EE 518: Machine Learning Theory & Algorithms **Spring 2024**

Mini Project 3

Due: June 7 (official) /June 14, 2024, 11:59PM PT

Student Name: *Instructor Name: John Lipor*

Note: The "official" due data will be June 7, meaning I will not give feedback or answer questions after that date. However, you are welcome to take the extra week to continue working on your project, turning it in on June 14.

Introduction

In this mini project, you will use machine learning to make some money (maybe. . . well, probably not) by predicting future stock prices [1]. Future predictions of this sort are known as time-series prediction and are a common application of regression techniques. While we have not studied time series data, this problem can be easily cast as a supervised regression problem (see e.g., [2]). This project involves predicting future stock returns from the Tokyo Stock Exchange. However, the definition of a "return" is nuanced and includes not only return but the degree of uncertainty in this return. The competition also directly accounts for your ranking of stocks.

Notes:

1. DSS rules apply for all mini projects.
2. Be sure to cite any sources you used, including/especially Kaggle notebooks.
3. You are free to use any algorithm(s) you choose, without exclusion.
4. While our boosting method of choice has been `XGBoost`, you may also wish to consider reading about or using `LightGBM` or `CatBoost`.
5. Getting your bearings in this project can be difficult. I recommend sorting the public code by *Most Votes* for helpful insights.

Requirements

You must create a report in either LaTeXor a Jupyter notebook that contains the following sections:

1. Problem description
   - Explain what data you have to work with, what algorithms you will use, what your goal is, and why anyone should care.
2. Exploratory data analysis (EDA)
   - There are many excellent tutorials on EDA online. Be sure to cite whichever ones you found helpful and comment on them in Sec. 6.
   - Be sure to make any significant findings stand out and try to keep this concise.

3. Challenges
    - What were the challenges you encountered when applying machine learning to this dataset?
    - Did these challenges mainly result from the data? From results? Installing libraries to perform preprocessing?
4. Approach
    - Provide a thorough but concise description of your approach, including (but not limited to) your approach to wrangling, preprocessing, and improving the performance of your predictor.
    - This should be much clearer than what is typically found in Kaggle notebooks, but make sure it stays concise. A summary paragraph at the beginning of this section could be helpful.
    - Since we have now formally studied model selection, you must use either a validation set or cross validation to tune your model parameters. Be sure to describe how you went about this and why.
5. Evaluation and summary
    - **(NEW)** Be sure to carefully describe your understanding of the competition's evaluation metric and how you accounted for this.
    - **(NEW)** In addition to your evaluation plots, you should make a submission to the competition and report your score.
    - Describe why/where your approach falls short in terms of estimation and approximation error. You may wish to generate a learning curve, as described by Ch. 11 of the text.
    - Consider diving into the data to see if you can determine any trends regarding which points were misclassified.
    - Likewise consider if there were any features that were particularly important or unimportant.
    - Summarize your solution, describing what worked, what didn't, what the main limitations are, and any general conclusions about the dataset.
6. What I learned
    - Describe the main skills/tools that you learned and used for this project and how you learned them.

Grading

The grading breakdown is below. Each section will be graded according to technical correctness, effort, and creativity. Note that clarity of writing is a major component. You should put yourself in the place of writing for a boss or senior in a workplace. If your writing is terrible, they will assume your work is terrible.

Item	**Percentage**
Description	5%
EDA	20%
Challenges	10%
Approach	30%
Evaluation	20%
What I learned	5%
Clarity/conciseness of written communication	10%

References

[1] (2022) Jpx tokyo stock exchange prediction. [Online]. Available: https://www.kaggle.com/competitions/jpx-tokyo-stock-exchange-prediction

[2] (2016) Time series forecasting as supervised learning. [Online]. Available: https://machinelearningmastery.com/time-series-forecasting-supervised-learning/