype is ‘object’ then we will store it in categorical\_columns.
* Getting Dummy values for Categorical Columns.



* Then we have created a new dataset for removing all null values and getting numerical values across all the Columns. Then, we have used Binary Equivalent value for the term Feature. If the term is ‘36 months’ then the value changes to 0 or else it will be 1
* Replacing Strings in emp\_length column with numeric value. For this, we have filled zeros if the emp\_length is empty or else we have just filled it with the number of years mentioned.
* Then we have used Simple Imputer to fill in missing values. We have use fit and Transform functions of SimpleImputer () on our DataFrame to fill in empty/null values.
* We then have approached to perform Normalization on the dataset using Standard Scaler to bring the data set on the same scale. We have split the dataset into 2 data frames out of which 1 needs to be normalized. Later, we performed normalization then summarized the transformed data. We have then calculated the correlation with interest rate after normalization for performing analysis and visualizations.
* Summarizing the transformed data np.set\_printoptions(precision=4)
* Displaying transformed data print(normalizedData[0:5,:]

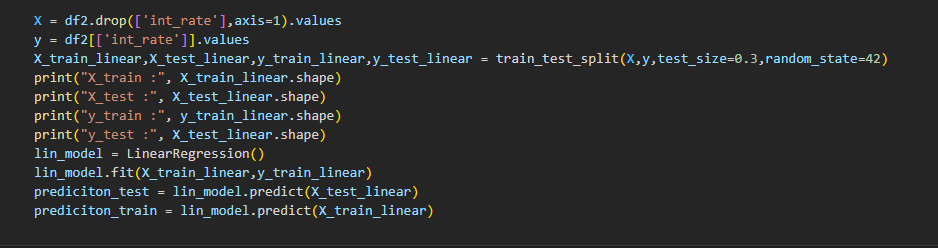


* Lasso (“Least Absolute shrinkage and selection operator”) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. The data is then categorized into three sets and treated differently: train, test and CV. We then split the data into dependent and independent variables to perform lasso CV for variable selection. Thus, we have successfully implemented feature selection by performing dimension reduction using LassoCV. At the end, we have used lassoCV algorithm for feature selection and dropping columns having zero coefficient.

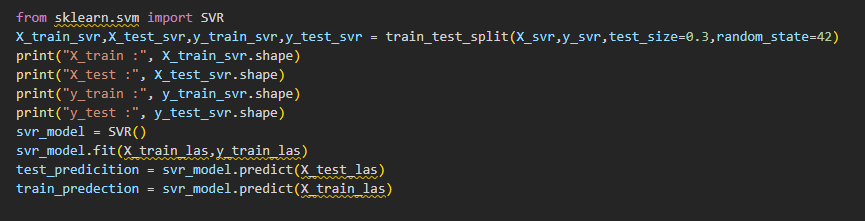


1. Modelling and Evaluation: Once we have a clean dataset with all the necessary features and records with values, we performed a test-train split on it to train various models, compare their accuracies, and use the metrics MAPE (Mean Absolute Percentage Error), MSE (Mean Squared Error), Accuracy Score and R2 Square to choose the model that fits the test dataset the best. The models that we will use are as follows:

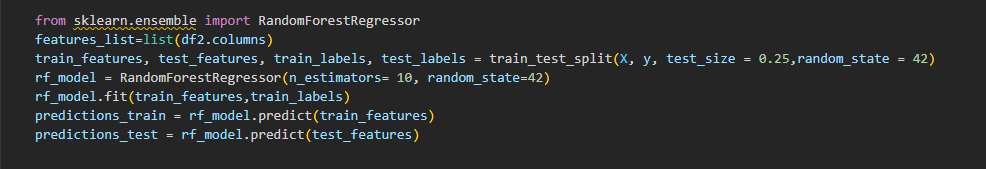
* Linear Regression: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Linear regression is a statistical approach for modelling the relationship between a dependent variable with a given set of independent variables. Thus, linear regression technique finds out a linear relationship between x (input) and y (output). It is used to predict a quantitative response Y from the predictor variable X. It is made with an assumption that there's a linear relationship between X and Y. The equation of the above line is: Y= mx + b Where b is the intercept and m is the slope of the line. So basically, the linear regression algorithm gives us the most optimal value for the intercept and the slope (in two dimensions). The y and x variables remain the same, since they are the data features and cannot be changed. The values that we can control are the intercept (b) and slope (m). There can be multiple straight lines depending upon the values of intercept and slope. Basically, what the linear regression algorithm does is it fits multiple lines on the data points and returns the line that results in the least error. This same concept can be extended to cases where there are more than two variables. This is called multiple linear regression. For instance, consider a scenario where you have to predict the price of the house based upon its area, number of bedrooms, the average income of the people in the area, the age of the house, and so on. In this case, the dependent variable (target variable) is dependent upon several independent variables. We have Performed Linear Regression with a test\_size of 30% and keeping the random state as 42, we have used both test prediction and train prediction for Error and MAPE calculation.



* Support Vector Regression: Secondly we have used Support Vector Regression Algorithm to train our Model. It uses the same idea of SVM but here it tries to predict the real values. This algorithm uses hyperplanes to segregate the data. In case this separation is not possible then it uses the kernel trick where the dimension is increased and then the data points become separable by a hyperplane. We have Performed Support Vector Regression with a test\_size of 30% and keeping the random state as 42, we have used both test prediction and train prediction for Error and MAPE calculation.



* Random Forest: A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees. The random forest algorithm combines multiple algorithms of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks. In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote. The random forest algorithm is not biased, since there are multiple trees and each tree is trained on a subset of data. Basically, the random forest algorithm relies on the power of "the crowd"; therefore, the overall biasedness of the algorithm is reduced. Even if a new data point is introduced in the dataset the overall algorithm is not affected much since new data may impact one tree, but it is very hard for it to impact all the trees. The algorithm works well when you have both categorical and numerical features. It works well when data has missing values, or it has not been scaled well. We have used Random Forest as our Last Algorithm to train-test the model. We have kept the test\_size to 25% and Random State as 42, for Random Forest as well we have calculate both test prediction and train prediction for Error and MAPE calculation.



# Description of Dataset:

Link - <https://www.kaggle.com/datasets/wordsforthewise/lending-club>

These files contain the latest payment information and all the necessary and appropriate loan data for all the loans issued through 2007-2018 which also includes current loan status (Current, Fully Paid, Late, etc.) The file containing loan data through the "present" contains complete loan data for all loans which are issued through the previous completed calendar quarter. There are certain additional features that include credit scores, addresses including zip codes and states, number of finance inquiries, and various other collections. We have only used the csv file which has details of all Accepted Loans between 2007 and 2018 for this project. The file is a matrix that includes a total of 151 variables (with ‘int\_rate’ being our Target label) and about 2260701 observations. Also, The Total Missing Values in Dataset is 108486249.

Specifications of the Dataset -

We have 105 columns pre-processing of data out of which 13 are Categorical Variables and 92 are Numeric Variables.

38 of 92 columns allow NULL values.

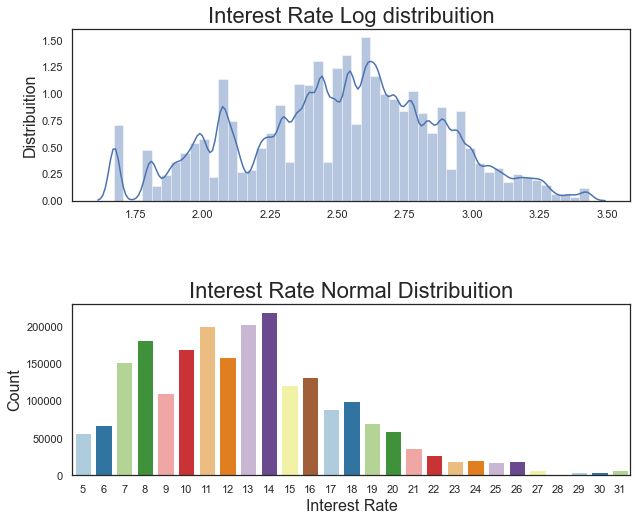
Most Categorical Variables are Descriptions while others are Dates i.e., Days of Week and Months.

Some Numeric Variables are derived based on the values of Other Numeric Variables.

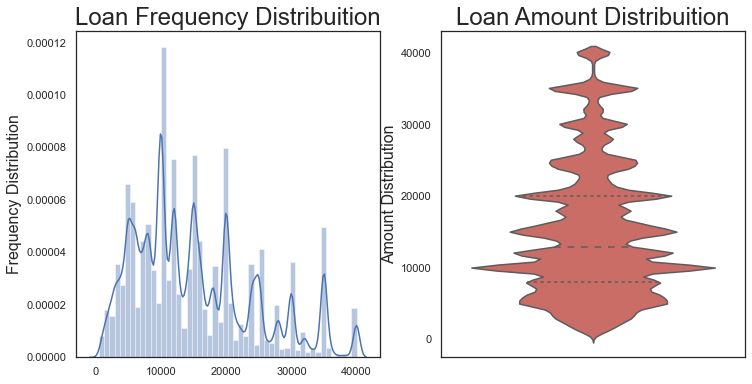
Now, we have visualized the Dataset with a few charts and graphs post analysis the Dataset.

Please see Below Visualizations:

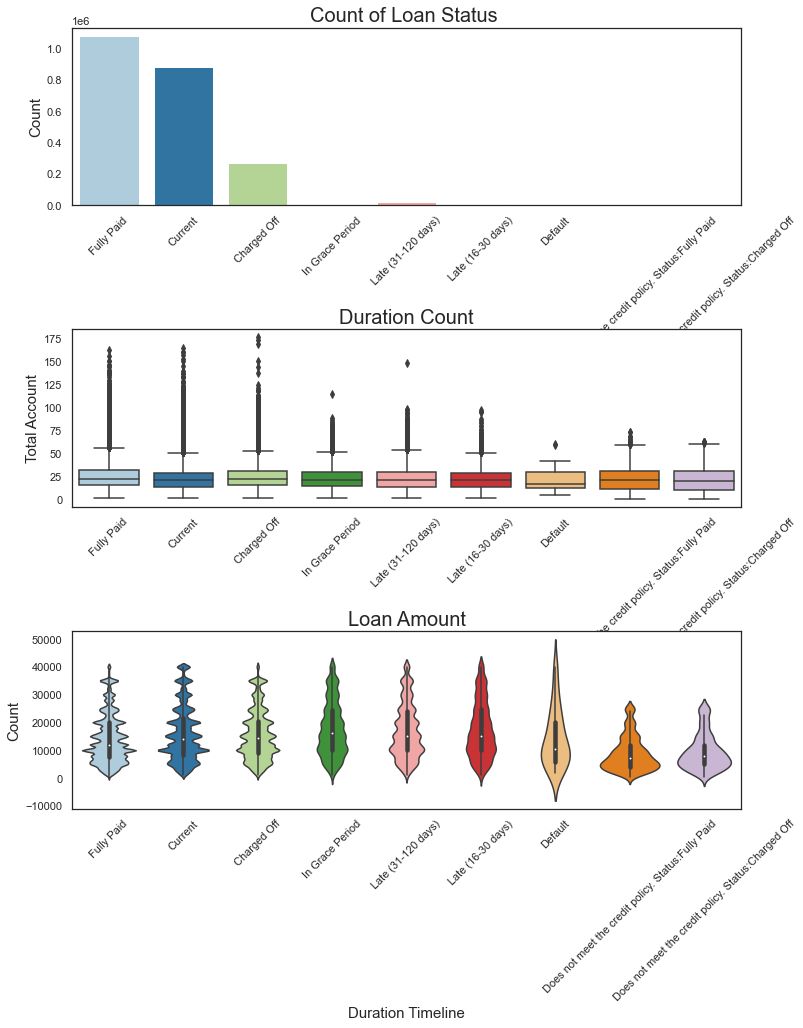
1. **Interest Rate Log Distribution and Normal Distribution**



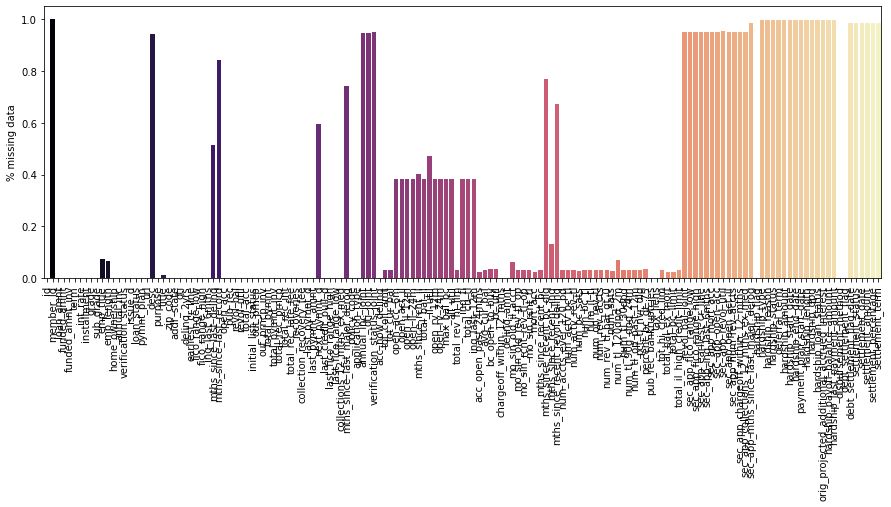
1. **Distribution of Loan Amount**



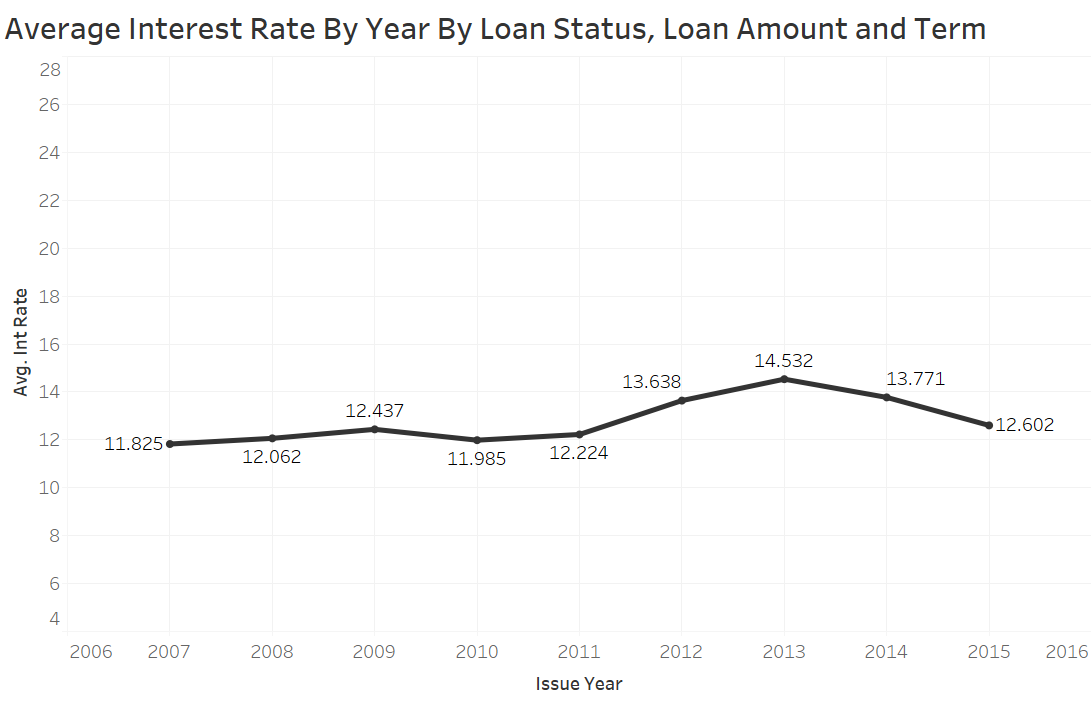
1. **Distribution of Loan Status Count, Duration Count, Loan Amount Count**

****

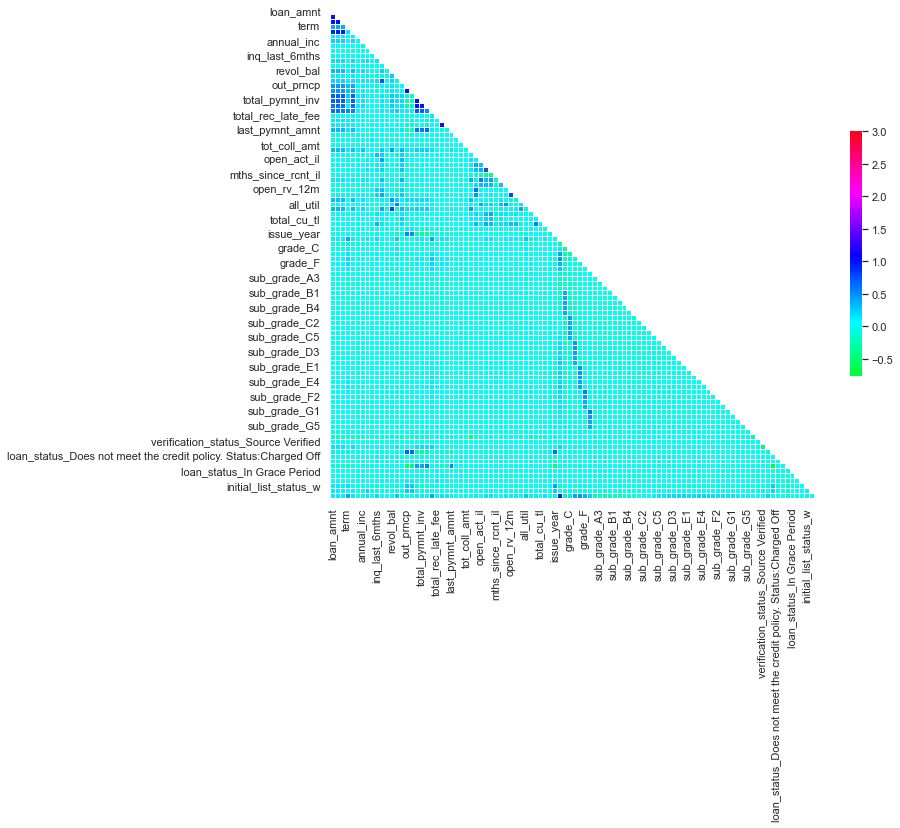
1. **% of Missing Values**

****

1. **Average Interest Rate by Year**

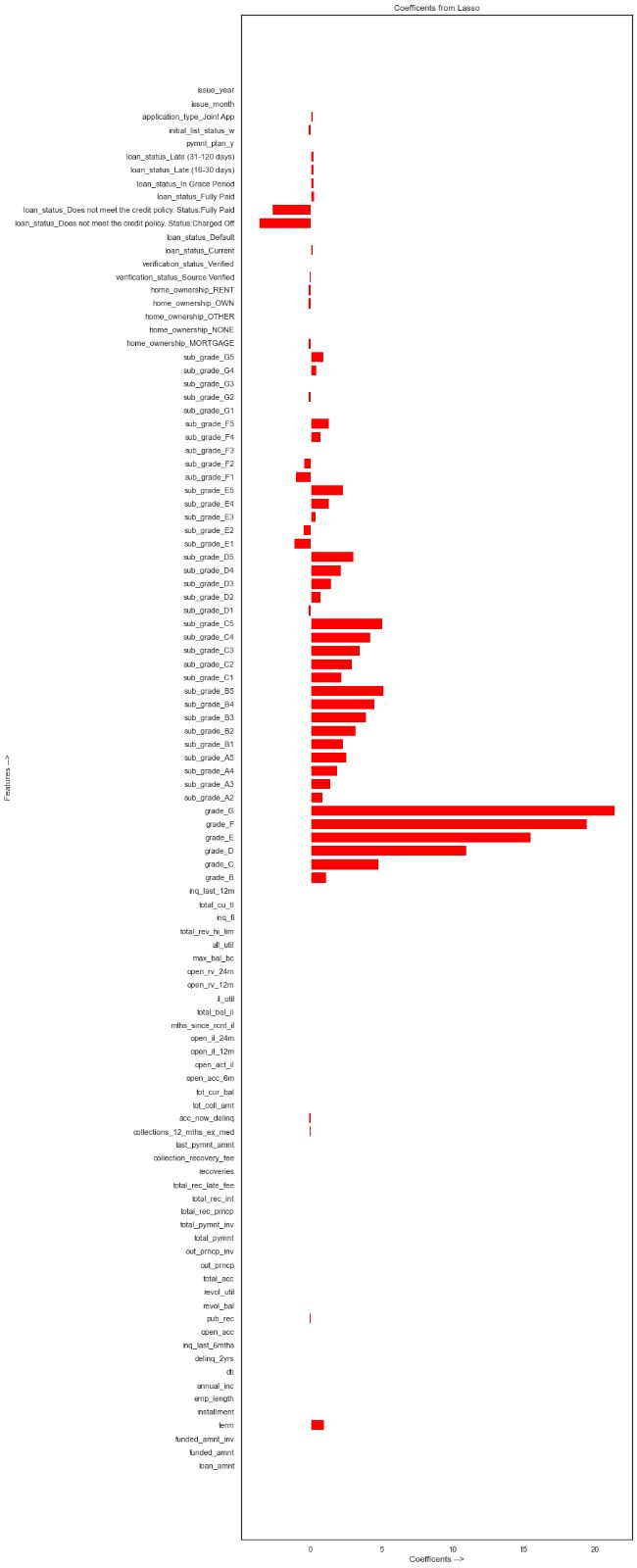
****

1. **Correlation Matrix**

****

# Result and Analysis:

Coefficient score of lassoCV –



We have run the below code for all the three algorithms and calculated Average Error, Accuracy, MAPE and RSquare Score:

error\_train = mean\_absolute\_error(y\_train\_linear,prediction\_train)

error\_test = mean\_absolute\_error(y\_test\_linear,prediction\_test)

mape\_train = 100 \* np.mean(error\_train / y\_train\_linear)

mape\_test = 100 \* np.mean(error\_test / y\_test\_linear)

accuracy\_train = 100 - mape\_train

accuracy\_test = 100 - mape\_test

r2score = r2\_score(y\_test\_linear,prediction\_test)

print('Model Performance')

print('Average Error(Train Data): {:0.4f} of int rate.'.format(np.mean(error\_train)))

print('Average Error(Test Data): {:0.4f} of int rate.'.format(np.mean(error\_test)))

print('Accuracy(Train Data) = {:0.2f}%.'.format(accuracy\_train))

print('Accuracy(Test Data) = {:0.2f}%.'.format(accuracy\_test))

print('Mape(Train Data): {:0.4f} of int rate'.format(mape\_train))

print('Mape(Test Data): {:0.4f} of int rate'.format(mape\_test))

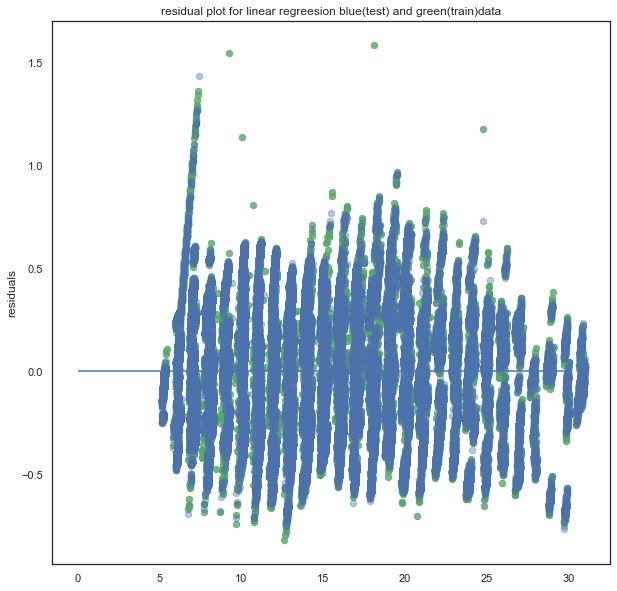
print('Rsquare Score(Test Data): ',r2score)

Now, let’s see the score for Each Algorithm used.

* Implementing Linear Regression:

Text, letter

Description automatically generated



**Fig 1: Residual Diagram for Linear Regression**

* Implementing Support Vector Regression:

Text, letter

Description automatically generated

Chart, scatter chart

Description automatically generated

**Fig 2: Residual Diagram of Support Vector Regression**

* Implementing Random Forest:

Text, letter

Description automatically generated

Chart, scatter chart

Description automatically generated

**Fig 3: Residual Diagram of Random Forest**

# Conclusion

After Running the Dataset on all three above mentioned Algorithm’s, the MAPE for each Model were as follows:

* Linear Regression: 1.8
* Support Vector Regression: 1.5
* Random Forest: 0.3

Amongst all the three models we can see that Random Forest gives us the best predictions. Given, any data from the Dataset, the predicted values can be seen satisfactory as compared to the actual data of interest rates provided by the lending club. So, if we test, our model, against any scenario, be risk averse, risk taking or any other scenario, our model will give a satisfactory predicted estimate of the interest rate. Also, it tells us about the most optimized ML Model that is Random Forest with a MAPE of 0.3 and an Accuracy of 99.72%

# References

https://www.lendingclub.com/public/how-peer-lending-works.action

<http://cs229.stanford.edu/proj2015/199_report.pdf>

<https://www.debt.org/credit/loans/personal/lending-club-review/>

<https://en.wikipedia.org/wiki/LendingClub>

<http://cs229.stanford.edu/proj2018/report/69.pdf>

<https://docplayer.net/39484619-3-algorithms-and-results-where-s.html>

<https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/>

<https://en.wikipedia.org/wiki/Random_forest>

<https://en.wikipedia.org/wiki/Artificial_neural_network>

<https://www.geeksforgeeks.org/ml-linear-regression/>

<https://en.wikipedia.org/wiki/LendingClub>

https://github.com/PriyankaM06091994/Data-Science-Engineering-Methods-and-Tools

https://github.com/jaminaveen/INFO6105\_DataScience