Advancing Asset Pricing Models with Machine Learning: A Comprehensive Analysis from 2013 to 2024

Group Project Report #2 for AFM 423

Binghong Chen, Wenjie Shen

April 21, 2024

Table of Contents

Introduction

In the rapidly evolving field of financial analytics, machine learning has emerged as a important tool for studying complex asset pricing dynamics. This report builds upon the foundational work of Gu, Kelly, and Xiu (2019), *Empirical Asset Pricing via Machine Learning*, which harnesses machine learning techniques to predict asset risk premia. Our study extends this approach by incorporating the latest market data and introducing a binary variable to account for the economic impact of the COVID-19 pandemic. By integrating this variable, we aim to enhance the predictive accuracy of asset returns. This report details our methodology, from data preprocessing to model evaluation, and presents a comparative analysis of machine learning methods in predicting asset prices, underscoring the significant role of modern data anomalies, such as COVID-19, in financial modeling.

Methodology

Data Description

This study uses data sourced from Bloomberg, focusing on monthly records from August 30, 2013, to March 29, 2024, for 25 stocks selected from the Dow Jones Indices. We excluded five stocks from the financial sector due to incomplete data, aiming to maintain data integrity. The analysis features 20 variables, informed by a review of relevant literature. These variables encompass a range of financial metrics including momentum indicators like 1-month momentum, volatility measures such as 30-day volatility, and traditional financial metrics like the price-to-equity ratio and market capitalization. This selection is designed to capture a comprehensive snapshot of each stock's performance and market behavior, facilitating a nuanced analysis of asset pricing dynamics using machine learning techniques.

Data Processing

The dataset is divided into three distinct periods to ensure comprehensive model evaluation:

- Training Set (August 30, 2013 - March 31, 2022): Used to train the models, with an extended timeframe to incorporate substantial pre- and post-COVID-19 outbreak data, enhancing the models' ability to learn pandemic-related patterns.
- Validation Set (April 1, 2022 - March 31, 2023): Helps in fine-tuning the models and selecting the best parameters without overfitting.
- Testing Set (April 1, 2023 - March 29, 2024): Serves to evaluate the models' performance on unseen data, ensuring the effectiveness of the predictive algorithms.

Records with missing values are excluded to maintain data integrity and model efficiency. Also, for the sake of regression models, the factor variable (ID) is transformed into dummy variables to facilitate numerical analysis.

Models Used

To analyze patterns in our dataset, we employ a series of machine learning algorithms:

- Ordinary Least Squares Regression (OLS): With traditional variable selection via Variance Inflation Factor (VIF) to address multicollinearity (Gale, 1960).
- LASSO (Least Absolute Shrinkage and Selection Operator): A generalized linear model that

enhances feature selection and model simplicity (Tibshirani, 1996).

- Random Forests (RF): An ensemble learning method known for its robustness and accuracy in handling complex datasets (Breiman, 2001).
- Neural Networks (NN1 to NN5): Configurations with one to five hidden layers, designed to capture non-linear relationships in data (Rumelhart, Hinton, & Williams, 1986).

Training and Validation

Models are trained on the training set and validated using the validation set. This process ensures that each model is optimized for both accuracy and generalizability before final testing.

Performance Evaluation

Model effectiveness is assessed using three key metrics. The Average Prediction Squared Error (APSE) derived from the testing set evaluates prediction accuracy. Portfolio performance is analyzed through two strategies: a long-only portfolio, constructed from the top five stocks predicted by each model, and a long-short portfolio strategy, which involves going long on the top-predicted stocks and short on the bottom-predicted stocks. The returns from long-only portfolios are also compared against an equal-weighted portfolio as benchmarks.

Variable Importance

To identify key variables influencing the predictions, we implemented model-specific methods for assessing variable importance:

- OLS and LASSO: The importance of variables is evaluated based on the absolute values of their coefficients. Larger absolute values indicate a greater impact on the model’s predictions, highlighting the variables' significance in determining asset prices .
- Random Forest (RF): Variable importance is assessed using the built-in importance function. This method calculates the contribution of each variable to the homogeneity of the nodes and leaves in the tree, reflecting how critical a variable is for the model's decision-making process.
- Neural Networks (NN): We analyze the first layer's weights to determine variable importance. This involves studying the magnitude of weights assigned to each input variable, where larger weights signify a stronger influence on the network’s predictions.

These tailored approaches ensure a comprehensive analysis of how each variable affects the models' outputs, enabling better interpretation and validation of the predictive models used in the study.

Results

APSE (Average Prediction Squared Error)

The analysis of APSE indicates that the Random Forest (RF), LASSO, and Ordinary Least Squares (OLS) models achieve superior performance compared to the Neural Network (NN) models. Among the NN models, NN3 and NN4 show better performance than their counterparts NN1, NN2, and NN5. The APSE rankings are as follows:

Rank	Method	APSE
1	RF	0.005909925
2	LASSO	0.005913862
3	OLS	0.006046081
4	NN3	0.009313400
5	NN4	0.009774450
6	NN1	0.009643970
7	NN2	0.010703060
8	NN5	0.016760685

Table 1: APSE Rankings for Various Machine Learning Models.

Portfolio Performance

Despite the higher APSE, the portfolio suggested by the NN5 model was the only one to generate a positive return, contrasting sharply with all other model-generated portfolios. Notably, while the RF model exhibited the lowest APSE, its corresponding portfolio performed the worst among those evaluated. The equal-weighted (EW) portfolio significantly outperformed all model-suggested portfolios. The returns are summarized as follows:

Rank	Method	Return
1	**EW**	**0.850323651**
2	**NN5**	**0.005909434**
3	**NN3**	**-0.067138299**
4	**NN4**	**-0.068705607**
5	**NN2**	**-0.092337787**
6	**OLS**	**-0.107053793**
7	**LASSO**	**-0.153641523**
8	**NN1**	**-0.162836901**
9	**RF**	**-0.176753339**

Table 2: Comparison of Returns from Model-Suggested Portfolios.

Long-Short Portfolio Performance

In an evaluation of long-short portfolio strategies, models were tasked with selecting their top-five and worst-five performing stocks. Except for NN1, all models generated positive returns from their long-short strategies, with NN models (NN2 to NN5) outperforming others in terms of return difference between the best and worst predicted portfolios. These results underscore the potential utility of using a long-short strategy based on machine learning predictions:

Method	Top 5 Return	Worst 5 Return	Difference
NN5	0.005909434	-0.2244395	0.230348934
NN4	-0.068705607	-0.2275748	0.158869193
NN2	-0.092337787	-0.2391308	0.146793013
NN3	-0.067138299	-0.2108188	0.143680501
OLS	-0.107053793	-0.2952407	0.188186907
LASSO	-0.153641523	-0.2636148	0.109973277
RF	-0.176753339	-0.2312554	0.054502061
NN1	-0.162836901	-0.1277849	-0.035052001

Table 3: Performance of Long-Short Portfolios Based on Model Predictions.

Variable Importance

Our analysis of variable importance across various machine learning models shows that volatility measures, notably 30-day volatility, and momentum indicators, such as 36-month momentum, are consistently identified as important variables.

Contrary to our expectations, stock-specific factors like the stock identifier (ID) and the COVID-19 variable did not emerge as influential in our models.

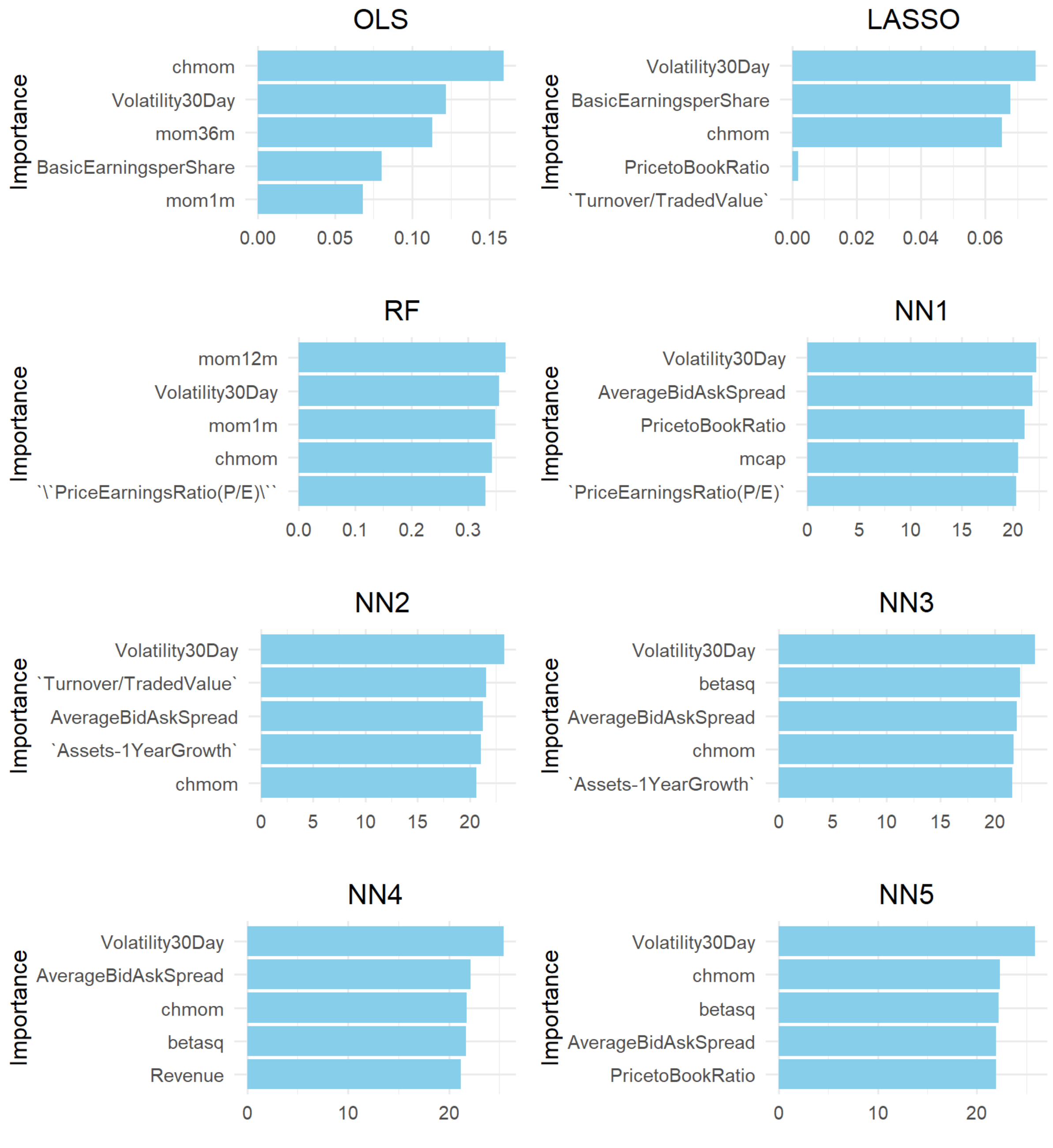

Figure 1: Variable Importance across Different Machine Learning Models.

Discussion

Our findings indicate that all model-selected portfolios for the long-only strategy underperformed compared to the equal-weighted portfolio. This unexpected result can be attributed to several factors:

1. Data Completeness: During data processing, we excluded all entries with missing values, which significantly reduced the dataset size. This likely resulted in a loss of critical information, preventing the models from capturing the full pattern of the dataset.

2. Insufficient Variables: The inclusion of only 20 explanatory variables may have been inadequate for the machine learning models to fully understand the complexities of financial data, which often features high inter-variable correlations. This limitation might have constrained the amount of information the models could leverage during learning.

3. Limited Dataset Size: The dataset spanned from 2013 to 2024 and included only 25 stocks, which is considerably smaller in both duration and scope compared to the foundational study that covered from 1957 to 2016 and nearly 30,000 stocks. The limited data could restrict the generalizability and robustness of the models.

4. Economic Conditions - Interest Rates: The significant rise in interest rates from 0.08 in January 2022 to 5.33 in August 2023 presented a unique challenge. The majority of the training data precedes 2022, while the validation and testing sets follow this period, potentially limiting the models' ability to adapt to the rapid changes in economic conditions.

Despite these challenges in the long-only strategies, the models excelled in the long-short strategies. Almost all models effectively identified stocks that outperformed or underperformed, suggesting that while they may not have captured the full pattern of the data, they successfully learned enough to distinguish between stocks with differing return potentials.

Interestingly, although Neural Network (NN) models showed lower prediction accuracy compared to OLS, LASSO, and RF in APSE metrics, their portfolio performances were superior. This aligns with the findings from the foundational study, where NN models, particularly NN4 and NN5, demonstrated robust performance. This observation underscores the potential of NN models to capture complex, nonlinear interactions that may be missed by more traditional approaches.

Also, our variable importance analysis indicated that market-wide indicators such as volatility and momentum variables significantly influenced asset pricing over individual stock identities or pandemic-related impacts during the analyzed period. The unexpected minimal impact of the COVID-19 variable may reflect the market's adaptability to abrupt economic changes. These insights demonstrate the predominant role of broad economic factors in asset pricing and the potential misalignment between theoretical assumptions and actual market behavior.

Conclusion

In this study, we built upon the work of Gu et al., applying their methodology to the latest market data and incorporating the impact of COVID-19. Contrary to our initial hypotheses, we discovered that market-wide factors such as volatility significantly overshadowed the influence of stock-specific identifiers and the pandemic. The success of long-short strategies and the exceptional performance of Neural Network models demonstrated the power of machine learning to decipher complex relationships within financial data, despite the inherent limitations posed by market variability and dataset constraints.

Moving forward, enriching the dataset with a greater number of stocks and extending the analysis to cover a longer time frame could provide more depth and stability to our conclusions. Incorporating additional metrics, such as trading volume, could enhance the models' ability to capture market sentiment and liquidity. Furthermore, experimenting with advanced modeling techniques like autoencoders offers a promising direction to potentially uncover subtler patterns in the data.

These enhancements could fortify the robustness of future models, ensuring that they not only reflect the multifaceted nature of financial markets but are also equipped to adapt to unprecedented economic events.

Appendix-1 Variables used in Analysis with explanations

Variable	**Description**
date	Date of the observation
LastPrice	Last recorded stock price
Turnover/TradedValue	Turnover or traded value of the stock
CurrentSharesOutstanding	Current shares outstanding
AverageBidAskSpread	Average bid-ask spread
Volatility30Day	30-day historical volatility
OverridableAdjustedBeta	Overridable adjusted beta
PriceEarningsRatio(P/E)	Price to earnings ratio
Revenue	Company revenue
Assets-1YearGrowth	Assets growth over the last year
BasicEarningsperShare	Basic earnings per share
PricetoBookRatio	Price to book ratio
NetWorkingCapitalInvestment	Net working capital investment
FinancialLeverage	Financial leverage
GrossProfit	Gross profit
mom1m	1-month momentum
mom12m	12-month momentum
mom36m	36-month momentum
chmom	Change in 6-month momentum
R1M_Usd	Future return, our dependent variable
betasq	Beta squared
ID	Stock identifier
mcap	Market capitalization
covid	COVID-19 variable (0 = before 2020, 1 = from 2020 onwards)

Reference

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Gale, David. (1960). *The theory of linear economic models.* -. McGraw-Hill.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature (London)*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *73*(3), 273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x