Assignment 2

Aditya Daftari (ad6173)

Question 1: a. Compose a Python script to download the dataset for each year. Your function should take the URL as input, use techniques like web scrapping to extract the data automatically, and return the data in *Pandas.DataFrame* format or save it locally. Note that you are not allowed to open web pages and click 'download the data' by hand.

b. Get data from the links above, and we would be happy to see as much data as you can get. Conduct basic data pre-processing, for e.g., dealing with missing values. Make sure your data is well prepared for the following steps.

Solution: The web scraping code is attached in the source code file. It takes just one year's URL as input and returns the data for that year and for all years before that if available. For data cleaning, I drop all columns (features) with missing values so that the specific feature is not used at all. For salary data with a range, I take the average of the lower and upper values. All feature names are made consistent across the years. Data for 2017 is in image format so could not be fetched.

Question 2: Time Series Data visualization. Plot and analyze the ranking trends of QF programs. How does NYU MFE's ranking vary over the years? Is there a significant relationship between features and ranks? Note your key findings in the notebook.

Solution: The graph below shows NYU MFE's ranking over the last 4 years.

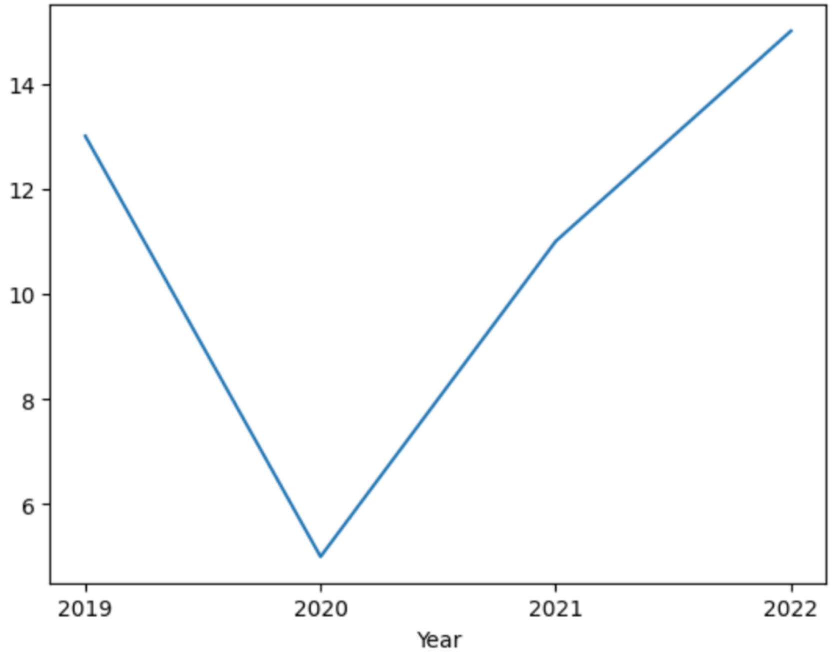
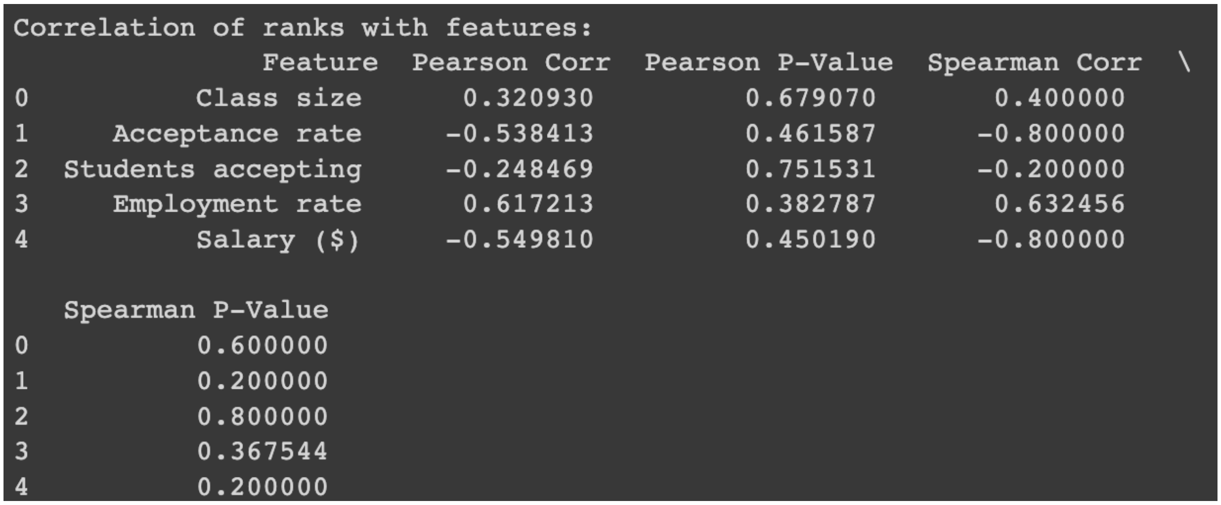


Fig 1: NYU MFE ranking over the years.

On inspecting the Pearson and Spearman correlation of Rank data with the other 5 features we see that positive and negative correlations are present but not statistically significant as the p-values are much higher than the standard threshold of 0.05.



```
Correlation of ranks with features:
             Feature  Pearson Corr  Pearson P-Value  Spearman Corr  \
0         Class size      0.320930         0.679070       0.400000   
1    Acceptance rate     -0.538413         0.461587      -0.800000   
2  Students accepting    -0.248469         0.751531      -0.200000   
3    Employment rate      0.617213         0.382787       0.632456   
4         Salary ($)     -0.549810         0.450190      -0.800000   

   Spearman P-Value  
0          0.600000  
1          0.200000  
2          0.800000  
3          0.367544  
4          0.200000  
```

Fig 2: Correlations of Rank with features

Pearson Correlation looks for linear relationships while Spearman correlation can find out nonlinear relationships. Looking at p-values for Pearson Correlation, they are all very high. Looking at p-values for Spearman Correlation, even though all of them are still above the threshold of 0.05, the most significant features out of these seem to be Acceptance rate and Salary ($) with p-value of 0.2 each and correlation of -0.8 each. This shows that there might be a nonlinear relationship between the two factors and the rank. A negative correlation of rank and salary is intuitive as higher salary would demand in a lower rank. However, negative correlation of Acceptance rate is surprising, as intuitively a lower acceptance rate means the program is more selective and therefore would demand a lower rank.

Question 3: Demonstrate factor changes of NYU MFE over the years. What types are these factors? Find out which factors (features or variables) of NYU MFE changed most and make hypotheses about how these changes would affect the rankings. Hint: you can try different data normalization schemes and analyze how it impacts the ranking.

Solution: Fig3 shows the time series for factor values after applying standard scalar. Fig 4 shows factor changes (factor_t – factor_{t-1}) after scaling those changes by [max(diff) – min(diff)], to find out the factors that have changed the most. Looking at Fig 4, in 2020 we can see the salary had a big increment as compared to 2019 which seems to have driven the drop in rank (drop in rank is good) in 2020. In later years, salary has dropped slightly probably driving the increase in rank(high rank is bad). Along with the salary in 2020, the class size had a huge drop as well which also would have improved the rank significantly. After that year class size seems stable so would not drive the changes in rank. The other factors don't seem to have any significant impact on the rankings. Therefore, the hypothesis is that salary is the main driver behind rank changes.

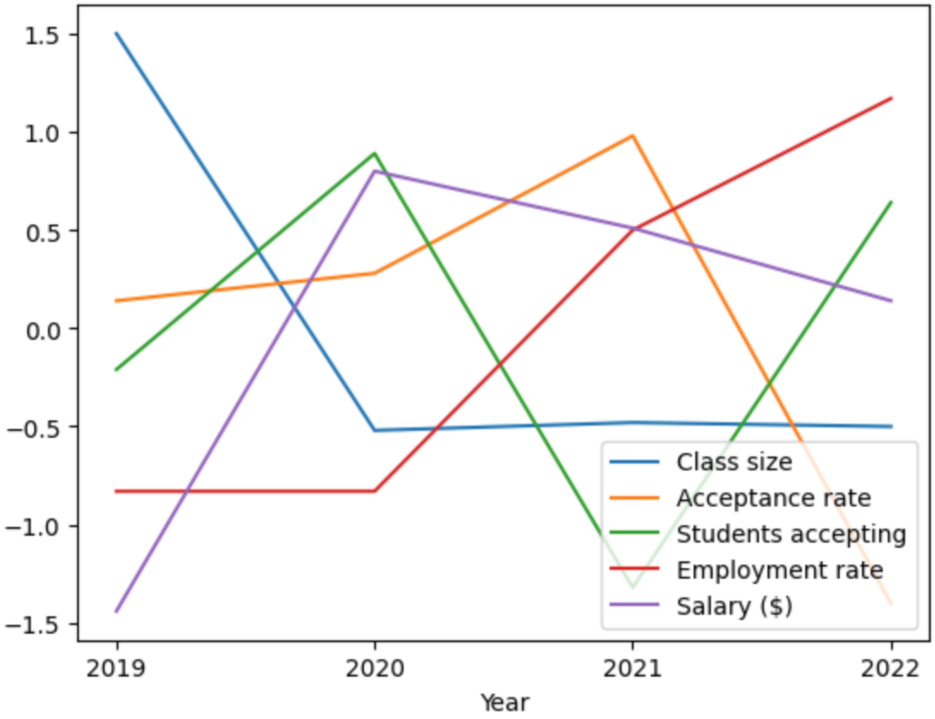


Fig 3: Factor values time series over the years.

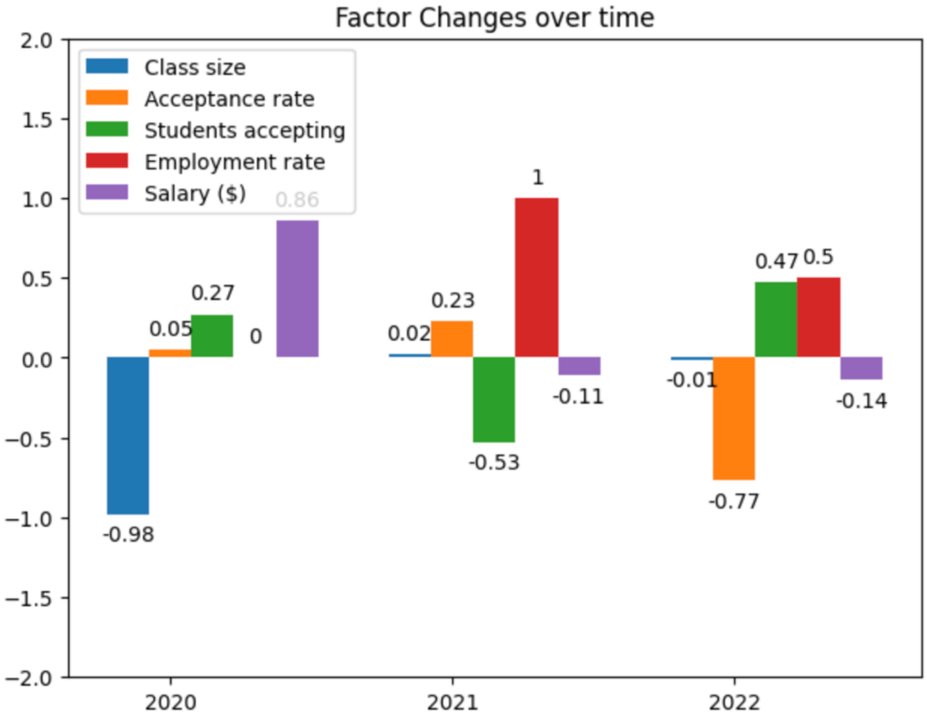
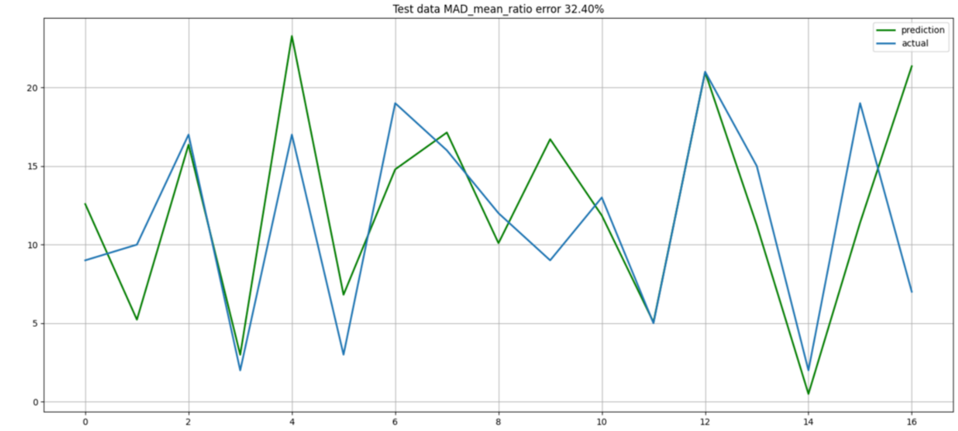


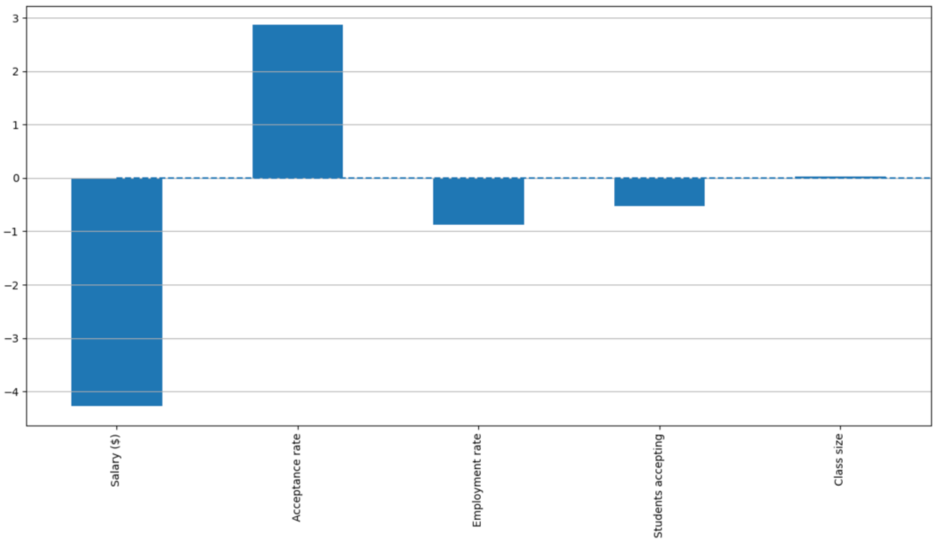
Fig 4: Factor changes over the years.

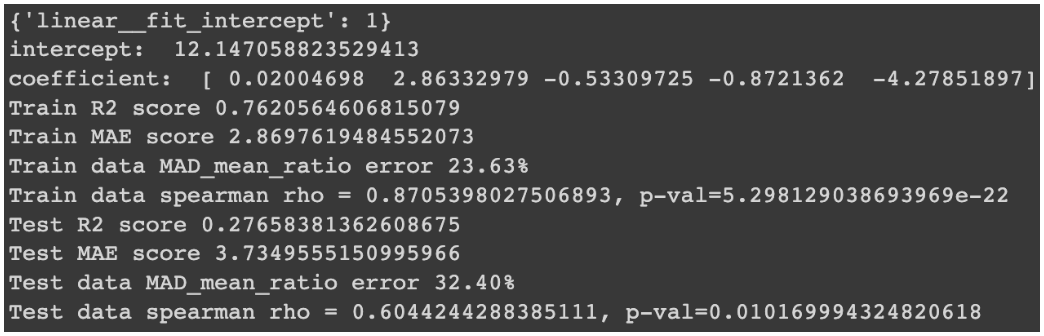
Question 4: Train the factor model to predict the program's rank with given features. Output coefficients of each factor. Is this consistent with your hypothesis in question 3?

Solution:

Factor Model Linear Regression Outputs and coefficients are shown in the following plots. The hypothesis that salary is the main driver is true. However, Acceptance rate also has a large coefficient.





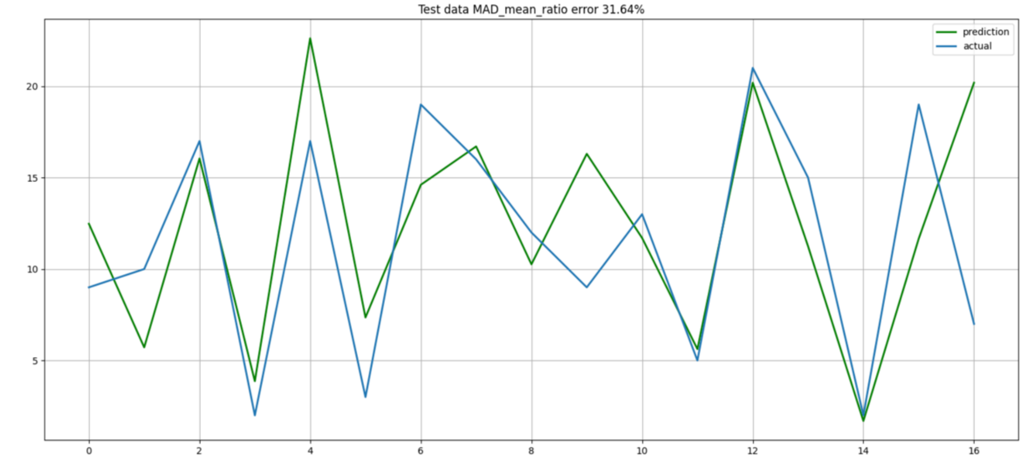


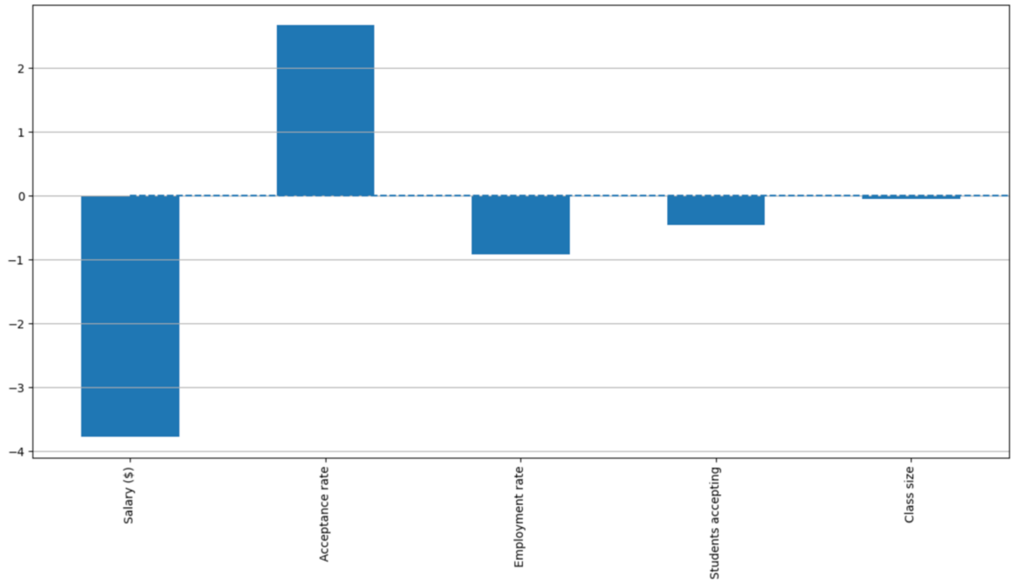
```
{'linear__fit_intercept': 1}
intercept:  12.147058823529413
coefficient:  [ 0.02004698  2.86332979 -0.53309725 -0.8721362  -4.27851897]
Train R2 score 0.7620564606815079
Train MAE score 2.8697619484552073
Train data MAD_mean_ratio error 23.63%
Train data spearman rho = 0.8705398027506893, p-val=5.298129038693969e-22
Test R2 score 0.27658381362608675
Test MAE score 3.7349555150995966
Test data MAD_mean_ratio error 32.40%
Test data spearman rho = 0.6044244288385111, p-val=0.010169994324820618
```

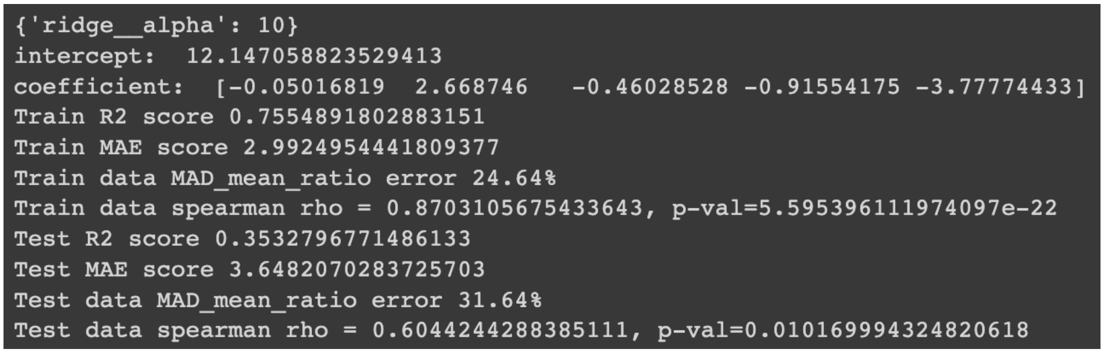
Question 5: Try using other machine learning techniques to fit the rankings. Use proper metrics to evaluate and compare with the results of question 4. What are the pros and cons of these models?

Solution: Alternate models -

1. Ridge Regression







```
{'ridge__alpha': 10}
intercept:  12.147058823529413
coefficient:  [-0.05016819  2.668746   -0.46028528 -0.91554175 -3.77774433]
Train R2 score 0.7554891802883151
Train MAE score 2.9924954441809377
Train data MAD_mean_ratio error 24.64%
Train data spearman rho = 0.8703105675433643, p-val=5.595396111974097e-22
Test R2 score 0.3532796771486133
Test MAE score 3.6482070283725703
Test data MAD_mean_ratio error 31.64%
Test data spearman rho = 0.6044244288385111, p-val=0.010169994324820618
```

Pros

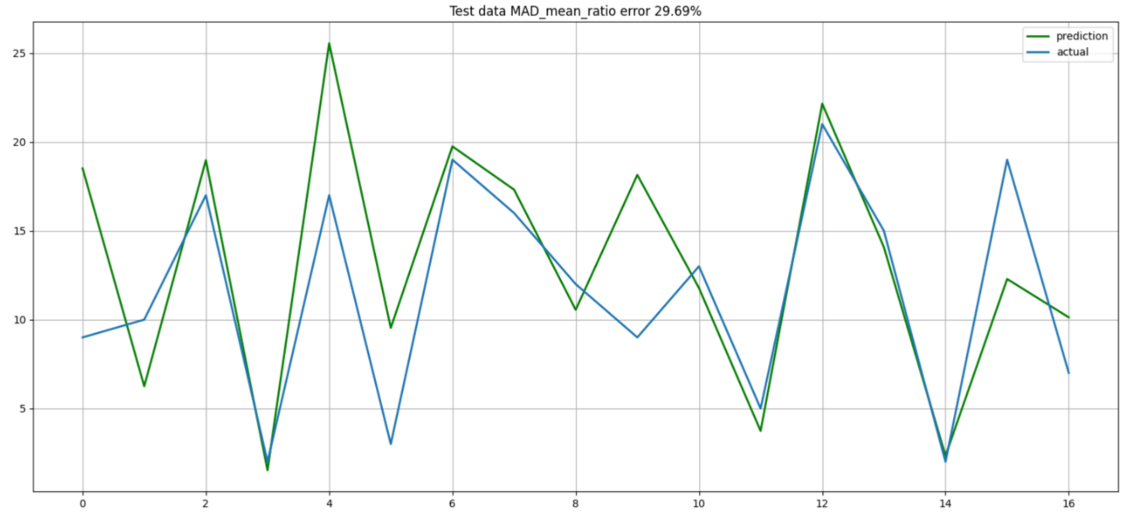
- It helps reduce overfitting by adding L2 penalty in the cost function. We can see it improved the test data performance slightly.

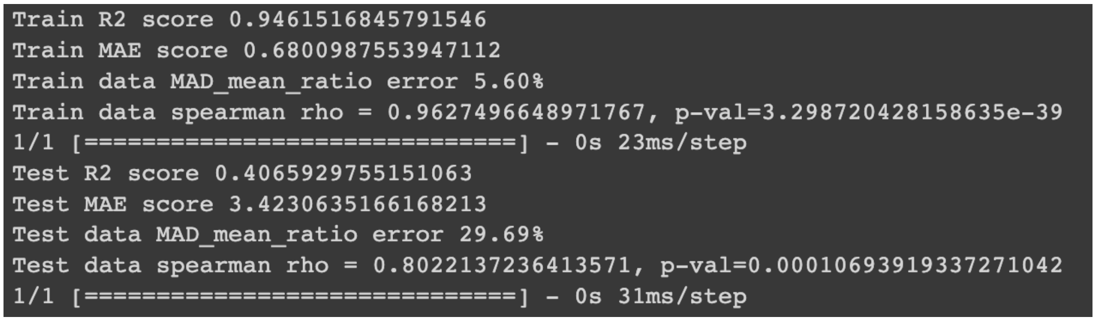
Cons

- Increases bias.

2. Deep Neural Networks

Here I have used NN in a regression form, but it can also be used as a classifier in this problem.





```
Train R2 score 0.9461516845791546
Train MAE score 0.6800987553947112
Train data MAD_mean_ratio error 5.60%
Train data spearman rho = 0.9627496648971767, p-val=3.298720428158635e-39
1/1 [==============================] - 0s 23ms/step
Test R2 score 0.4065929755151063
Test MAE score 3.4230635166168213
Test data MAD_mean_ratio error 29.69%
Test data spearman rho = 0.8022137236413571, p-val=0.00010693919337271042
1/1 [==============================] - 0s 31ms/step
```

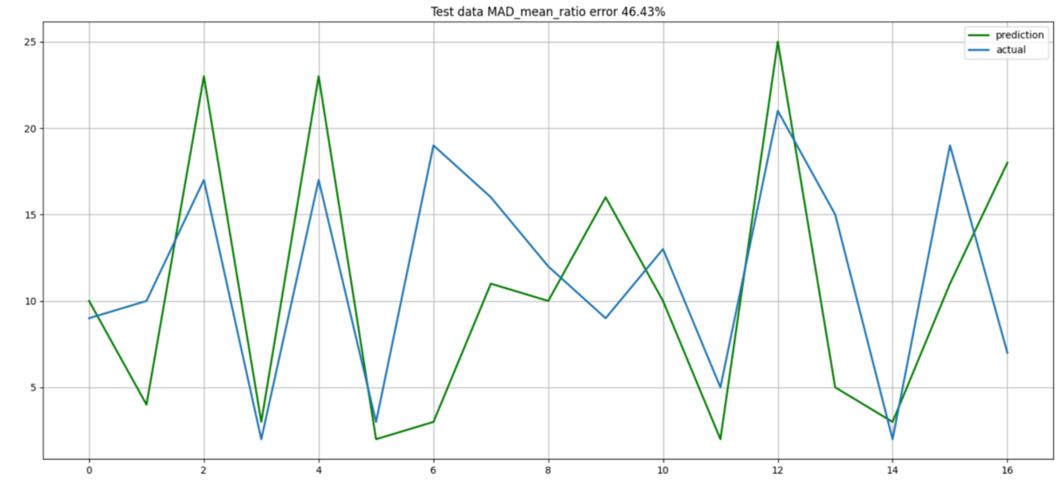
Pros

- Fits to nonlinear data very well. Improved the performance in our case.

Cons

- Does not help in understanding relationship of rank with features.

3. Random Forest



Pros

- Fits training data very well with high accuracy, precision, and recall.

Cons

- Performs poorly in test data because of overfitting.
- Less interpretability of results.