FAMA 70 Factor Model using Modern Machine Learning Techniques

***Abstract*—In this paper we investigate the superior machine learning method for the prediction of the American stock return base on its past data. The algorithmic trading space is rich with new strategies. Companies have spent billions in infrastructures, research, and development to be able to jump ahead of the competition and beat the market. Still, it is well acknowledged that the buy and hold strategy is able to outperform many of the algorithmic strategies, especially in the long-run. However, finding value in stocks is an art that very few mastered, can a computer do that? We developed a method called FAMA 70 and found the surprising high performance of non-linear models.**

***Index Terms*—Market Forecasting, Feature Selection, Feature Investigation**

I. Introduction

The stock market is a complex and chaotic dynamic system with both the characteristics of randomness and systematic component. And since the market is not a perfect ideal environment only made of rational people, the noises that cannot be explained by traditional economic theories which have the assumption that people are all rational should be addressed. Meanwhile, there are many different kinds of factor could be used to forecast the stoke value. The field of financial forecast has been working on developing methods that could efficiently catch the hidden relationship of those factors. Varies Artificial Neutral Networks(ANN) have been developed in order to achieve better results. To analyze the data across a relatively long period of time, Real-time recurrent learning[2] was the most simple method, however, its constant back-propaganda action makes it extremely insufficient to deal with long term market with many factors. Thus, Long Short-Term Memory[1] has been introduced as a common method, which utilizes its special gate mechanism, and removes the unnecessary gradient to produce a highly efficient but also low error rate result. Also, the Support Vector Machine[3] is a classic classification method able to produce a robust result by learning to split the high-dimensional feature space as far as possible. The active need from the quantitative analysis field is an example of the importance of the study in ML financial forecast and the need to develop methods with higher performance.

II. Literature

The Capital Asset Pricing Model[4] is one of the fundamental theories in the modern economy to predict the capital market, and it suggests using only one factor, the systematic risk β, to predict the stock return with the regression method. Fama-French Three-Factor Model[5] is an expansion on the CAPM, while noticing the deficiency of only using β as a factor, the Fama-French Three-factor model added two extra factors,(i)market capitalization and (ii)book-to-market ratio, to make the model more well rounded. Later, the Fama-French Five-factor Model[6] has been developed in addition to the original three-factor model by adding the two factors: (i) profitability and (ii) investment, to fill the drawbacks of the original method. In general, CAPM, Fama-French Three-Factor Model, and Fama-French Five-Factor Model are all linear regression methods with different number of factors.

A. Lasso, Ridge, and Elastic Net

While considering machine learning methods, linear models are commonly used, aside from the simple linear regression, regularized regression is a more advanced method with higher performance and fewer weak points. The Lasso[7] is a classic L1-regularized linear model with an effective effect on dealing with data with multicollinearity. Such as:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{p} w_j \times x_{i_j})^2 + \lambda \sum_{j=0}^{p} |w_j|$$

For a data set with M instances and p features, the Lasso adds a penalty for weight with large magnitude compare to the Ordinary Least-Square(OLS) cost function. Because the penalty is base on the magnitude of the weight coefficient, thus, it could reduce some coefficients to zero and produces a sparse model, so it helps feature selection encountering a huge number of features. The Ridge[8] is an L2-regularized linear regression, similar to the Lasso, it also added a penalty to the OLS cost function as such:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{p} w_j \times x_{i_j})^2 + \lambda \sum_{j=0}^{p} w_j^2$$

The Ridge adds the square of the weight coefficient as the penalty, compare to OLS and the lasso, it has a high efficiency at reducing over-fitting, however, it does not reduce the coefficients to zero. While each Lasso and Ridge has its own advantages, the Elastic Net[9] combines the two and formed as such:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{p} w_j \times x_{i_j})^2 + \lambda_1 \sum_{j=0}^{p} w_j^2 + \lambda_2 \sum_{j=0}^{p} |w_j|$$

With both the L1 and L2 regularization embedded into the cost function, the elastic net has both features from the two methods.

B. Decision Trees and Random Forest

For nonlinear models, the decision trees are the fundamental elements. It is a tree structure split by information entropy and a simple fundamental model could be easily interpreted. The random forest[10], as its name suggests, creates many trees with each one learns from a piece of the sample that is randomly selected through the bootstrapping method. And averaging each tree's prediction in the end. Such a bagging method could have a better result against the noise compare to one single decision tree.

III. Contributions

The FAMA 3/5 models have their own limitations, and we optimized them in mainly two ways. On the one hand, we added more factors to tune the model; on the other hand, instead of using only linear regression like what the FAMA 3/5 models did, we use both linear & non-linear machine learning methods. The model that we developed using 70 factors and much more sophisticated machine learning methods is named FAMA 70 factor model.

IV. Structure

In section V we describe the strategy we use to clean the raw data and the methodology to choose the super-parameters for the models. In addition to that, we also show our approach to test the models we got. In section VI we present the computational result. Specifically, in section VI.A, we present the comparison of the four linear models. In section VI.B, we have our two nonlinear model results compared. In section VII, we discuss the results among all the models and show an unusual finding. We also did a feature investigation on the linear and nonlinear models to explain the possible reasons for our finding.

V. Methodology

We use various fine-tuning methods to find the most optimized parameter for each model. An advanced testing method is applied in order to value the performance. The dataset we are using is created by Carbone[11] who publicly released this dataset on Kaggle. The code and the dataset could be addressed at https://github.com/Deemocean/FAMA-70. The computation is powered by a single RTX 3090 with various run times when training different models.

A. Data Processing

For the purpose of reducing outliers, we investigated shares with unusual gains, as we set the threshold to a 500% increase, we plot those "top-gain" stocks individually. And we found some stocks have a flat increase curve which is natural for the market, and that indicates the high possibility of these stocks being outliers due to mistyping in the process of data collecting. We also realized the existence of 0-value and missing value in the data. For entries of data with more than 5% Nan or 0-value, we drop it, otherwise, we fill in with the average value.

B. Testing Methodology

To avoid over-fitting, we used the cross-validation method. Specifically, we split the data into five folds of training&testing sets(with a ratio of 8 to 2) Because the samples are not balanced in terms of the state of increase, we made each fold has the same percentage as the initial data. For the fine-tuning of the λ weight coefficient for the L1 and L2 regularization and other parameters, we did grid search cross-validations to find the most optimized values.

VI. Computational Results

In this section we first present results on Linear Models, followed by the performance of the nonlinear models. And we also did a feature investigation on the models as well.

A. Linear Model Results

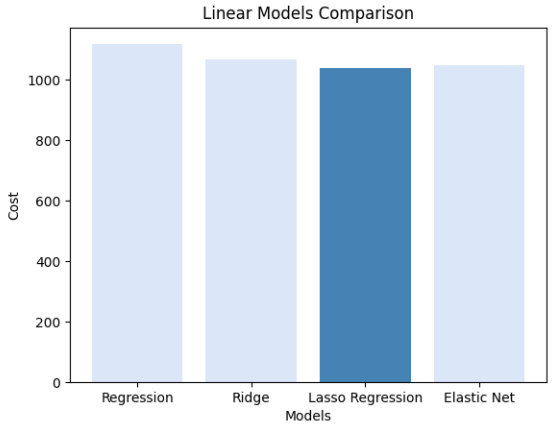

Fig. 1. Cost Comparison of Linear Models

Figure 1 presents the results on linear models including simple linear regression with OLS, the Ridge, the Lasso, and the Elastic Net, with a score of 1117.037, 1067.968,**1039.405**, and 1048.039, respectively. We noticed that the Lasso regression has the lowest cost result as representing the best-performed model among all the Fama-70-factor linear models.

B. Non Linear Model Results

As shown in Figure 2, Among the two nonlinear models Random Forest behaves the best with the lowest cost score of **913.519** compare to Decision Tree's 1008.848.

VII. Discussion

As shown in Figure 3, among all the models, decision trees and random forest having the lowest cost score which indicates the nonlinear models generally outperform the linear models including OLS, Ridge, Lasso, and Elastic Net. And we also noticed the surprising high performance of the decision trees. Consider decision trees has a relatively simple structure compare to Ridge, Lasso, and Elastic Net. In theory, those more complex regression methods should have a better result. As the simple nonlinear based regression method results better

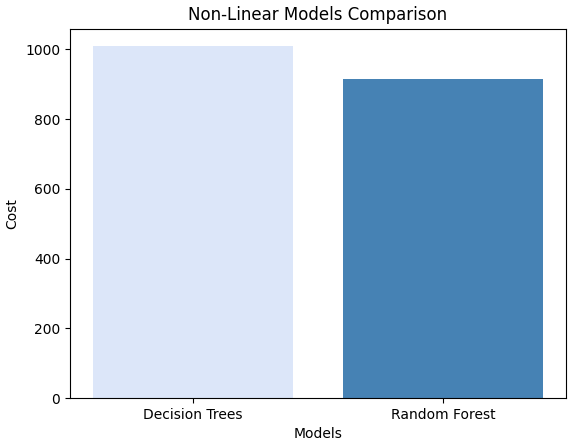

Fig. 2. Cost Comparison of Non-Linear Models

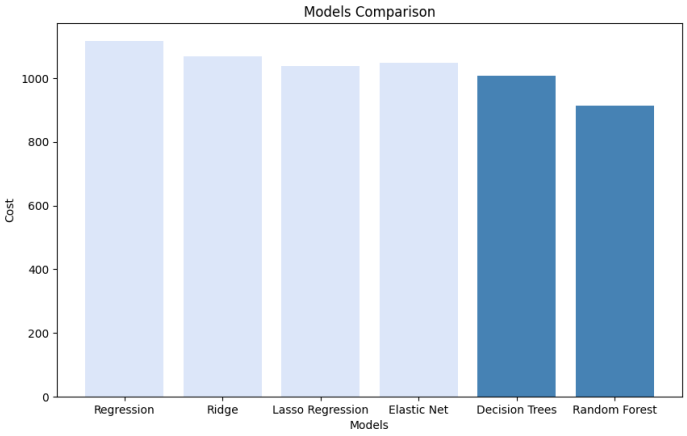

Fig. 3. Cost Comparison of All Models

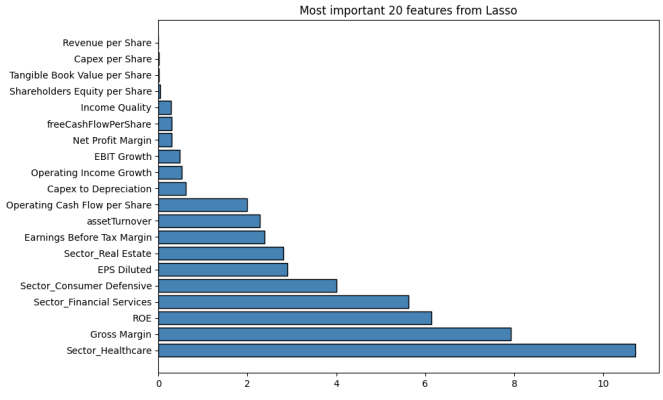

Fig. 4. Feature importance of the Lasso Model

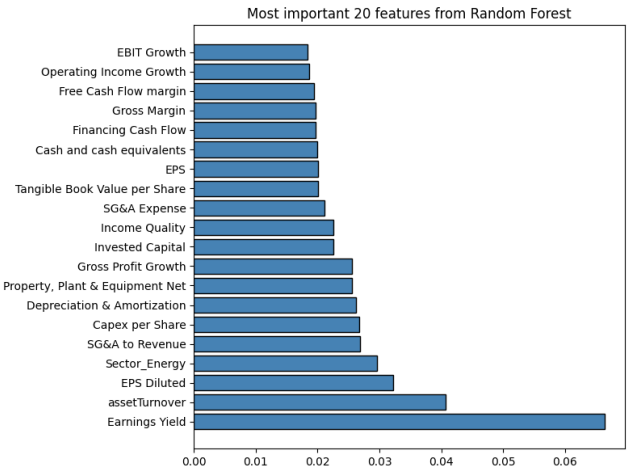

Fig. 5. Feature importance of the Random Forest Model

than complex linear based regression methods, it indicates the existence of feature interactions and the nonlinear characteristics in the data. Figure 4 and Figure 5 give the top 20 important features selected by the Lasso Model(The best behaved linear model) and the Random Forest Model(The best behaved nonlinear model). Even though the data itself is linear, but the two models have very different feature selection results, which further proves the non-linearity existence in the stock market indicates. While current models are generally being linear, it is vital to capture all the richness of financial data by using nonlinear models.

VIII. Conclusion

The stock market has its patterns while having noises and randomness. There are many linear forecast models are being used to predict the market, but most of them are based on linear models. However, as the result we got from our experiment, the nonlinear method outperformed those linear models, which suggests the nonlinearities in the market. Meanwhile, nonlinear models are often criticized for the reason of being hard to interpret compare to linear models, however, using the feature importance plots, we can identify the critical features as obvious as investigating linear models. Thus, there is a considerable richness for developing nonlinear methods for ML trading strategies.

References

[1] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[2] Robinson, A. J., and Frank Fallside. The utility driven dynamic error propagation network. Cambridge, MA: University of Cambridge Department of Engineering, 1987.

[3] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.

[4] Sharpe, W.F. (1964), CAPITAL ASSET PRICES: A THEORY OF MARKET EQUILIBRIUM UNDER CONDITIONS OF RISK*. The Journal of Finance, 19: 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

[5] Fama, Eugene F.; French, Kenneth R. Common Risk Factors in the Returns on Stocks and Bonds. Journal of Financial Economics. 1993, 33 (1): 3–56. doi:10.1016/0304-405X(93)90023-5.

[6] Fama, Eugene F., and Kenneth R. French. ”A five-factor asset pricing model.” Journal of financial economics 116.1 (2015): 1-22.
[7] Tibshirani, Robert. ”Regression shrinkage and selection via the lasso.” Journal of the Royal Statistical Society: Series B (Methodological) 58.1 (1996): 267-288.
[8] Hoerl, Arthur E., and Robert W. Kennard. ”Ridge regression: Biased estimation for nonorthogonal problems.” Technometrics 12.1 (1970): 55-67.
[9] Zou, Hui, and Trevor Hastie. ”Regularization and variable selection via the elastic net.” Journal of the royal statistical society: series B (statistical methodology) 67.2 (2005): 301-320.
[10] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
[11] Carbone, Nicolas. (2020, Jan.). 200+ Financial Indicators of US stocks (2014-2018), Version 1. Retrieved 2021, March from https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018.