Predicting a Country's S&P rating

IME672A - Project Report

Group No 5

Group members :-

Amit Tiwari (170094)
Bharat Swami (170207)
Sanjay Singh Rana (170627)
Saroj Anjesh (170634)

Professor
Faiz Hamid
IME Department
IIT Kanpur

Introduction : -

There are many financial service companies which release the credit rating of all countries on a specific time interval. From those top three major companies are Standard and Poor Global Rating, Moody's and Fitch Rating. These companies have been releasing country ratings since 1949. And this project involves predicting credit rating based on past credit ratings using machine learning models. These credit ratings of a country are also known as Sovereign credit ratings.

Original Data Sheet

https://docs.google.com/spreadsheets/d/1WxaujcjDd71XbK7xX9hBrzDXd4UfxlNowNvDc-xC3J0/edit#gid=19

ISO code	Country	S&P Rating	S&P Outlook	Moody's rating	Moody's Outlook	Fitch Rating	Fitch Outlook
IL	Israel	A+	STA	A1	STA	A	STA
PL	Poland	A-	STA	A2	STA	A-	POS
MY	Malaysia	A-	STA	A3	STA	A-	STA
SI	Slovenia	A-	STA	Baa2	NEG	A-	NEG
IT	Italy	BBB+	NEG	Baa2	NEG	A-	NEG
CZ	Czech Republic	AA-	STA	A1	STA	A+	STA
EE	Estonia	AA-	STA	A1	STA	A+	STA
SK	Slovakia	A	STA	A2	NEG	A+	STA
MT	Malta	BBB+	STA	A3	NEG	A+	STA
CL	Chile	AA-	STA	Aa3	STA	A+	STA
CN	China	AA-	STA	Aa3	POS	A+	STA
JP	Japan	AA-	NEG	Aa3	STA	A+	NEG
SY	Taiwan	AA-	STA	Aa3	STA	A+	STA
KW	Kuwait	AA	STA	Aa2	STA	AA	STA
BM	Bermuda	AA-	STA	Aa2	STA	AA	STA

Data Understanding

We were given ratings of 135 countries on 11 different dates i.e from April 2010 to March 2013. This data includes 1822 rows and 9 attributes. With the help of pandas python library we dropped the columns of moody and fitch because that was not in our problem statement. After dropping them we are left with ISO Code, Country name, S&P rating, S&P Outlook and Date.

ISO code - Unique code give to each country
Country name - name of the country
S&P Rating - credit rating that varies from AAA to D
S&P outlook - status of rating

S & P credit rating is an ordinal variable. We converted this rating to numeric rating which ranges from 22 to 1, where 22 is an extremely good and top notch rating i.e AAA and 1 is very poor rating and risky for investment i.e D. Shown in table no. 3

The credit rating AAA means that a particular country will surely repay the debt if investment is done in it. For any country with a BBB rating in 2009 means, that particular country will default and won't pay the debt has a chance of 0.55%. Look at Table no. 1, S & P default rate chart.

Year	AAA	AA	A	BBB	BB	B	CCC/C
2009	0.00	0.00	0.22	0.55	0.75	11.01	49.46
2010	0.00	0.00	0.00	0.00	0.58	0.87	22.73
2011	0.00	0.00	0.00	0.07	0.00	1.68	16.42
2012	0.00	0.00	0.00	0.00	0.30	1.58	27.52
2013	0.00	0.00	0.00	0.00	0.10	1.65	24.67
2014	0.00	0.00	0.00	0.00	0.00	0.78	17.51
2015	0.00	0.00	0.00	0.00	0.16	2.41	26.67
2016	0.00	0.00	0.00	0.06	0.47	3.75	33.33

Table 1 - S & P default chart

Attribute 1

Rating	Type	Type
AAA	Investment	Extremely strong
AA+, AA, AA-	Investment	Very Strong
BBB+, BBB, BBB-	Investment	Strong
BB+, BB, BB-	Speculative	Adequate
B+, B, B-	Speculative	Faces major future uncertainties
CCC	Speculative	Currently vulnerable
CC	Speculative	Currently highly vulnerable
C	Speculative	Has filed bankruptcy petition
D	Speculative	In default

Table 2 - Credit Rating and its type

Rating	Numeric Rating	Type
AAA	22	Top Notch for investment
AA+	21	Invest UNDER OBSERVATION
'	'	'
'	'	'
CC	3	Bad for investment
C	2	Bad for investment
D	1	Bad for investment

Table 3 - Numeric rating

Attribute 2

Outlook type	Interpretation
Positive	Rating may be raised next month
Negative	Rating may be lowered next month
Stable	Rating is not likely to change

Table 4 - S & P Outlook

S & P outlook gives additional information about rating. Positive means rating may rise next month, Negative means rating may go down next month and stable means rating will not likely to change next month.

Data Preprocessing

Filling the missing values.

To fill the missing values first we convert the S & P rating to integer scale rating with the help of mapping. Then characterize each rating to quality subgroups which helps us to check whether a rating changes to the same group or jumps to another group.

Then for each country we calculate the number of different past ratings and there count also the number of different outlook and there counts during then we are dropping NaN values (which we want to fill). Now we find the mean rating for a given country and if it's integer type then we replace all NaN values for that particular country to that value and same for outlook.

But if its floating point number (i.e, its ceil value is not equal to floor value) we check the characterization mapping and if both ceil_charaterization and floor characterization value matches we replace NaN rating values with mode of values and the same for outlook.

If characterization for ceil and floor value is not same then we take the mode of outlook and if mode if “Positive” we fill NaN rating values with ceil value of mean rating, if it's “Negative” we replace nan rating values for floor value of mean rating and for Stable outlook we set rating to mode of ratings . And same for outlook.

Out of the 8 attributes, the value of the **overall** attribute was to be predicted. It can be seen from the dataset that there should be a total 1822 values for each attribute. However, there were many attributes with missing values. All the missing values needed to be filled which will clean the dataset and help in better understanding of the data. In the next step, the total number of the null values for each attribute was calculated. It can be observed from the data set that there were a few rows that had missing values for 48 rows and had ‘62’ as the overall rating. To avoid duplication, these rows were dropped, instead of trying to fill the missing values.

Among the non-numeric, attributes, namely, **Preferred foot**, **work rate** and **the player position** were retained to plot them against the other attributes to see if they contributed anything significant to the overall rating of the player.

Next, to find out the relation among the numeric variables, a correlation matrix was plotted and based on that, a number of attributes which had high correlation among themselves, were clubbed to form a single attribute, while those that were clubbed were dropped. The following attributes were clubbed by taking the mean of every attribute corresponding to one entry to form a new single attribute, and the correlation matrix with new attributes were plotted

Problem faced and solution approach

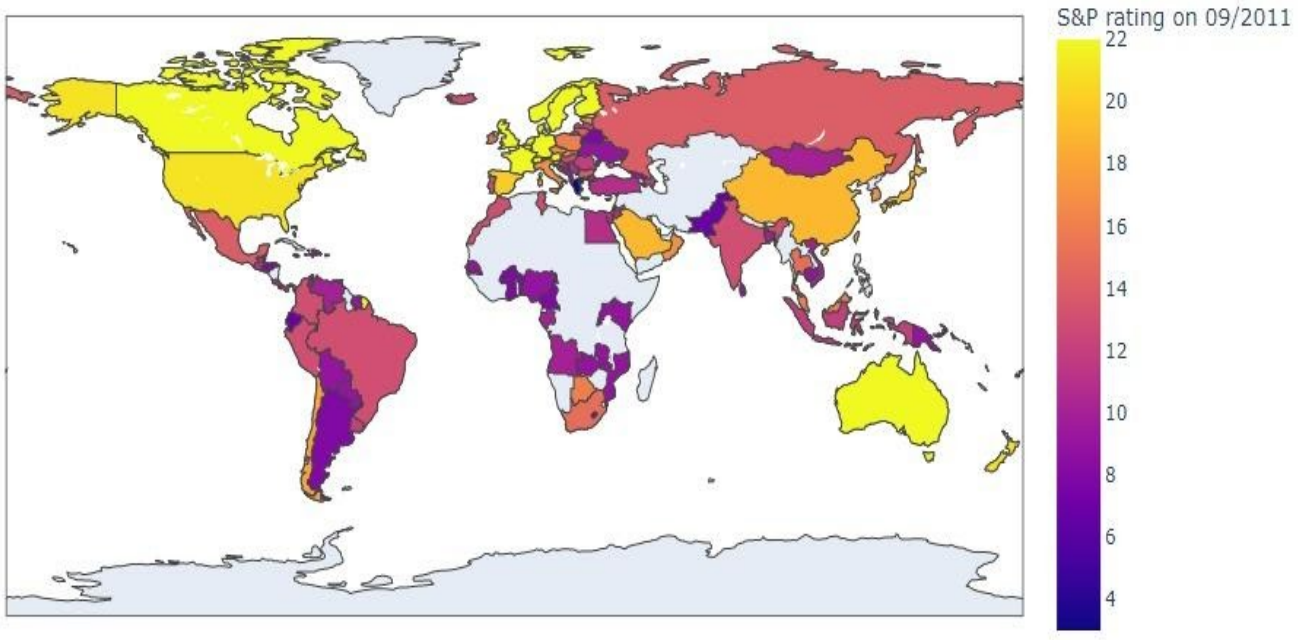
After we finished the data understanding and data preprocessing we found that data provided by the professor was very low in number for applying any kind of machine learning model. So we used a web scraping technique to get more data from a website mentioned below. And we used that data for our further analysis. “fresh-data-1.xlsx”

Example - India
https://tradingeconomics.com/india/rating

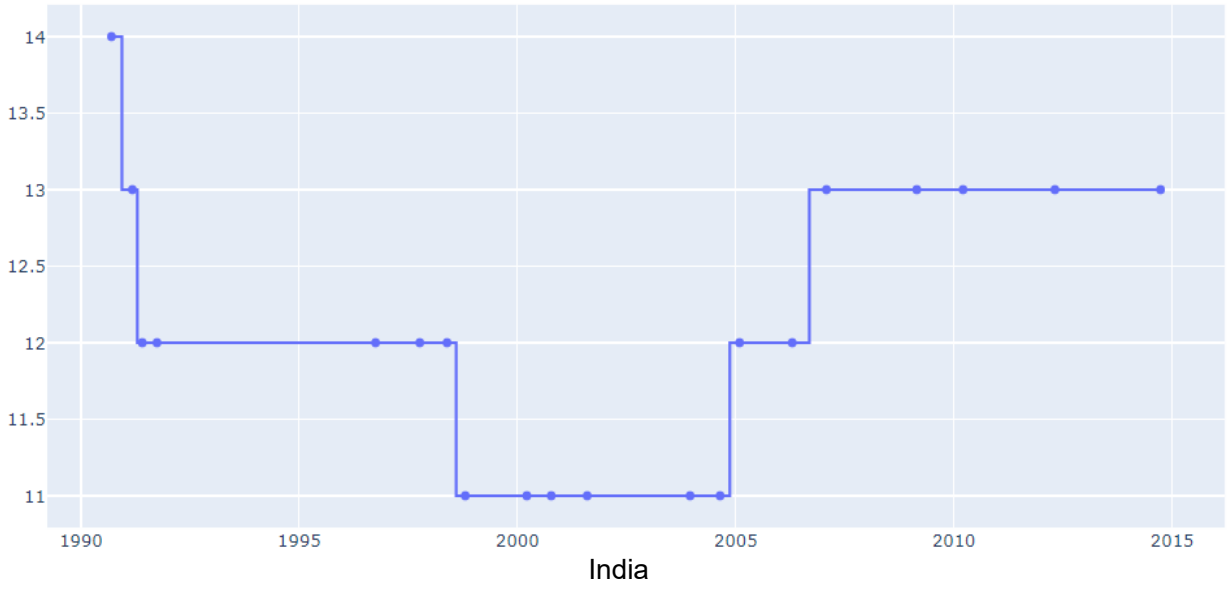
Data Visualisation

Geographical heat map

To begin visualisation on our dataset, we first started with geographical heat map visualisation. To plot this we used numeric ratings based on their S & P rating. Highest rating 22 for AAA rating (Top Notch) which is yellow coloured in the graph and lowest rating 1 for D rating (Junk) which is violet colour in the graph. This is the graph for September 2011, similarly we can make graphs for months as well by changing month in the code.



Baseline plot of S&P rating



This plot shows how the credit rating of India changes from 1990 to 2015. where 14 means BBB refer table no 3. And similar graph can be obtained for all the 135 countries using the code in "main_code.ipynb"

Model Implementation

1. Autoregression Models (AR Models)

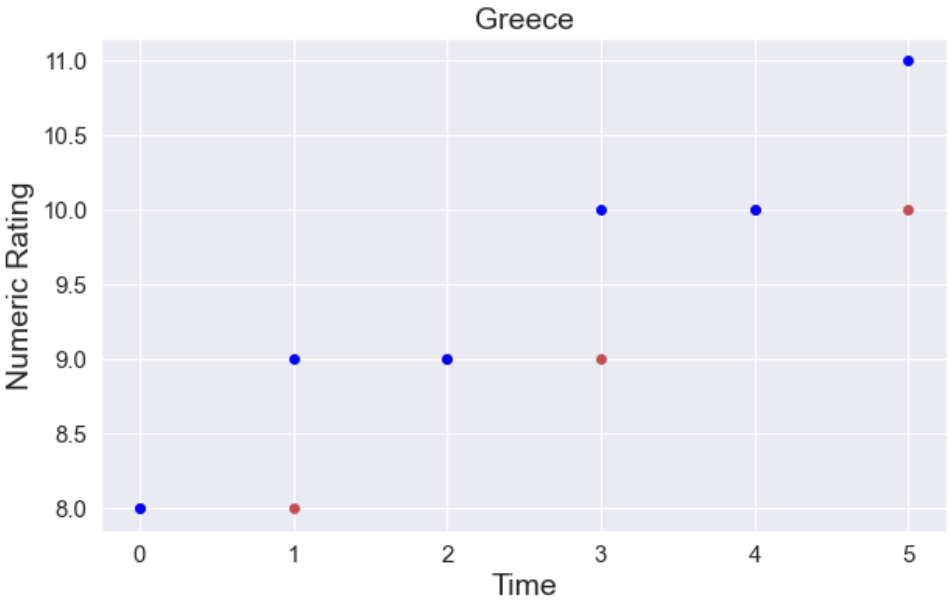
$$Y_t = \beta_0 + \beta_1 * Y_{t-1} + \beta_2 * Y_{t-2} \quad (2^{nd} \text{ Order AR model})$$

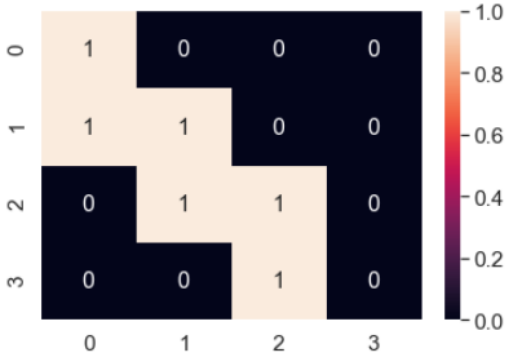
- β_0, β_1 and β_2 are constant
- Y_t is rating at time t
- Y_{t-1} is rating at time t - 1
- Y_{t-2} is rating at time t - 2

We divided the code of this model into two parts:

- First part of the code will run 135 times to predict the rating of each country individually by taking input of all past ratings of that particular country.
- Second part of the code runs once depending on the input country's ISOcode. Demo for the country Greece is shown in "main_code.ipynb". The whole data of Greece is divided into two: training and test data and a confusion matrix is obtained.
- Similar graph and confusion matrix of all the 135 countries can be obtained by changing variable "pISOcode" in the "main_code.ipynb"

Result of AR model





Result of AR model

- Blue points are actual data and Red points are predicted data
- 6 predicted rating 3 matches with the original rating. Which means AR model is 50% accurate for country Greece with the number of data that we get after web scraping

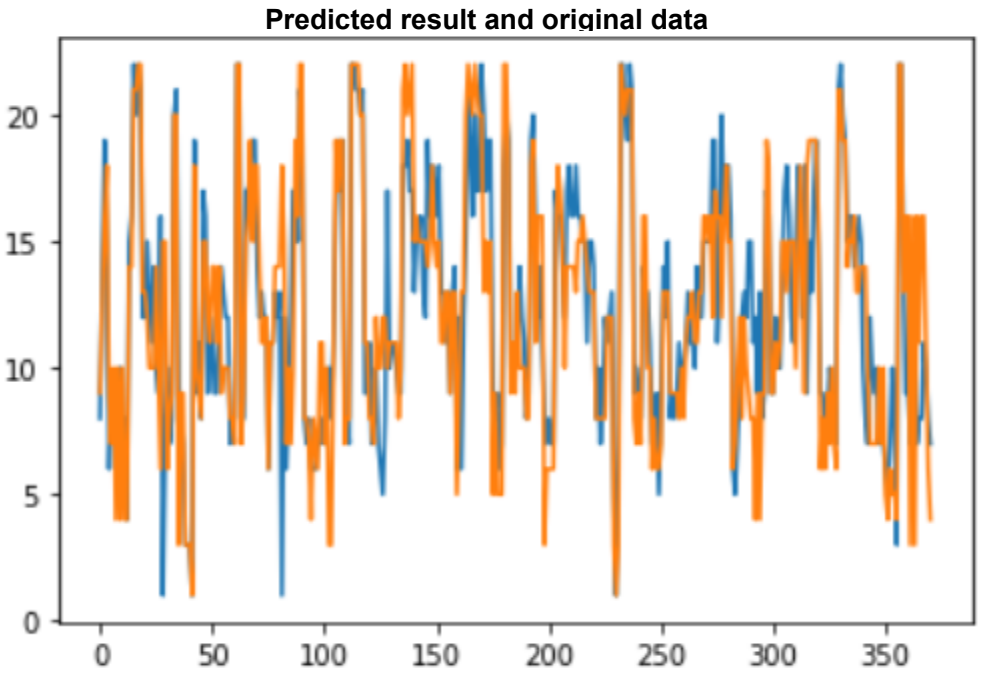
Conclusion of AR model

- AR model is one of the best model for predicting time series data. It's accuracy keeps on increasing as we increase the order of the model and number of input data. We are with constrained with less number of data hence the max accuracy we obtained was 50% for Greece

2. Bayes Classification:

As we know Bayesian classification is based on Bayes' Theorem and the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. So first we are going to split the rating from the dataset then converting one outlook column to three columns namely Positive ,Stable and Negative which have binary values. Then, splitting the rating from the dataset as a target set and for each country we apply the bayes model (python module: GaussianNB) on 80% of data , remaining 20% for testing. And we predicted ratings on the testing dataset .
Then we compare the predicted dataset to the original dataset and find the accuracy.

We get about 0.2 accuracy which is not good but confidence_matrix shows the predicted result is around the original result. If we have three-four times more data we can get good accuracy.



Result :- The plot above shows that the predicted values are around the original values.

Final Conclusion

We need more data for better results for bayes classification. As we can see accuracy is very low but the predicted result is very close to original data, there's less fluctuations between results.