Evaluation of Stock Features Using Machine Learning Techniques

Lo Siu Fai | Chinese University of Hong Kong | November 22, 2025

1. Introduction

The 'data mining' game started when CAPM was proposed in 1960s (Sharpe, 1964; Lintner, 1965) - the first factor model of return. And then, Fama and French (1992) proposed the famous 3 factor-models with factors including 'market return', 'size' (SMB), and 'book-to-market ratio' (HML). This sparked an extensive search for factors—often termed the "zoo of factors" (Cochrane, 2011)—that can explain the cross-section of expected stock returns.

2. Descriptions of Machine Learning Models and Prediction Power

Traditional Ordinary Least Squares (OLS) regression faces several limitations in high-dimensional financial data (Gu, Kelly, & Xiu, 2020):

1. Multicollinearity: features may interact with themselves, resulting in a lot of redundant features as well as large variance in coefficients.
2. Over-fit: OLS minimize the in-sample bias but often with the cost of large out-of-sample variance.
3. Interpretability: it's usually hard to identify the true drivers among many redundant factors.
4. Feature Selection: OLS lacks a built-in mechanism to exclude irrelevant features.

To address these issues, we employ three shrinkage methods popularized in the context of "Statistical Learning" (Hastie, Tibshirani, & Friedman, 2009):

1. **Ridge regression**: add a penalty term to the Mean Squared Error (MSE) argument to shrink the size of some coefficient, but keep all features. The penalty effect is controled by a constant that multiplies the penalty term – L2. [Result: The minimum MSE is 0.9754, corresponding to alpha (lambda) = 0.1]

$$\hat{\beta}^{ridge} = arg \min_{\{\beta_0,\ \beta\}} \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

2. **Lasso regression**: similar to Ridge, but allows coefficients to be zero, i.e. can remove some features. The penalty effect is controlled by L1 (Tibshirani, 1996). [Result: The minimum MSE is 0.9709, corresponding to alpha (lambda) = 0.001.]

$$\hat{\beta}^{lasso} = \arg \min_{\{\beta_0,\ \beta\}} \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

3. **Elastic Net regression**: combination of the above two, but instead of controlling L1 and L2, we usually control the 'sum L'=L1+L2 and the 'L1-ratio'=L1/L (Zou & Hastie, 2005). [The minimum

MSE is 0.9705, corresponding to alpha (lambda) = 0.001, L1 Ratio = 0.1]

$$\hat{\beta}^{enet} = \arg\min_{\{\beta_0,\ \beta\}} \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2$$

3. Economical Interpretation of Features

1. evm: Multiple of Enterprise Value to EBITDA

Coefficient: -0.016905 | Absolute Impact: 0.016905 | Direction: Negative relationship with returns

Economic Interpretation: Higher EVM predicts lower returns, implying that companies with elevated EV/EBITDA ratios should be avoided. It is important to invest in companies that have reasonable prices relative to their operating earnings.

2. opmbd: Operating Income before Depreciation as a Fraction of Sales

Coefficient: 0.015270 | Absolute Impact: 0.015270 | Direction: Positive relationship with returns

Economic Interpretation: Companies with higher operating profit margins are expected to deliver superior returns since this factor reflects the company's economic moat and pricing power. It suggests seeking companies with strong and sustainable operating margins.

3. equity_invcap: Common Equity/Invested Capital

Coefficient: -0.014278 | Absolute Impact: 0.014278 | Direction: Negative relationship with returns

Economic Interpretation: Higher equity proportion in capital structure predicts lower returns. Over-reliance on equity financing may dilute returns, while reasonable debt leverage with careful risk management can increase capital efficiency. This may also indicate immature companies with less established credit capacity. Therefore, we prefer companies with balanced capital structures.

4. debt_ebitda: Total Debt/EBITDA

Coefficient: 0.013623 | Absolute Impact: 0.013623 | Direction: Positive relationship with returns

Economic Interpretation: This factor reinforces the economic interpretation of factor three from a debt perspective, suggesting that higher debt levels yield higher predicted stock returns. This could be due to efficient use of financial leverage which enhances shareholder returns. However, it is worth noting that excessive leverage should be avoided.

5. ps: Multiple of Market Value of Equity to Sales

Coefficient: 0.013390 | Absolute Impact: 0.013390 | Direction: Positive relationship with returns

Economic Interpretation: Higher price-to-sales multiples predict better performance, which seems counterintuitive to value investing. However, we should take growth expectations into consideration because the market pays a premium for return potential. Therefore, valuation naturally varies with revenue quality and growth prospects, which positively correlate with returns.

Conclusion: We believe these factors generally correspond with well-known trading strategies, such

as Warren Buffett's value investment approach. Value investing considers profitability (opmbd), overvaluation (evm), and financial health (debt_ebitda and equity_invcap), which is consistent with four out of the five most important factors in our model.

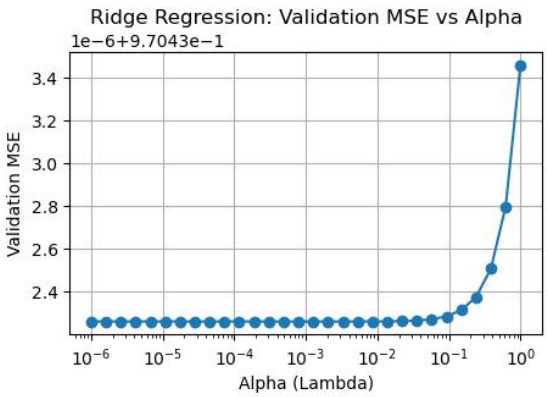
Conclusion: These findings largely resonate with the quality-value framework popularized by investors like Warren Buffett, focusing on profitability, valuation, and financial health (Asness, Frazzini, & Pedersen, 2019)

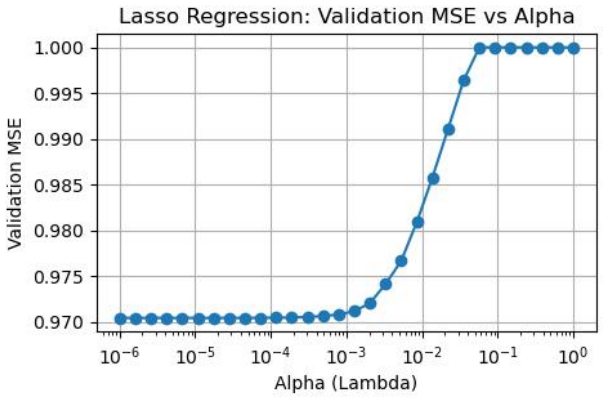
6. Discussion about the Limitation of the Models

The best alphas for all of the three models are the smallest in the selected range, implying that:

Ordinary Lest Square (OLS) linear regression may be the best fitting model: since the OLS already minimizes the sum of squared error (SSE), any penalty term will enlarge the SSE. This is common when the "signal-to-noise ratio" in financial time series is extremely low (Israel, Kelly, & Moskowitz, 2020). In this situation, the "whisper" of the stock market is so faint that aggressive regularization (high lambda) will suppress the few valid signals along with the noise.

To verify the market's characteristic as described by Israel et al. (2020), the Ridge and Lasso regressions were re-evaluated given alphas in a log-space with smaller lower bound (10e-6). The validation results (see the charts below) demonstrate a monotone decrease in MSE as the alpha goes to zero. This empirical evidence implies that under a highly efficient market, the bias introduced by adding penalty terms outweighs the shrinkage of variance in prediction. Therefore, OLS regression is considered the best predicting model for this data set.





7. Future Studies

The signals from the fundamental features in the financial markets are usually obscured by their noise and require more sophisticated studies, including but not limit to:

1. Non-linear study: use models such as random forest and XGBoost to captures the complex interconnections between features (Gu, Kelly, & Xiu, 2020).

2. Dynamic Alpha: implementing rolling-windows study to account for alpha decay and regime switching (McLean & Pontiff, 2016).

Reference

Asness, C. S., Frazzini, A., & Pedersen, L. H. (2019). Quality minus junk. *Review of Accounting Studies*, 24(1), 34-112.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, 66(4), 1047-1108.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, 47(2), 427-465.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223-2273.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Science & Business Media.

Israel, R., Kelly, B. T., & Moskowitz, T. J. (2020). Can machines learn stock returns? *AQR Capital Management Working Paper*.

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1), 5-32.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.