MATH 6380Q Project 1 Peer Review

Report 1. PCA analysis and prediction on the return of SNP500 dataset

[ Reviewer 1 ]

1. PCA analysis and prediction on the return of SNP500 dataset

Summary:

In the project, the author intends to compare the performance of two machine learning algorithms, namely neural network and random forest regression, on the prediction of stock price based on the SNP500 dataset. Before feeding the data directly into the learning algorithms for training, the author employed PCA in a pursuit of dimensionality reduction of the dataset; however, due to the under-performance of it, he chose to directly feed the entire dataset for training the random forest regression and neural network models. And later found that random forests in general perform more robustly than the neural network.

Strength of the report:

The grouping and normalization of the dataset are very reasonable. It is a good assumption that stocks that belong to the same type should behave similarly.

Weakness of the report:

The author concluded that random forest regression in general out-performed the neural network model in this case study. However, he did not vary the number of neurons nor did he adjust the number of hidden layers which are crucial parameters to determine the neural network’s architecture, hence its performance. It is suggested that before jumping to the conclusion, he could compare the cross-validated error and variance of a subset of neural networks with different hidden units and hidden layers, then choose the best one to compare with the random forest regression model.

|  |  |
| --- | --- |
| Evaluation on Clarity and quality of writing (1-5): | 4 |
| Evaluation on Technical Quality (1-5): | 3 |
| Overall rating: | 4 |
| Confidence on your assessment: | 1 |

[ Reviewer 2 ]

In poster one, “PCA analysis and prediction on return of SNP 500 dataset”, author found that PCA could not effectively reduce the dimension of this dataset. Therefore, he used random forest regression and neural network to obtain the result. Author used 452 different stocks’ closed prices in 1258 consecutive market days as the dataset and build two different models based on random forest regression and neural network to predict the return of stocks.

The strength of this poster is that it builds two different models and both two models are evaluated and compared. Finally, one of it gives a more stable result in different classes.

The weakness of this poster is the dataset could not be done with PCA method and the reason is not listed.

Evaluation on Clarity and quality of writing: 4. There is a good use of examples and figures and the report is clearly written. Suggestion is to try more segments to make the paragraphs clearer and make an abstract.

Evaluation on Technical Quality: 3.

Overall rating: 3.

Confidence on my assessment: 3.

[ Reviewer 4 ]

This report gets a stable result by random forest regression. It is clearly written, and the figures look very good.

Evaluation on Clarity and quality of writing: 5

Evaluation on Technical Quality:4

Overall rating:4

Confidence on your assessment:2

[ Reviewer 5 ]

Summary: In this project, the author adopted PCA to analyze the SNP500 data in 4 consecutive years. It was found that PCA could not effectively reduce the dimension of this dataset. Therefore, he then directly used the dataset to do the prediction of return with random forest regression and neural network. The performance of both models are evaluated and compared.

Strengths: Different techniques like random forest and PCA are applied.

Weakness: Even though the first few PCAs doesn’t explain 95% of the variation, they still can provide some information.

Evaluation on Clarity and quality of writing: 4. Typo: ‘as as to predict the return in a reasonable way’ in Results and Discussions.

Evaluation on Technical Quality: 3.

Overall rating: 3

Confidence on your assessment: 2

[ Reviewer 6 ]

1 PCA analysis and prediction on return of SNP500 dataset

1.1 Summary

For Xu Han’s work, he adopted PCA(Principal Component Analysis) on the SNP500 dataset. Initially, He evaluate the capability of PCA by divided the whole dataset into 10 subclass. The feature he used is the return for each subclass stock price. Besides the principle component analysis, he also applied Neural Networks and Random Forrest regression for the prediction, and do a comparison of these two method on the aspect of mean absolute error.

1.2 Strength and Weakness

The strength of this paper is that he involved two methods for prediction, Neural Networks and Random Forrest regression, which Prof.Yao did not cover in the class. The diversity of this report is good. The weakness of this paper is that he did not implement a parallel analysis on the data he used. Besides, the used of stock price return seems a bit doubtful for me, the return can shows changes within one subclass, but hard to reveal the big picture of the stock price trend.

1.3 Score

1.3.1 Clarity and Quality of Writing

The writing of Han is well-organized, the structure of the poster is clear. It is definitely good to include some briefly introduction for the PCA method. It is obvious that Han puts his effort on this project. However, the typesetting of this poster can be improved. I will give him 4/5 on this aspect.

1.3.2 Technical Quality

This poster constructs a thorough investigation on PCA and regression method for data regression. However, the type of Neural Network did not clarified. Furthermore, the content includes three method, more information for each method would be better. Also, parallel analysis can be a very suitable add-in for this report. I will give him 4/5 on this aspect.

1.3.3 Overall

The overall score for this poster is 4/5.

[ Reviewer 7 ]

Summary of this poster: In this report, they analyze the SNP500 dataset by using two models. They find that PCA couldn’t effectively reduce the dimension of their dataset. Thus, they use the dataset to do the prediction of return with random forest regression and neural network. Then the two models are analyzed and compared in the report.

Describe the strengths of the poster: The poster uses two different models and compares the performance of these two modes, which is rigorous and exhaustive.

Describe the weaknesses of the poster: This poster doesn’t provide the detail information and laws from the dataset. It would be better if they extract the useful information from the dataset by analyzing the distributions of PCA components.

Evaluation on Clarity and quality of writing (1-5): 4

The report is generally in well written. The figures are well organized. But there are some typos in this poster: 1) Data dimension reduction is very important to analyzing (analyze) the given data and predicting the unknown result. 2) Then, as as to predict the return in a reasonable way.

Evaluation on Technical Quality (1-5): 4

This report uses two models to analyze the dataset: one is PCA and the other one is the random forest regression and neural network. It also gives the detail methods of these two models. Thus, theoretical analysis is good.

Overall rating: 4

Confidence on your assessment: 2

I am not very familiar with the netural network, so I just browse the paper without checking the technical details.

[ Reviewer 8 ]

• Summary of the report.

This report applies PCA on SNP500 data and does prediction using the random forest and neural network.

• Describe the strengths of the report.

It is clearly written.

• Describe the weaknesses of the report.

About half of the report is repeating the existing theory, lack of own analysis. The author did PCA, but failed to take advantage of it. Since financial data tend to have much noise, it might be better to apply some extended PCA methods to capture the weak signal.

• Evaluation on Clarity and quality of writing (1-5): Is the report clearly written? Is there a good use of examples and ﬁgures? Is it well organized? Are there problems with style and grammar? Are there issues with typos, formatting, references, etc.? Please make suggestions to improve the clarity of the paper, and provide details of typos.

4-

Little suggestion: 1. Change a little bit the arrangement of the figures instead of all on the bottom. 2. PCA is principal component analysis, no need of “PCA ANALYSIS”. Also PC means principal component, no need of “PCA component”.

• Evaluation on Technical Quality (1-5): Are the results technically sound? Are there obvious ﬂaws in the reasoning? Are claims well-supported by theoretical analysis or experimental results? Are the experiments well thought out and convincing? Will it be possible for other researchers to replicate these results? Is the evaluation appropriate? Did the authors clearly assess both the strengths and weaknesses of their approach? Are relevant papers cited, discussed, and compared to the presented work?

2

About half of the report is repeating the existing theory, very lack of own deep analysis.

• Overall rating: (5- My vote as the best-report. 4- A good report. 3- An average one. 2below average. 1- a poorly written one).

3

• Conﬁdence on your assessment (1-3) (3- I have carefully read the paper and checked the results, 2- I just browse the paper without checking the details, 1- My assessment can be wrong)

3

[ Reviewer 9 ]

1. **Summary**

This report studied SNP500 dataset which contain 452 different stock’s closed prices in 1258 consecutive market days. The author used the PCA method to do the dimension reduction however it could not effectively reduce the dimension of the dataset. So the author applied random forest and neural network to predict the return of the dataset.

1. **Strength of the report**

The analysis of the data is reasonable and clear. The methods applied are clearly explained.

1. **Weakness of the report**

The report did not give a conclusion related to the trend of the data, the readers could not get a big picture of the features of the data.

1. **Evaluation of clarity and quality of writing**

The report is clearly written and well organized. The figures are clearly explained. No obvious problems of the style, grammar, typo, and formatting are examined.

1. **Evaluation on technical quality**

The results are technically sound and no obvious flaws in the reasoning are found. The report provides the methods for analyzing similar dataset. Both the strength and weakness of the PCA method is analyzed. Proper papers are citied however did not compare with present work.

1. **Overall rating**

An overall rating of 4 is given.

1. **Confidence on your assessment**

2- I just browse the paper without checking the details

[ Reviewer 10 ]

*Summary:*

Conducting PCA preprocessing before train a random forest model or neural network model to predict the performance of different industry sectors’s stocks in SNP500.

*Strength:*

Background introduction is in detail. Predict the stock value using random forest and neural network.

*Weakness:*

Since the project requires using PCA or MDS on those dataset, the PCA in this poster seems not so important. The analysis shows the PCA is not so useful. Deleting the PCA part and showing the predicting result seems like a better story.

Did not show the prediction result or performance when two sections involving the random forest model and neural network model.

*Evaluation on Clarity and quality of writing (1-5): 3*

Since the work is conducted by one person, the poster sometimes use ‘we’ and sometimes use ‘I’, maybe replacing ‘we’ with ‘me’ or ‘I’ is more accurate.

It would be nice to uniform the figure format, like have a line for x-axis for all figures instead of one figure.

The explanation of extreme values in historical returns is “Those major events include stock split, dividend and etc”. For the given data, the extreme value is hundreds times than the return low band, it’s likely caused by more influential events like M&A.

“The comparison can be further studied in Fig. 4, which is a scatter plot of two components.” May be the writer want to say, the historical return is projected on the space of first two principal components by PCA.

The name of ‘return’ is not so easy to understand, maybe the percentage of changes in stock price is more like ‘price growth rate’.

*Evaluation on Technical Quality (1-5): 3*

The method part introduced PCA, random forest and neural network in detail. It would be good if shows the gap between these methods (PCA, random forest and neural network) and how the writer design his/her experiment at this part.

It seems the writer is familiar with the PCA method already. And it would better to tell a story PCA is useful.

*Overall rating (1-5): 3*

*Conﬁdence on your assessment (1-3): 3*

[ Reviewer 11 ]

Summary: This report finds that PCA cannot effectively reduce the dimension of SNP500 dataset and then they do prediction on the raw data using random forest and neural network.

Strengths: Clearly states the methods and results.

Weaknesses: Lack of analysis.

Writing: (3) Clearly states the methods and results. The author talks too much detail so that the characters are too small for a poster.

Technical Quality: (3) The report does not explain why PCA cannot reduce the dimension. I think the author can use the PCA first and then apply random forest or neural network to do prediction and compare the prediction error with raw data.

Overall rating: (3)

Confidence on my assessment: (2)

[ Reviewer 12 ]

**Summary:** This project explores the application of PCA on SNP 500 dataset and finds it not suitable for the dimension reduction of this dataset. Besides, it also uses random forest regression and neural network methods to predict the return of a single stock based on the other stocks’ return.

**Strengths:** This project uses return (rate) of the stocks rather than the initial values as the initial data to process, which can better reflect the change of different stocks and make the results of different stocks comparable. After the unsuccessful trial of PCA, the project turns to other methods like random forest regression to process the data which gives us another perspective of the data.

**Weaknesses:** PCA is for dimension reduction while random forest regression and neural network are for regression problems. They are not consistent and comparable to some extent. Besides for the regression part, I think to predict the return of a stock is more likely to be a time series problem and it may be better to do the prediction with a series of data rather than only those of a single day.

**Evaluation on Clarity and quality of writing:** 4

There are some little grammatical mistakes in the poster. But overall the format of the poster is good.

**Evaluation on Technical Quality:** 3

I would suggest the author to further analyze the PCA results like explore the distinctive points that are far from the others in the principal components map.

**Overall rating:** 4

**Confidence on your assessment:** 2