COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

PREDICTING TENURE OF LOANS ISSUED BY WORLD BANK

IEOR 4650 - Business Analytics Project

Spring 2018

Kushagra Agarwal, Subham Kedia, Sumeet Kotaria, Mansi Sheth, Salil Vaidya

1. Overview

The World Bank was established in 1944 with the aim to help rebuild Europe and Japan after the 2nd world war. Since then it has been giving out loans to various public and private institutions for projects to improve and develop public infrastructure. The International Development association(IDA) within the World Bank is the primarily concern when it comes to lending to poor countries.

The team decided to work on a data set available on the world bank data platform. The data set lists various loans given by the IDA to public and private entities across the globe. The data set has around 8,000 instances and 30 features. It records the attributes such as start and end period of the loan, currency of loan disbursed, loan amount, current status of the loan, Project details for which the loan was assigned etc.

The main objective of the project is to propose a solution by which the world bank can predict the number of years in which a country will repay the loan. Our team has tried to used various Machine Learning techniques taught in the class to predict the duration of repayment of a loan based on various economic indicators present in our dataset.

2. Scope of the project

Due to the erraticity of policy making from US and crisis in several European countries, inflowing funds to the world bank are under pressure. The top officials need to understand the duration of the new loans disbursed so as to have an idea of cash flow in the upcoming years.

The group plans to get insights on statement of loans by the IDA, what type of projects are most common when it comes to disbursing loans and finding correlations in our data to predict the duration of a new loan. These insights can help leaders to assess how a particular region has performed and chart out policies for further disbursement of loans.

3. Dataset Information

The main dataset used in this project was obtained from the World Bank Group Finance Open Data website. IDA Statement of Credit and Grants contains data of public and publicly guaranteed debt extended by the World Bank Group. It has 8174 instances with details of loan extended to countries from the year 1961 to year 2016.

To further enrich our data, we have used a supplementary data source called “World Development Indicator” from the World Bank Group Finance Open Data website. This dataset has several development indicators for each country. From these, we have selected 39 economic factors like GDP, GNI, Foreign Direct Investment and several others for different countries.

We have combined these two datasets to form a merged dataset where we have structured economic factors according to countries for different years. These factors used as independent variable to decide our Y i.e. Response variable which in this case is “Duration taken to repay the loan”.

We have created a new “Duration” feature for the data by taking the difference in date values of “Last Repayment Date” suggested by the World Bank and the “Board Approval Date”.

4. Exploratory Data Analysis

We performed exploratory data analysis to analyze the loans data and draw insights related to loans.

- Geographical distribution of the loans : India has taken the most number of loans from World Bank followed by Pakistan, Bangladesh and China.
- Amount of Loan : India has borrowed the maximum amount (USD 52 Billion) followed by Bangladesh (USD 28 Billion) and Pakistan (USD 23 Billion).
- Regional distribution of the loans : Most number of loans were given to countries in Africa followed by South Asia.
- Number of loans over time : The number of loans disbursed over time has been increasing since 1960.
- Loan duration over time : The average duration of loans is decreasing over time.
- Service Charge Rate over time : The average Service Charge Rate i.e. Interest Rate is increasing over time, which means that the World Bank is charging more for loans than it used to.
- Duration by Service Charge Rate : The duration of loans for a particular service charge was analyzed for top borrowers. No specific trend was observed.

5. Analysis

We used various techniques and models that we learnt in class in order to gain valuable insights and to achieve our goal to predict the duration of new loans based on the country and several factors affecting that country's economy. First of all, we cleaned our dataset and developed a correlation plot to understand the linear relationship amongst the different features.

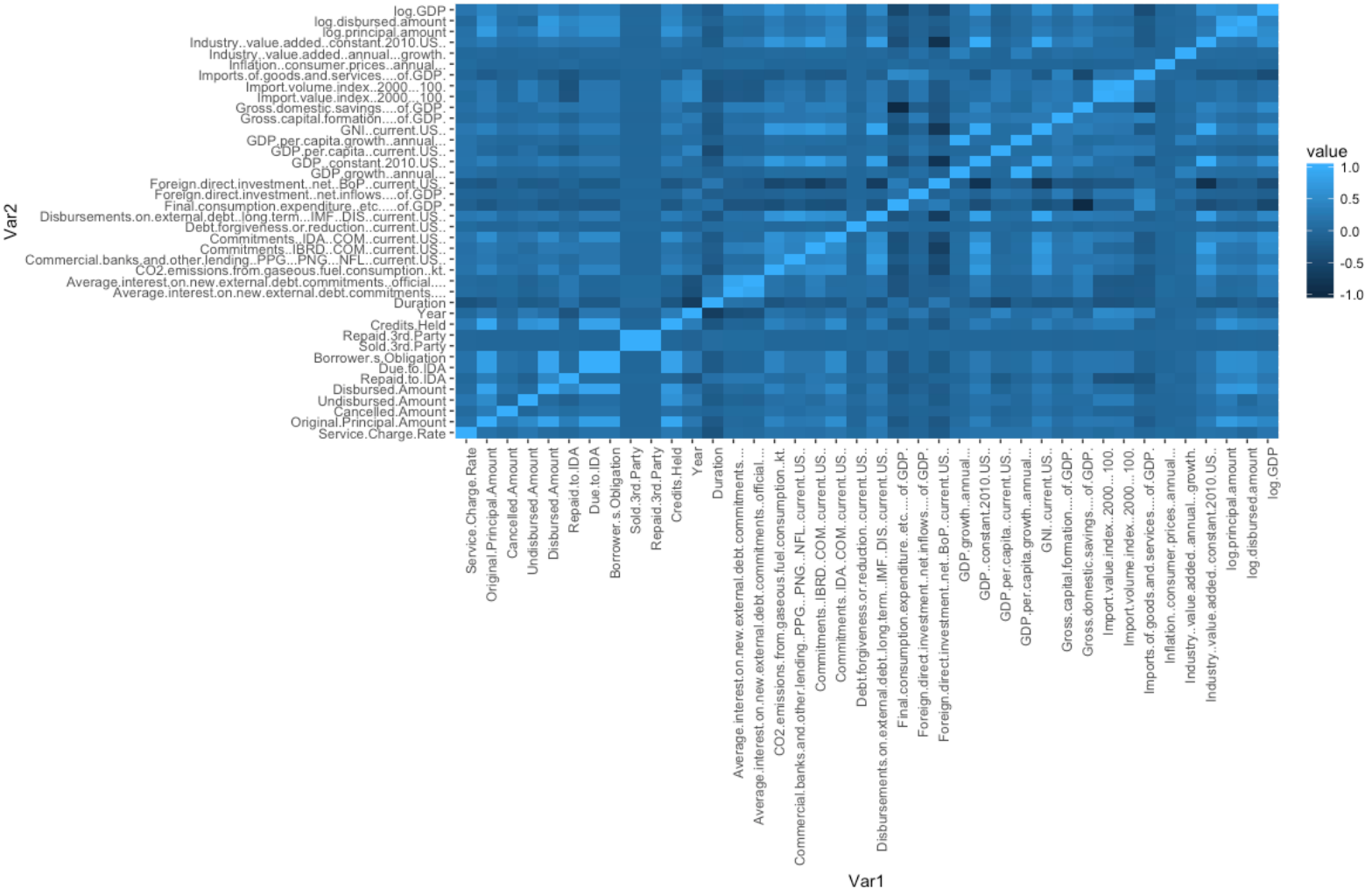

Figure 1: Correlation plot of all features

Then, we fit different models on the dataset as follows -

5.1 Linear Regression Results:

The results show that the features that affect the duration of the loans most are: log.GDP, GDP per capita and Inflation of consumer prices.

Looking at the summary of the linear model, our adjusted R-square is 0.3276 which is not a very plausible number. Hence, the linear model doesn't give us a good model to predict the loan tenure.

5.2 Lasso Regression Results:

Using 75% of the dataset as the training data and applying 5-fold cross validation on our training set, we performed lasso regression and found the best lamda to be 0.001 with a mean square error of 15.00078.

Plotting the histogram of the residuals of the lasso regression, we get the following plot:

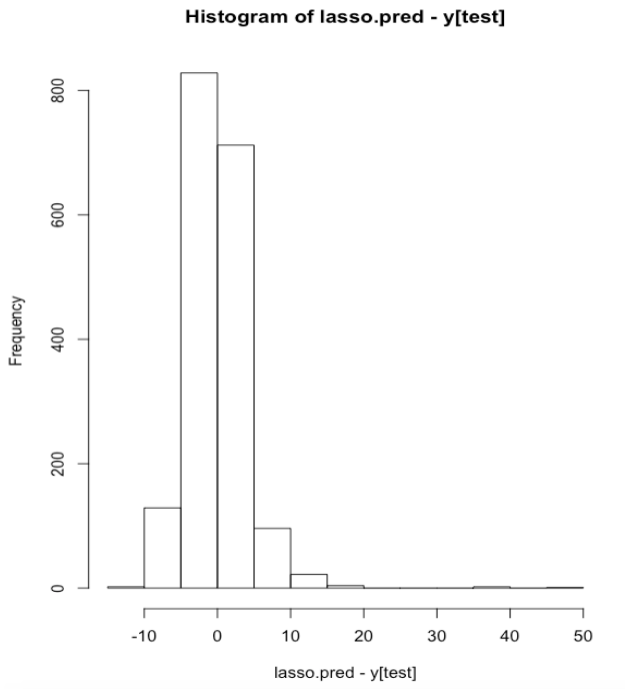

Figure 2: Histogram of residuals for Lasso Regression

Observing the histogram, we can see that for lasso regression our predicted value of response variable varies from the actual value mostly by values in between 0 and -5.

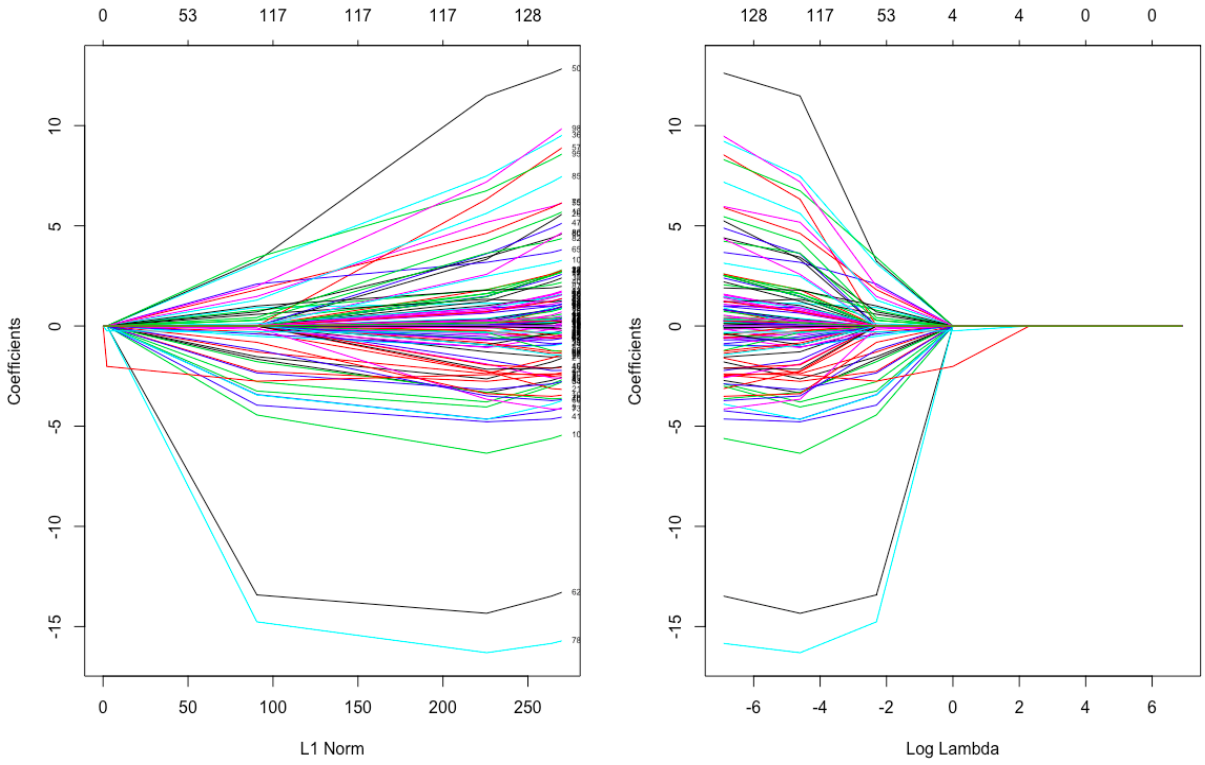

Figure 3: Variance explained by components

The above plot shows the variation of coefficients for different features with lambda and L1 norm for lasso regression.

5.3 Ridge Regression Results:

Using 5-fold cross validation on the training set, we performed ridge regression which resulted in a best lambda of 0.01 and a mean square error of 14.99921 years.

Plotting the histogram of the residuals of the ridge regression, we get the following plot:

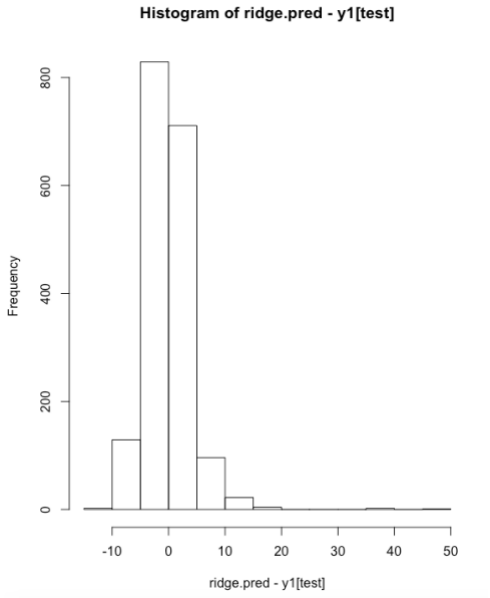

Figure 4: Histogram of residuals for Ridge Regression

Observing the histogram, we can see that our error for predicted values of loan duration lies highly in between -5 and 5.

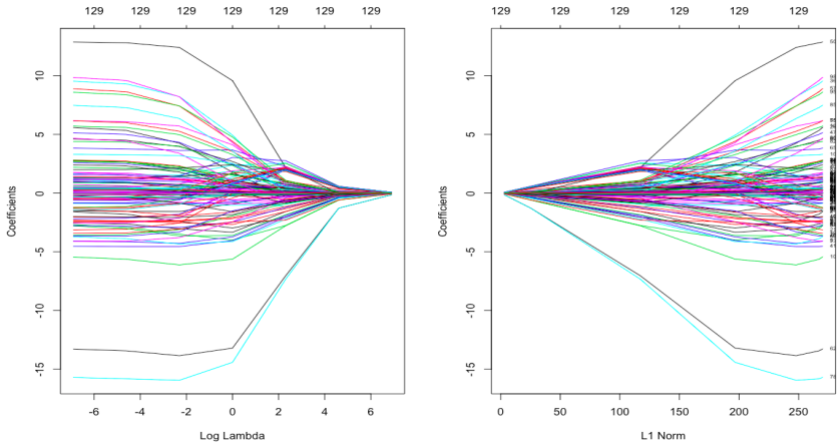

Figure 5: Variance explained by components

The above plot shows the variation of coefficients for different features with lambda and L1 norm for ridge regression.

5.4 K-Nearest Neighbor Results:

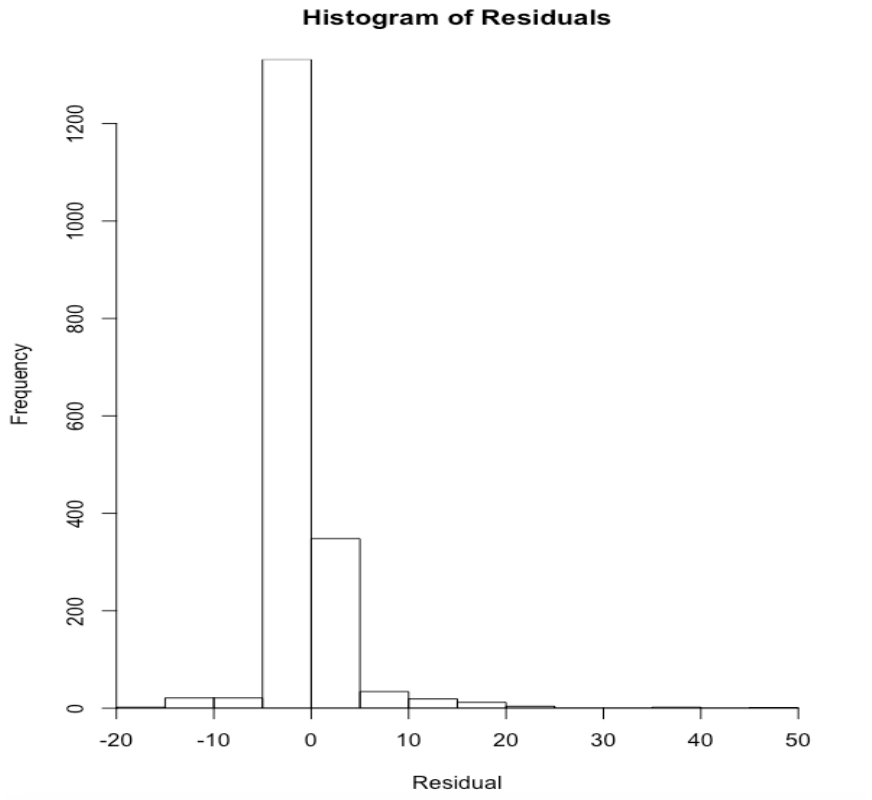

Figure 6: Histogram of residuals for k-NN

Using cross validation on the training data, we performed the k-Nearest Neighbors algorithm with best k being 13 and a RMSE of 3.6 years. The error rate with the knn model is 28%.

5.5 Principal Component Analysis Results:

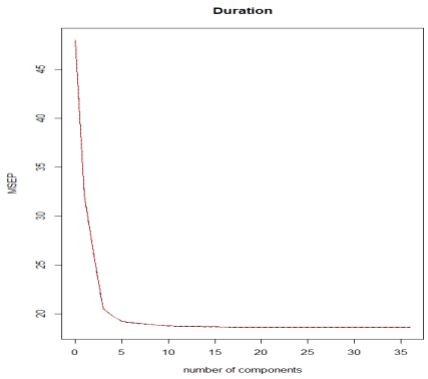

Figure 7: Histogram of residuals for Ridge Regression

We have used 5-fold Cross-Validation on the training data and we predicted the duration on the test data.

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps	10 comps	11 comps	12 comps
X	19.12	32.33	37.05	42.66	46.89	51.47	54.20	58.48	60.67	63.68	66.14	69.67
Duration	33.35	46.56	57.78	59.38	60.43	60.73	60.94	61.05	61.30	61.45	61.53	61.56
	13 comps	14 comps	15 comps	16 comps	17 comps	18 comps	19 comps	20 comps	21 comps	22 comps	23 comps	
X	72.93	74.74	75.77	78.21	80.06	81.97	83.35	83.99	85.91	87.26	88.76	
Duration	61.59	61.64	61.70	61.72	61.74	61.76	61.78	61.79	61.80	61.80	61.80	
	24 comps	25 comps	26 comps	27 comps	28 comps	29 comps	30 comps	31 comps	32 comps	33 comps	34 comps	
X	90.21	91.47	93.16	93.88	95.3	96.69	98.72	99.43	100.0	100.0	100.0	
Duration	61.80	61.80	61.80	61.80	61.8	61.80	61.80	61.80	61.8	61.8	61.8	
	35 comps	36 comps										
X	100.0	100.0										
Duration	61.8	61.8										

Looking at the above results we can see the 60.73% of the variance is explained by 6 Components.

We have Mean-Squared Error of 18.22076 on the predicted value.

6. Conclusion

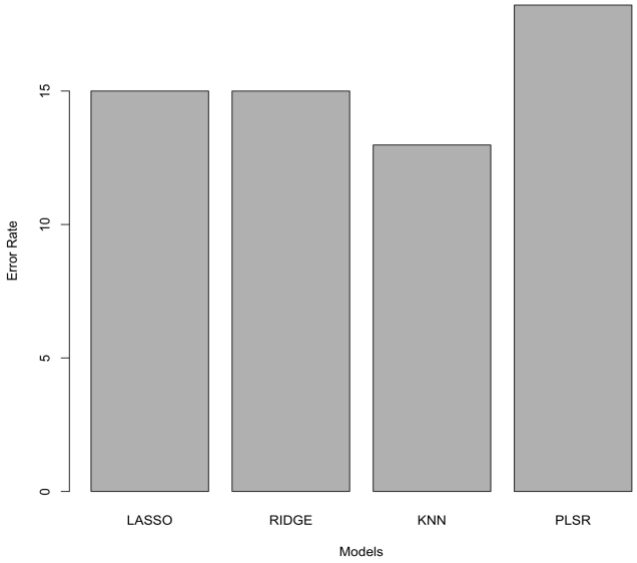

Figure 8: Error rate for different models

After working on models like LASSO regression, RIDGE regression, K-NN, PSLR on the dataset and plotting the respective Error rate, we observe the lowest Error rate for K-NN and highest error rate for PSLR

From our EDA we observed that duration of loan has been decreasing since the establishment of world bank. Also the duration of the loan is heavily depended on the country asking for the loan.

The data-set included 47 different social economic factors. Many of these features were sparsely populated and hence the final models were run on 38 features. Best performing model was K-NN model with k=13. It gave us the root mean squared error of around 3.6 years with a 28% error in prediction. The model will help world bank to predict duration of any new approved loan. This in turn will help the world bank to plan their cash flow.

A major limitation to this methodology is that it does not take into account several political factors which might impact the economic stability of these borrowing countries.