**Springboard Data Science Capstone Project-**

**Predict Interest Rate**

Dan Wang

Jul 11th, 2020

**Introduction:**

Interest rate is a key factor in loan market since it is important for both borrowers and lenders. Borrowers shop around in loan market looking for lower interest rate while lenders focus on tradeoffs between higher interest rate and lower default risk and also being competitive. Knowing the interest rate in advance have many benefits to both parties and definitely have positive impact to loan market. In this project here, I am about to train a model that can make good predictions on the interest rate by providing other key factors. In this way, that both borrower and lenders can know the interest rate in advance.

**Client:**

The clients are borrowers who are interested in getting a loan as well as lenders, investment banks, mortgage companies and other companies engaged in lending businesses. Below are some benefits that borrowers and lender can achieve if they can use a machine learning mode to predict interest rate.

From borrowers’ perspective:

1. Knowing the interest rate in advance can help borrower to compare a variety of offers from lenders and thus, reduce the cost of borrowing.
2. Knowing the interest rate in advance can help borrower in making decision about if it is the right time to get a loan. Predictions showing a lower interest rate in the near future may imply a hold sign for the borrower while predictions showing a higher interest rate indicate a good time to get a loan.
3. Knowing the interest rate in advance can help borrower have a better financial budget plan, like how much repay it is on a monthly basis and how many years it takes to pay off the loan.
4. Knowing the interest rate in advance can help borrowers better understand how to improve their personal information in order to have a desired interest rate, like boosting up their credit score or reducing current loan balance or applying for fewer loan amount, etc.

From lenders’ perspective:

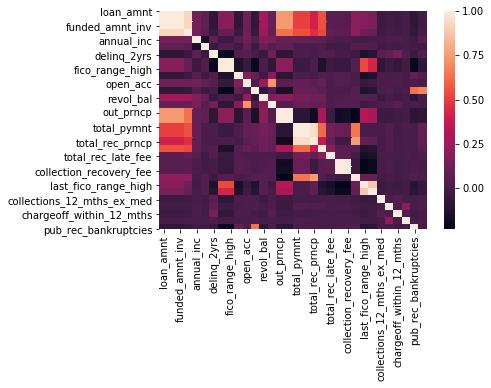
1. Knowing the interest rate in advance can make better financial plans. Lenders can limit of total lending at current lower interest rate and save some for the near future if lenders can predict a higher interest rate in the near future.
2. Knowing the interest rate in advance can have better picture of the tradeoffs between higher interest rate and the default risk. The ultimate goal of leader is to make profits by doing loan business, default loans with higher interest rate is not what they want.
3. Knowing the interest rate in advance that lenders can put himself into the loan market to see if himself is competitive enough to attract borrowers.

**Data Acquision and Cleaning:**

My dataset was acquired online through Kaggle.com. It is a dataset collected from a loan company based in San Francisco, California. Two files are provided online, one accepted loan file and one rejected loan file. The dataset was used for a binary classification prediction whether a loan application will be accepted or rejected based on the borrowers’ personal information. In my project, I will analyze the dataset from a different aspect that is how much the interest rate will be based on all the other facts. The rejected loan file does not include the target feature and will be disregarded, only the accepted loan file will be in use here.

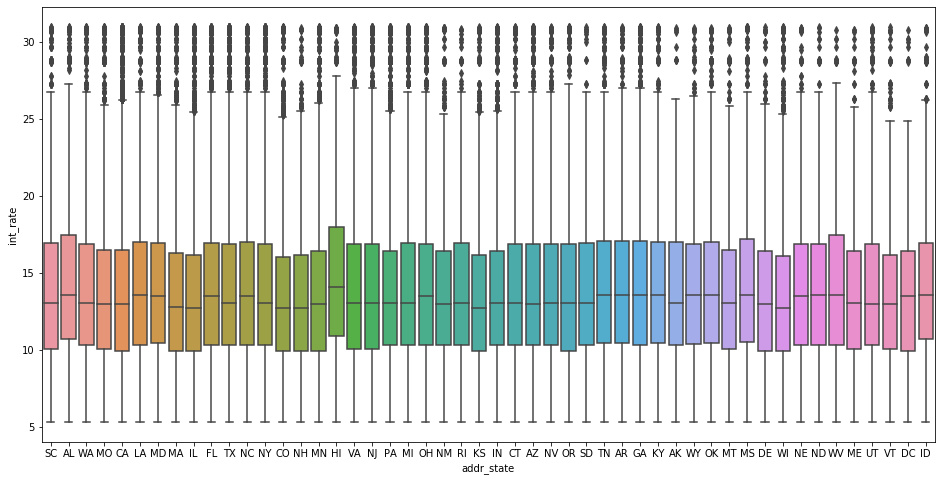
The dataset contains information collected from the year 2013 to 2019, totally more than 2.6 million records and more than 150 features. In order to keep accurate and meaningful dataset, thresholds have been set to eliminate incomplete records and attributes. Attributes have more than 10,000 missing records in a column and records have more 40 missing attributes in a row are dropped from dataset. Thus, the size of the dataset was reduced to 0.3 million records by 55 attributes.

The 55 attributes were composed of numerical attributes, date time attributes and also string attributes. For the numerical attributes, Pearson correlation coefficient was utilized to illustrate if two variables are highly correlated. The threshold was set to absolute number 0.9. Any attribute has absolute value of Pearson correlation coefficient greater than or equal to 0.9 are removed from dataset because of the collinearity.



There are total 4 date time attributes in the dataset, last payment date, last credit pulling date, loan issue date and also earliest credit line date. The last payment date and the last credit pull date are highly correlated with earliest credit line date. It is acceptable to remove first two mentioned features from dataset. Comparing to month and day information, year information of the rest two date time features would be more meaningful since one represents interest rate trend during the past seven years and the other represents earliest credit establishment year. The year information of the loan issue date and earliest credit line date were kept and replace the original two feature values.

The string attributes have to be either converted to numerical values or deleted from dataset. For the attributes, like “id”, “url”, as well as those have only one identical values are removed from dataset since model has low possibility to learn meaningful information from them. For the attributes, like “zip\_code”, “addr\_state”, which represent the origins of loans, are analyzed with box plot to illustrate the relationship between interest rate and their origin states.



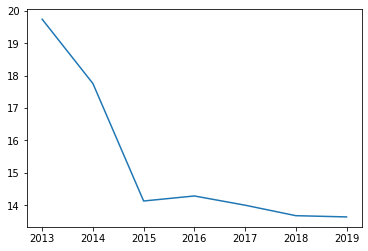
For reading above plot, there is no big difference in interest rate no matter where the origin state is. Based on the fact of that, the zip\_code and addr\_state attributes are also dropped from dataset. The rest of other string attributes are transformed into numerical values according to their distinct values.

The last step in data pre-processing stage is to fill in missing values. The technique that I utilized here is KNN imputer which completes missing values using the mean value from the most similar neighbors found in the dataset. The KNN imputer derives higher accuracy in comparison with filling in missing values with mean of the entire attributes.

After the data pre-processing state, the attributes were further reduced to 39 and the total number of records stays the same as 0.3 million.

**Data Split:**

Before any model was trained and compared to each other, I first need to split my dataset into training and test section. The average interest rate trend is plotted and listed below.



By observing above interest rate trend plot, the interest rate dramatically dropped from 19.74 % to 13.63 during the past seven years. In order to train my model properly, I will use 2015 and 2016 as testing dataset and use the rest of years as training dataset.

With all the interest rate known in the dataset, I will use supervised learning regression algorithms to build prediction models. 70% of the data (year 2013,2014,2017,2018 and 2019) will be selected to train models and the rest 30% (year 2015 and 2016) will be used to evaluate the performance of models. Models selected in the project here are linear regression, lasso regression and random forest regression. Linear regression is implemented first in order to give me a general idea of how well all independent features doing in predicting the target variable. Lasso regression comes after linear regression is used to optimize linear regression by reducing model complexity and prevent over-fitting. Random forest regression is also trained just in case dataset is in a non-linear shape and the linear regression and lasso regression cannot capture the non-linear features.

**Evaluation Metrics:**

The task in this project is a regression task, the output of this prediction will be continuous value in a given range. Three evaluation metrics will be utilized here to access the performance of each model, the mean squared error, R squared value and the adjusted R squared value.

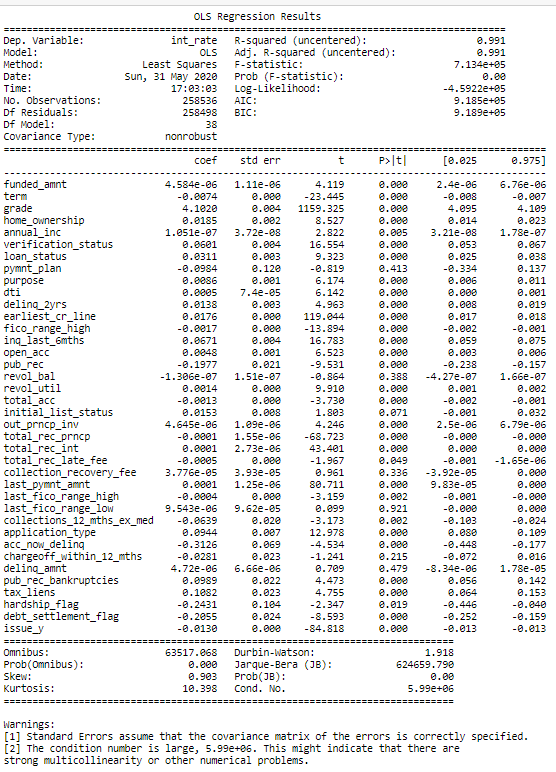
Mean Squared Error: It is the average of the squared difference between the predicted value and the actual value. As it squares the error, the lower the mean squared error value, the better the model performs.

R-Squared: It measures how well the fitted regression line fit into the dataset. R-squared value always ranges between 0% and 100%. 0% means the regression line does not explain the target variable at all while the 100% represents the model fits the dataset without any error. Thus, the larger the R squared value, the better the regression model fits into the dataset.

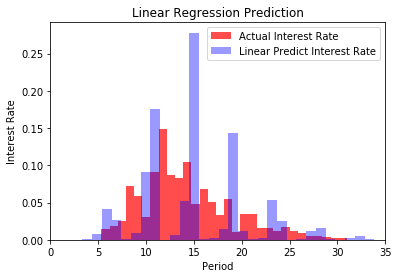
Adjusted-R-Squared: R-squared value keep increasing by adding additional features into model, however, the adjusted R squared increases only when the new feature improves the model more than expected by chance only. The adjusted R squared value decreases when the new feature does not improve the moment sufficient.

**Linear Regression:**

At the very beginning, all the 39 features in the training dataset are passed into linear regression model to train the model and then a summary has been generated. Any features having P-value greater than 0.05 are removed from training and testing dataset since higher P-value suggests changing in that specific feature will not associated with changes in the target variable. In other words, features having P-value greater than 0.05 are non-essential features in predicting target variable.



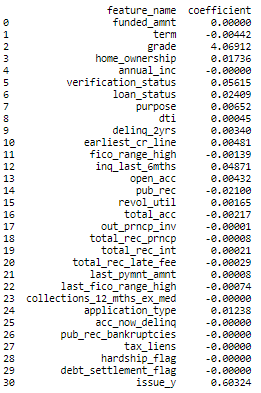
After removing non-essential features from dataset, the training dataset was reduced to 32 features and same number of records. Putting the reduced size dataset in the model again and the test dataset to evaluate the model, it gives me the mean squared error, the R-squared value and the adjusted R squared values are 2.6593, 0.8881 and 0.8882 which indicates the linear regression model did a good job in predicting the target variable. Below is the illustration of linear model prediction, the red bar represents the actual interest rate while the purple bars represent the predicted values.



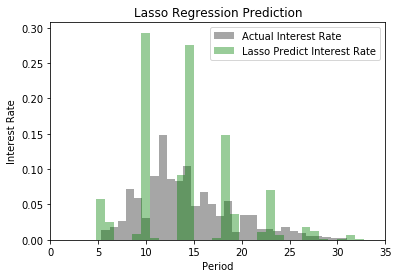
**Lasso Regression:**

Lasso Regression is utilized to optimize linear regression by adding penalties to linear regression. Lasso Regression performs L1 regularization which can reduce some features’ coefficient to zero and get eliminated from the model. Hyper parameter alpha determines the degree of the penalty to the model. Higher alpha will reduce all features’ coefficient to zero while lower alpha does not provide sufficient penalization. A proper alpha value is a key successor of launching lasso regression. The grid search cross validation technique is utilized to select the best alpha for the model and has decided the alpha score to be 0.01.

The reduced size training dataset was passed to lasso regression model with pre-determined alpha value (0.01), a coefficient report with its feature name has been printed to further reducing the size of both the training and testing dataset. For those features have coefficient value of 0 can be safely removed from training and testing dataset to simplify the model and also to avoid overfitting.



There are total 11 features having 0 coefficient that can be removed from dataset which further reduced dataset to be the size of 21 features and 0.3 million records. Passing this dataset to lasso regression model give me better mean squared error, r-squared values and also adjusted R squared value which are 2.1617, 0.9091 and 0.9092. These three evaluations metric shows the lasso regression did a better job than simple linear regression model. Below is the plot showing lasso regression prediction, the gray bars are the actual interest rate while the green bars represents the lasso regression prediction.



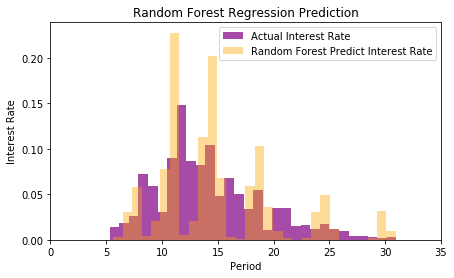
**Random Forest Regression:**

While linear regression mode is used to deal with dataset that are in linear shape and lasso regression is used to optimize linear regression, random forest regression is implemented to deal with chances that the dataset is in non-linear shape.

Just like lasso regression, random forest regressor has hyper parameters that can be tuned in order to better fit into the dataset. I decided to tune the most important two parameters, ‘n\_estimators and max\_depth’, in grid search cross validation technique to find the best parameter for this dataset. Due the constraint of time and capacity, I have to separate the process into two phases. At phase one, I put two possible parameters in each parameter and get one pair of best parameter. Then, I put another two possible parameters in each parameter again and get another pair of best parameter. At phase two, I put the two pairs of best parameters into grid search cross validation again to make a final selection. The best number for the above two parameters are 130 and 75.

Unlike the linear regression and the lasso regression, random forest regression does not require feature selection. It automatically ranks each feature by the purity of nodes. Features that can separate dataset with higher purity nodes will have higher rank while features that separate dataset with least purity nodes will have lower rank.

The original 39 features dataset was passed into the model to train the model and the test dataset was used to evaluate the model. The mean squared error, the R squared value and the adjusted R squared values are calculated as 2.2306, 0.9062 and 0.9062. Below is the plot of random forest regression prediction versus actual interest rate. The purple bars represent the actual interest rate while the yellow bars represent the random forest regression prediction.



**Model Comparison:**

There are total three models implemented in this project to make predictions on the interest rate. They are linear regression, lasso regression and random forest regression. Below is a summary of how these models performs in predicting the interest rate of the year 2015 and 2016.



From reading above summary, all three models are doing great jobs in making prediction on the interest rate of the year 2015 and 2016. Among all three models, my preference will be lasso regression since it uses smaller dataset than linear regression and random forest regression, but better prediction values and less time consuming.