FA 590 Statistical Learning in Finance

2024 Fall

Final Project

The primary goal of this project is to design, implement, and evaluate statistical learning models to predict stock-level risk premiums. You will work with a comprehensive dataset covering 60 years of stock data, comprising 30,000 stocks and 94 firm-specific characteristics for each stock, as well as eight aggregate time-series variables representing broader macroeconomic conditions. A detailed explanation of all dataset features is available in the "Feature Dictionary.xlsx" file. You can access the dataset through the following link: https://stevens0-my.sharepoint.com/:f:/g/personal/zyang99_stevens_edu/EtzdCCW11WJHjvwxytwQMCEBLSSUhjZDAJ6Ir_UOofKN1g?e=cCzp8f

Submission Guidelines

1. Written Report
    a. Submit a written report in PDF format, with a maximum length of 30 pages. This limit includes all figures, tables, and references.
    b. The report should be written with either 1.5 or double line spacing, using a 12-point font size.
    c. Your report will be the primary basis for grading and should clearly explain your modeling approach, results, and conclusions.
2. Code: Submit your code in .py / .ipynb (Python), .R / .Rmd (R), or other code files, ensuring it includes all necessary components to replicate your main results.

Project Scope You may choose to pursue one of two project directions:

1. Regression: Predict stock-level risk premium (a continuous variable).
2. Classification: Classify stocks into categories based on their risk premiums.

Appendix A provides further instructions on the final project.

Group Work

The final project is a group project. You may form groups of up to three members. All members of a group will receive the same grade. To sign up, please use the following link to register your group: https://stevens0-my.sharepoint.com/:x:/g/personal/zyang99_stevens_edu/EUKqQIcG3I9OtqfzqdOJiy8BsBGvzgRHTjlSfQX5Q4yJ5g?e=cYsxIL

Final Project Award

An award will be presented to one group in each project direction (regression and classification). The winning groups will have the opportunity to present their projects in class, and each member of the winning group will receive 5 bonus points added to their final grade (conditional on completing the presentation).

Important Dates

- 3 PM, September 24, 2024: Deadline for signing up the group members and group direction.
- 3 PM, November 26, 2024: Deadline for the final project. You should submit your written report and code no later than this deadline.
- December 3, 2024: Announce the Final Project Awards.
- December 10, 2024: Award-winning groups are invited to present their projects during the lecture time.

Computing Resources

If your personal laptop does not have sufficient computing resources to handle the data size for this assignment, you have access to the computing resources at the Hanlon Financial Systems Center, available at two locations:

- **Hanlon Lab 1** (4th floor, Babbio): 30 workstations equipped with Intel i9-11900 CPUs, 64GB RAM, RTX 3070 GPUs (8GB GPU memory), 1TB SSD for Windows 11, and 2TB HDD for Ubuntu 20.04.
- **Hanlon Lab 2** (1st floor, Babbio): 33 workstations equipped with Intel i9-14900 CPUs, 32GB RAM, RTX 4000 Ada GPUs (20GB GPU memory), 1TB SSD for Windows 11, and 1TB SSH for Ubuntu 22.04.

The center is open Monday through Saturday. You can view detailed availability on their calendar [Link to Calendar]. Reservations are not required—simply log in with your Stevens account to use any available workstation.

Appendix A Further Instructions

1 Define Your Target Variable

Risk premium is defined as the excess return of a stock over the risk-free rate. Specifically, the risk premium is calculated as the difference between the stock's return, which represents the price change from one month to the next, and the risk-free rate, which is approximated by the three-month U.S. Treasury Bill rate. The risk premium is provided in the dataset with feature name "*risk_premium*". The dataset has been processed in the way that you can directly use the predictors in the same month to predict the risk premium of that month.

The target variable for regression is the risk premium provided in the dataset.

The objective of classification is to classify the stock-level risk premiums into different baskets. For example, you can classify the risk premiums to be positive or negative (a binary target). Or you can classify the risk premiums into deciles, i.e., ten baskets of the same size. If you choose the classification direction, you need to specify and justify your predictive target.

2 Statistical Learning Models

Implement linear models, tree-based models, and neural networks to predict/classify the risk premium, and tune the model hyperparameters.

3. Model Evaluation

The model evaluation is divided into two components: predictive performance and portfolio performance. Predictive performance measures how accurately the model predicts outcomes, while portfolio performance assesses how well those predictions translate into portfolio gains. For regression tasks, predictive performance is evaluated using metrics such as R-squared and mean squared error (MSE), while classification tasks are assessed using metrics like the F1 score and area under the curve (AUC).

Using the predicted risk premiums from the statistical learning models, construct an equal-weight portfolio consisting of the top 100 stocks. To do this, sort the stocks by their predicted risk premiums for each month and uniformly allocate investments across the top 100. Calculate the average portfolio risk premium, volatility, and Sharpe ratio to assess the portfolio's performance.

Denote the risk premiums for the top 100 stocks at month t as $r_{1,t}, r_{2,t}, \dots, r_{100,t}$. The portfolio risk premium at month t of an equal-weight portfolio is

$$r_{p,t} = \frac{1}{100} \times \sum_{i=1}^{100} r_{i,t}.$$

Assume there are a total of T months, the average portfolio risk premium is

$$\bar{r}_p = \frac{1}{T} \times \sum_{t=1}^{T} r_{t,p}.$$

The volatility is

$$\sigma_r = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} \left(r_{t,p} - \bar{r}_p\right)^2}.$$

The Sharpe ratio is calculated as

$$S = \frac{\bar{r}_p}{\sigma_r}.$$

In your report, you need to report both the in-sample and out-of-sample performance.

4. The Written Report

4.1 Open Questions

While building the statistical learning models, consider exploring the following open questions (but not limited to these):

- How do nonlinear models compare to linear models in predicting stock risk premiums? What impact do nonlinearities and interaction effects have on improving the accuracy of these predictions?

- Does more accurate model prediction necessarily lead to better portfolio performance? If not, what factors might cause a portfolio based on more accurate predictions to underperform compared to portfolios built on less accurate predictions?
- As an equity trader, how would you convince your supervisor to adopt complex statistical learning algorithms, such as neural networks, for stock analysis?

4.2 Report Content

The written report should include, but is not limited to, the following components: exploratory data analysis, experiment design (e.g., sample splitting, feature engineering, and evaluation metrics), modeling (e.g., model selection, feature selection, and hyperparameter tuning), model evaluation, and the takeaways from the analysis.

Additionally, the report should disclose whether and how generative AI tools, such as ChatGPT, were utilized. If applicable, please highlight effective use cases of generative AI to share with the class.

Appendix B Final Project Groups and Directions

Group	Member 1	Member 2	Member 3	Project Direction
1	Andre Sealy	Peng Fu	Vincent Bin	Regression
2	Adam Moszczynski	Alex Weigel		Regression
3	Federica Malamisura	Lucía de Alarcon		Regression
4	XIAO JIN	Daiki Ishiyama		Classification
5	Darshan Nanjegowda	Remoun Salib		Regression
6	Trevor Lenig	Nazarii Tretiak		Classification
7	Kangping Zeng	Ruoyao Pei	Shijie Zhou	Classification
8	Zelin Chen			Classification
9	Elisa Colusso	Alma Nasic	Pooja Pande	Regression
10	Swapnil Pant	Hao Cai	Vedant Dhaval Vaidya	Regression
11	Dennis Vink	Vincent Grillo		Classification
12	Mrinal Gupta			Classification