ML-based US Stock Return Prediction and Asset Allocation

By: Tianze Shou, Wenhan Li, Lucy Hu, Mia Zhang

Advisors: Eli Ben-Michael, Cosma Shalizi

Clients: Lars-Alexander Kuehn, Mingjun Sun

INTRODUCTION

Background

- US stock market has evolved a sophisticated information system, enhancing market transparency and efficiency.
- Quantitative trading strategies are widely used by market participants to manage large volumes of information.
- Specifically, machine learning (ML) have been a key component of those quantitative strategies.
- With ML models, market participants can make better predictions of future stock prices and evaluate corresponding risks of their portfolios.
- Therefore, our group implemented different ML models to predict the the monthly excessive returns of these traded stocks, which helped optimize our portfolio construction.

Problem Statement

1. How to effectively predict excessive return of stocks in the future with historical data?
2. How to allocate portfolio such that wealth can be maximized at the end of a given period?

DATA

Data Source

Our data of analysis is the Center for Research in Security Prices (CRSP) dataset on stocks in the US from NYSE, AMEX, and NASDAQ together with several hundred hand-crafted features since December 31, 1925.

Data Processing

- Selected variables: response variable: `ret_exc_lead1m`; 21 predictor variables:

Valuation	Performance	Risk	Analysis	Other
- be_me - market_equity - ebit_sale - at_grl - sale_grl - cash_at	- ret_12_1 - ret_1_0 - ret_60_12	- rvol_252d - beta_252d - qmj_safety - rmax1_21d - chcsho_12m	- ni_me - ope_be - gp_at - at_be	- eq_dur - age - z-score

- Rank transformation: each characteristic is transformed into the cross-sectional rank

METHODS

Training Structure

- Our training approach for stock data uses a rolling structure: 10 for training, 5 for validation, and 1 for testing (figure 1).
- For the models that include hyperparameters (eg. regularization parameter), we tuned them using grid search on validation data.

Models

1. Elastic Net: a model combines Ridge regression's parameter shrinkage with Lasso regression's feature selection, effectively limiting the model's degree of freedom. The objective function is: $RSS + \lambda * [(1-\alpha) * ||\beta||_2 + \alpha * ||\beta||_1]$
2. Random Forest Regressor: an ensemble learning method that builds multiple decision trees with L2 regularization. We also set limitations on maximum depth of trees, number of trees, etc., to avoid overfitting.
3. Neural Networks: a deep learning model that aims to capture the deep, latent, and hierarchical representation of input features. We implemented three-layers and five-layers networks (we adapted NN5 to include the residual link and dropout technique), and their model architectures are shown respectively in figure 2 and figure 3.
4. Logistic Regression: the model applies multinomial classification with L1 regularization, combining log-likelihood maximization with feature selection to efficiently handle multi-class problems and maintain a sparse solution. Expression: $\min -\sum_{i=1}^{N}\sum_{k=1}^{K}[y_{ik}\log(p_{ik})] + \lambda\sum_{j=1}^{M}|\beta_j|_w$

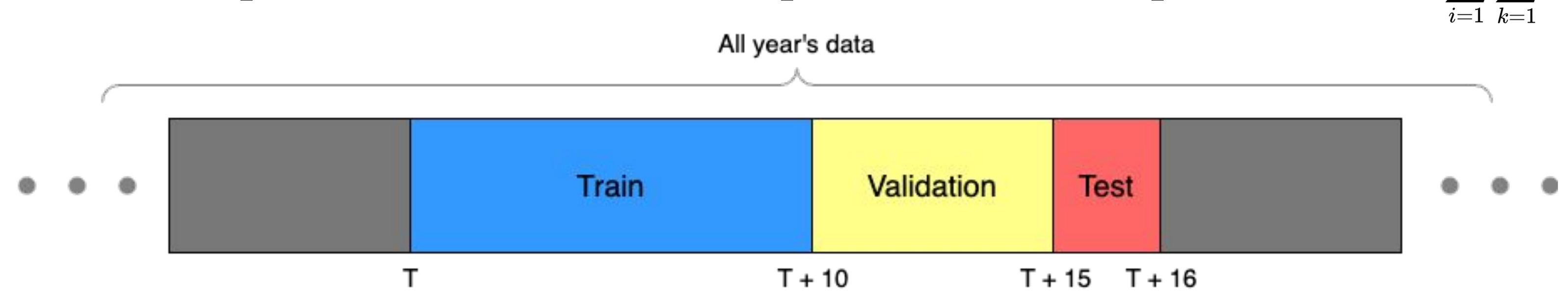


Figure 1: Train-Validation-Test Structure

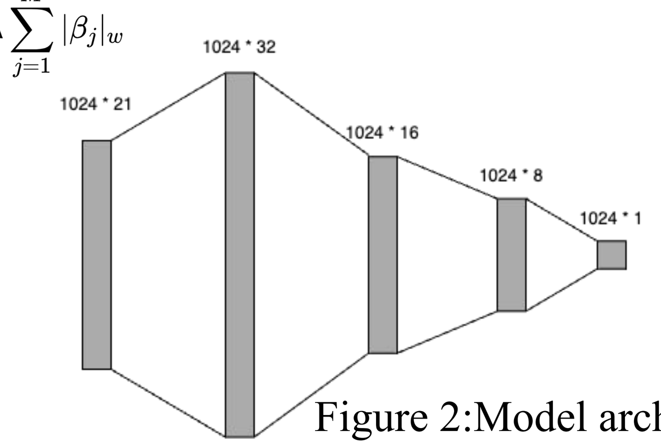


Figure 2:Model architecture of NN3

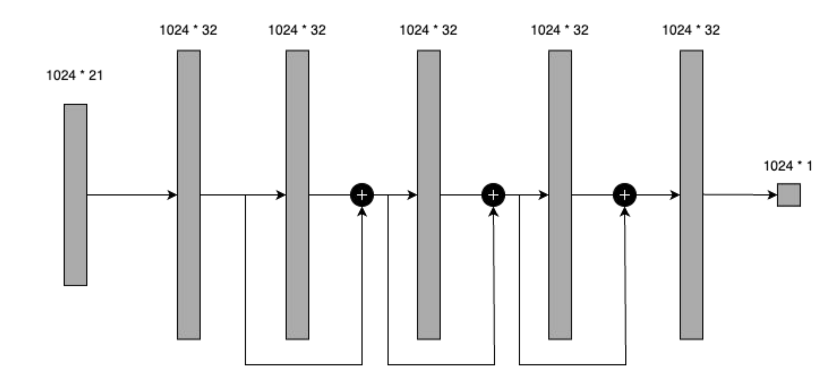


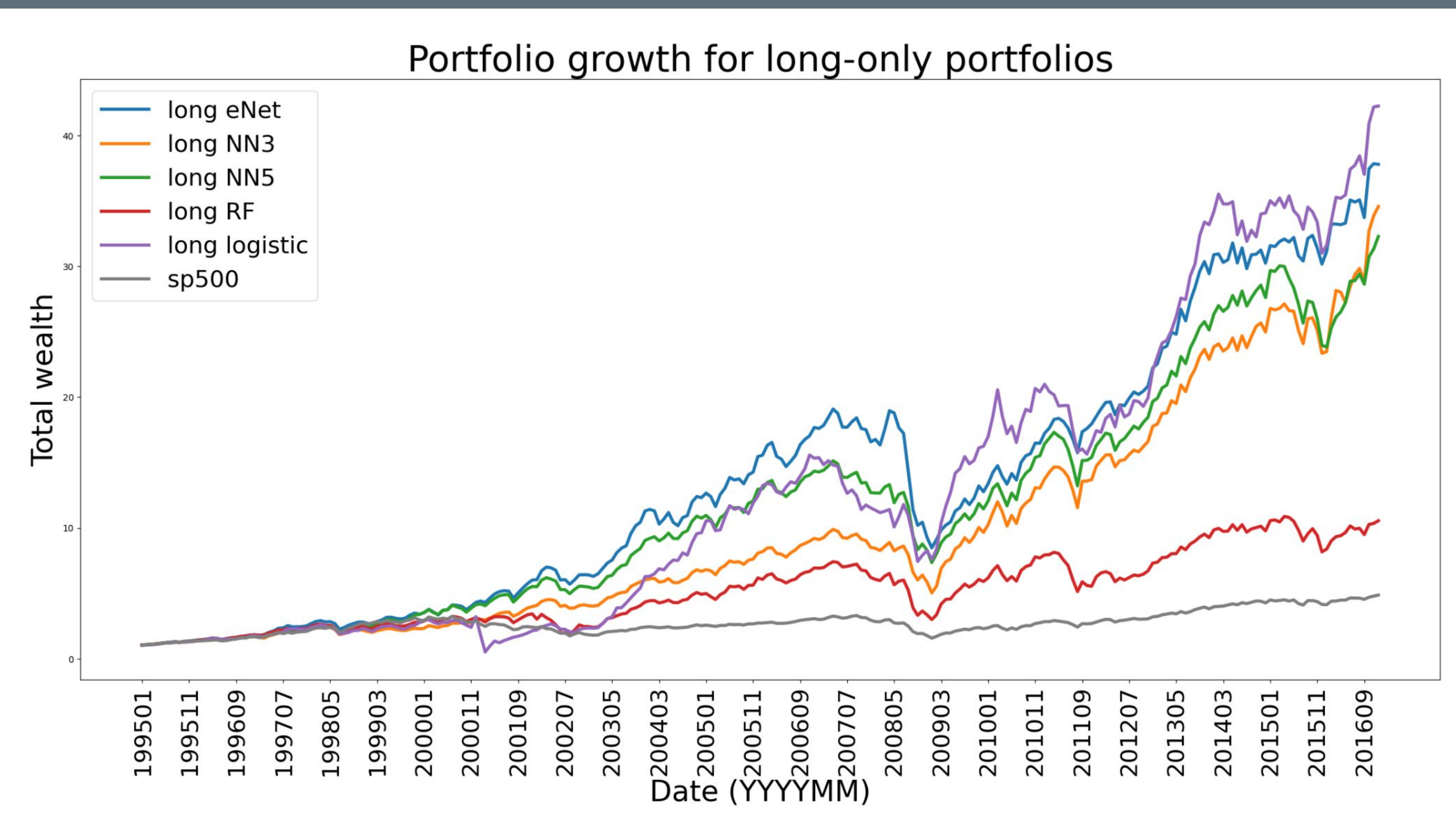
Figure 3: Model architecture of NN5

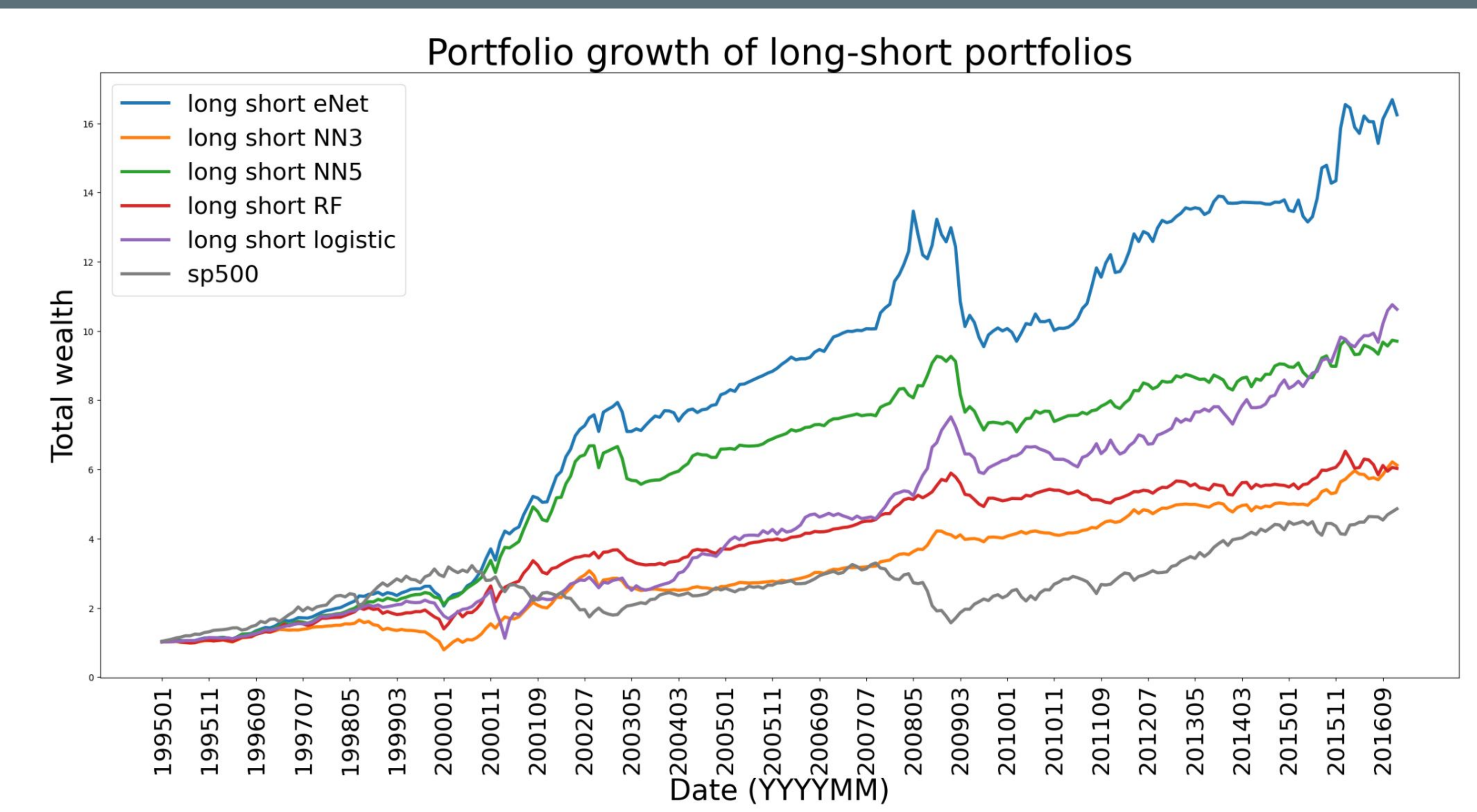
ANALYSIS & RESULTS

- Portfolio overall return:

$$r_p = \frac{\sum_i r_i Q_i}{\sum_i Q_i} - \frac{\sum_j r_j Q_j}{\sum_j Q_j}$$

- Long-only strategy: purchasing only the top 10% stocks (i-indexed)
- Long-short strategy: purchasing top 10% stocks while selling bottom 10% (j-indexed)





- With our predictions on excess returns, we used the two asset allocation strategies as described above.
- In 2008, the drops in long-short portfolio values were less pronounced than the drops in long-only.

CONCLUSION

- The long-only strategy using logistic regression stood out as the top performer (initial investment of $1 in 1995 grew to $43 by 2016 as compared to S&P 500 which grew to $4.5 by 2016).
- Elastic net performed the best among models using long-short strategy (growing to approximately $17 by 2016).
- Our work was limited to using R^2 as the primary metric for hyperparameter tuning. Future work can consider using the Sharpe Ratio.
- Our work only utilized the immediate cross-sectional feature. Future work can consider fitting a sequence model such as RNN to take into account of previous observations.

REFERENCES

- Matteo Bagnara. Asset pricing and machine learning: A critical review. *Journal of Economic Surveys*, 2022. Doi: 10.1111/joes.12532.
- Turan G. Bali, Robert F. Engle, and Scott Murray. *The CRSP Sample and Market Factor*. John Wiley & Sons, Inc., 2016.
- Eugene F. Fama and Kenneth R. French. The cross-section of expected stock returns. *The Journal of Finance*, 47(2):427–465, 1992.
- Shihao Gu, Bryan T. Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *SSRN Electronic Journal*, 2018. doi: 10.2139/ssrn.3281018.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- Narasimhan Jegadeesh and Sheridan Titman. Momentum strategies.*The Journal of Finance*, 48(3):979–1007, 1993.
- M I Jordan. Serial order: a parallel distributed processing approach. technical report, june 1985-march 1986. 5 1986. URL https://www.osti.gov/biblio/6910294.
- Daniel Poh, Bryan Lim, Stefan Zohren, and Stephen Roberts. Building cross-sectional systematic strategies by learning to rank. *SSRN Electronic Journal*, 2020. doi: 10.2139/ssrn.3751012.
- William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442, 1964. doi: https://doi.org/10.1111/j.1540-6261.1964.URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1964.tb02865.x.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, Jan 2014. ISSN 1532-4435.