Portfolio Optimization and Modeling Analysis for Portfolio Return

Jiantao Lei[1, *, †], Bowen Xiao[2, *, †], Yufei Xue[3 *, †]
[1*]Jlei3@albany.edu, [2*]Bowen.Xiao@bayes.city.ac.uk, [3*]Yxue16@alumni.jh.edu

[1]Collage of Arts&Science University of Albany Albany, the United States,[2]Bayes Business School, City, University of London London, the United Kingdom,[3]Carey Business School Johns Hopkins University Washington DC, the United States
[†]These authors contributed equally.

Abstract—This paper mainly focuses on two parts – portfolio optimization and modeling. The theory of efficient frontier and Sharpe ratio are used to optimize and select the portfolio. The paper's portfolio used to do further research is the frontier portfolio with the most significant Sharpe ratio. This paper also provides an analysis and evaluation of the significance of the features in the Fama-French 5-factor model by applying a series of machine learning models and comparing the Sklearn score. Based on the conclusions of the 5-factor model analysis, this paper also quantifies the impact of the pandemic, develops a 6-factor model and a new 3-factor model, and compares them with the Fama-French 5-factor model. The result shows that the 3-factor model is better than the other two models.

Keywords-Portfolio optimization, Machine learning, Factor model, pandemic

1 INTRODUCTION

According to the Global Financial Stability Report announced in April 2020, the COVID-19 pandemic has a significant impact on the financial systems, and further escalation of the crisis could affect the global financial stability [1]. Commenting on the impact of the pandemic in the financial markets, the report said, "Equity markets experienced the fastest drop in history …." [1]. In the UK, the pandemic also caused a sharp decline not only in the gross domestic product (GDP), 4.8% below February 2020 monthly GDP levels during December 2020 [2], but also had a blow to many companies. Some companies even went bankrupt. According to Wind Financial Terminal (WFT), European equities plunged more than -25%, the worst quarter since the financial crisis [3]. However, unprecedented central bank easing that started with the 2008 credit crisis and jumped in 2020 to offset Covid 19 pandemic has driven stock and bond prices higher and interest rates to record low levels.

Since stock investment is a significant activity in the financial market, portfolio construction and the prediction of stock prices are two inevitable topics. To optimize the portfolio, the theory of mean-variance analysis is a wide-using method. The other topics, stock market prediction

ICEMME 2022, November 18-20, Nanjing, People's Republic of China

DOI 10.4108/eai.18-11-2022.2327134

and stock selection, are two more complex and enduring analysis topics. The goal of stock market prediction is to determine the future movement of a stock value of a financial exchange [4]. Stock market prediction can reflect the level of prosperity of the overall economy. For investors, an accurate return prediction can help to develop their investment strategy and earn more profit. Predicting stock returns offers enormous chances for profit and is a significant motivation for research in this area; knowledge of stock movements by a fraction of a second can lead to high profits [5]. However, although accompanied by huge potential profits, stock return prediction is always challenging. In terms of the fundamental analysis, investors try to find the intrinsic value by looking at factors such as price-to-earnings ratio (P/E), price-to-book-value (P/B), and financial report analysis, which is very time consuming and laborious, and even sometimes misleading because many aspects must be considered, and the firm's potential manipulation can lead to a wrong valuation. Quantitative finance relatively explains how actors can calculate completely independently because the device replaces social cues [6]. This paper is mainly from quantitative finance, by looking for features, training models, and then testing. The Capital Asset Pricing Model (CAPM) is one of the most famous pricing models. The CAPM describes the relationship between risk and expected return used in the pricing of risky securities [7]. However, in 1993, Nobel laureate Eugene Fama and Kenneth French found that the stock return is based on a combination of market risk and firm size and value, which is known as the Fama French 3 Factors model. Fama and French argue that the "common habit" of using the CAPM to evaluate the portfolio performance and to estimate the cost of capital should be broken: "The CAPM is wanted, dead or alive." said Fama and French [8]. Twenty years later, in 2013, based on the three-factor model, they proposed a five-factor model. This new model took investment and profitability into consideration.

Based on the theoretical framework derived from the efficient market theory (EMT) and rational expectations intertemporal asset pricing theory (Chen et al., 1986; Merton, 1973), stock prices always fully reflect all available information, and whenever a particular asset is influenced by systematic economic news, no extra reward can be earned by bearing diversifiable risk [9]. Thus, we believe the COVID-19 pandemic can be a significant feature in predicting the stock market. A new model containing the pandemic as a factor of analysis is necessary for describing the stock market during 2020 and 2021. This paper develops a new model based on the original Fama French 5 factor model by adding relevant factors and deleting some that are not applicable based on the evaluation.

The contributions in this paper include: 1) The ability of the Fama-French 5-factor model in explaining stock returns in the UK stock market can be tested in this paper. 2) A new model that includes the factor of COVID-19 pandemic can be developed to explain and predict stock returns before the influence of pandemic on the stock market goes to a level that has not been considered. 3) Although the pandemic is still updating its effects on the economy, this paper can provide a recent and baseline conclusion to future research. 4) This paper can provide a systematic and statistical conclusion on how the pandemic influences the stock market, which can be used for the future development of the Fama French 5 factors model.

2 METHOD

2.1 Data preparation

The data we used in this paper include stock close prices and factors. We visited the Fama-French website and found the past ten years' factor data which was created using the Bloomberg database. The factor data includes five factors. They are market risk (Mkt-RF), size effect (SMB), the value effect (HML), profitability (RMW), and investment (CMA) factor.

We also obtained the close monthly prices of the 14 UK companies from Yahoo Finance website [10] and Wind financial terminal. The considered companies are from companies with relatively higher stock prices in different industries. The reason why we covered several industries is that we want to see how the pandemic influences the overall financial market, but not only one industry.

2.2 Portfolio optimization

The central ideas for this thesis include portfolio optimization, modeling analysis, and machine learning. The study aims to build an optimal portfolio and predict its future movement by modeling. The model was evaluated in this paper, and we also tried to improve it when we found detects in this model.

Looking for an efficient frontier is the way to find the optimal portfolio in this paper. First, we calculated the stock log returns, standard deviations (SD), and covariance matrix among these stocks. Those results can be calculated by the formula (1), (2), and (3).

$$Log\ Returns = \ln\left(\frac{Price_n}{Prices_{n-1}}\right) \tag{1}$$

$$SD = \sqrt{\frac{(r_1 - \bar{r})^2 + (r_2 - \bar{r})^2 + (r_3 - \bar{r})^2 + \cdots + (r_n - \bar{r})^2}{n}} \tag{2}$$

$$\Omega = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \dots \\ \sigma_{2,1} & \sigma_2^2 & \dots \\ \dots & \dots & \dots \end{bmatrix} \tag{3}$$

Where means the return of the stock at time period n; r¯is the risk-free rate; Ω represents the covariance matrix, and σ in formula (3) represents the covariance between each two of the considered stocks.

Then, to maximize the portfolio return, we drew an efficient frontier and chose the portfolio with the biggest Sharpe ratio. Sharpe ratio is the excess return divided by the SD of each stock. It can be calculated by the formula (4). In formula (4), "excess return" means the return on the next day minus return on the day before, and "standard deviation" is the SD calculated in a formula (2).

$$Shapre\ Ratio = \frac{Excess\ Return}{Standard\ Deviation} \tag{4}$$

2.3 Modeling

After portfolio optimization, we built a linear model based on the theory of the Fama-French 5-factor Ordinary least-squares (OLS) model, which is an expansion upon the three-factor Fama-French OLS model showing the portfolio's exposure to the five factors mentioned in session 2.1, i.e., how much of the portfolio returns are explained by these factors and therefore how much excess return it generates. The formula of Fama-French 5-factor model is shown in formula (5)

$$\begin{gathered} R_{it} - R_{Ft} = a_i + b_i(R_{Mt} - R_{Ft}) + \\ s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMA_t + e_{it} \end{gathered} \tag{5}$$

Where R_{it} is the return in month t of one of the portfolios; R_{Ft} is the risk-free rate; $(R_{Mt} - R_{Ft})$is the return spread between the capitalization-weighted stock market and cash [11].

When using the OLS model to estimate the coefficients of features, there will be some potential problems of such as overfitting or too large coefficients. Three models are used to solve the problems and improve the OLS model: Ridge regression model, Lasso model, and Multitask Elastic net model.

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where independent variables are highly correlated [12]. The Ridge coefficients can be estimated by formula (6).

$$\left|\left|X\omega - y\right|\right|_2^2 + \alpha\left|\left|\omega\right|\right|_2^2 \tag{6}$$

Lasso is a regression analysis method that performs both variable selection and some regularization to enhance the prediction accuracy and interpretability of the resulting statistical model. The Lasso coefficients can be estimated by formula (7).

$$min_{\omega} \frac{1}{2n_{samples}} \left|\left|X\omega - y\right|\right|_2^2 + \alpha\left|\left|\omega\right|\right|_1 \tag{7}$$

The Multi-task Elastic-Net model is an Elastic-Net model that allows fitting multiple regression problems jointly enforcing the selected features to be the same for all the regression problems, also called tasks [13]. The objective function to minimize is formula (8).

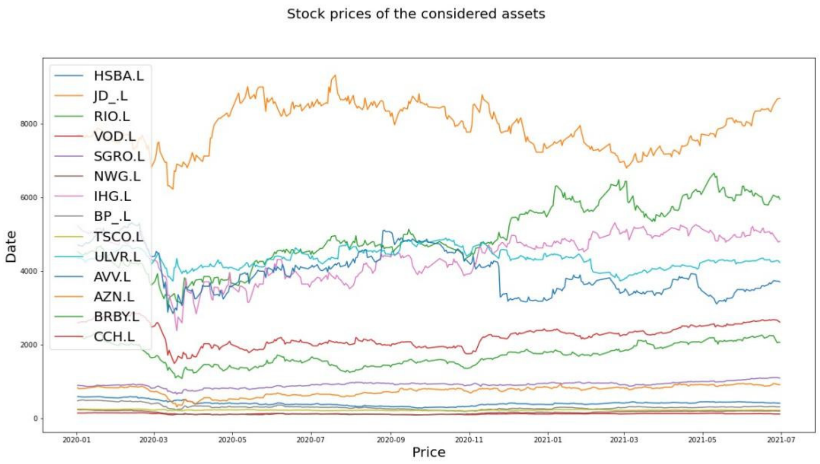

Figure 1. Stock prices of the considered assets from Jan.2020 to Jul.2021.

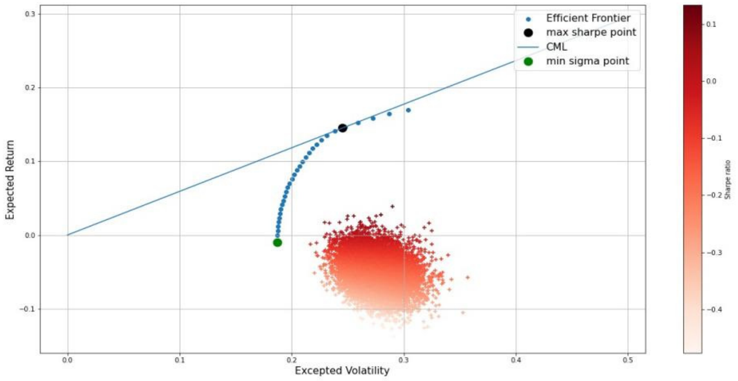

Figure 2. The mean-variance plot

$$min_{\omega} \frac{1}{2n_{samples}} \left|\left|XW - y\right|\right|_{Fro}^{2} + \alpha\rho\left|\left|W\right|\right|_{21} + \frac{\alpha(1-\rho)}{2} \left|\left|W\right|\right|_{Fro}^{2} \tag{8}$$

Sklearn scores of each model can be generated to compare the performance of the models. The higher the score is, the better the model performs to the data.

3 RESULTS AND DISCUSSION

In this paper, the mean-variance analysis is used to optimize the portfolio. Fama-French 5 factors model is used in the prediction of stock price and has been tested. The pandemic factor is calculated and added to the factors. An updated Fama-French 5 factors model is applied to improve the accuracy of stock price prediction. Some machine learning models have been done to evaluate the performance of the models.

Figure 1 shows the price trend of the 14 considered stocks from January 2020 to July 2021. Figure 1 shows that most of the considered stocks have fallen sharply during March 2020, which was exactly the start of the pandemic in Europe. During the next year, the stock prices of most considered stocks were slowly recovering. The price of AstraZeneca (AZN), which can be found as the orange line in the plot, shows that it recovered much faster than other stocks. This is because AZN is one of the most famous medical companies in the world. Since the pandemic led to a huge requirement of medical devices, the price of AZN was even higher than that in 2019, before the pandemic. From Figure 1, an assumption that the pandemic is an indispensable factor in the prediction of stock prices in recent years can be proposed.

3.1 Results of portfolio optimization

The plot of mean-variance analysis is shown in Figure 2. The red part represents the portfolio volatility versus the average return of 500,000 simulated random portfolios. More simulations are supposed to be performed to make the upper edge of the red area form a curve. However, because of the limitation of Python, 500,000 simulations are shown here. Investors are assumed to be risk-aversion which means they like mean-returns and do not like portfolio risk, which we measured as standard deviations. The blue points form a curve above the red area, which is drawn using minimize function. This curve is formed by a set of frontier portfolios. It shows the maximum return an investor can get for a set level of volatility or the volatility/risk that an investor needs to bear to earn a certain level of returns. The green point represents the portfolio with the lowest possible risk. Generally speaking, the higher the return investors want to gain, the greater the risk they need to undertake. The black point represents the portfolio with the maximum Sharpe ratio, which is the optimal portfolio that is required for further research and analysis later in this paper.

From the analysis above, Table 1 shows the optimized portfolio investment weight of the 14 considered stocks. Figure 3 shows the distribution of the considered stocks in the pie chart.

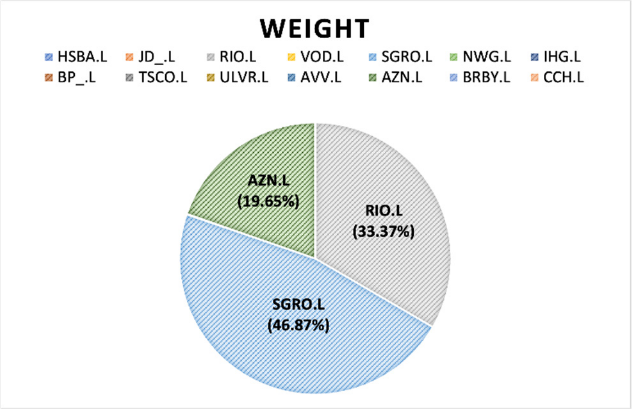

Figure 3. Pie chart of portfolio weights

Three of the 14 stocks are effectively invested, and other stocks are almost not in consideration. Only stocks of RIO, SGRO, and AZN can be seen in the pie chart. Based on the portfolio weight in Table 1, the optimal portfolio is constructed, and the expected return is 14.516%. Since most of the stocks are not invested in the portfolio, different models and factors are necessary to find a more reasonable result.

TABLE 1. THE OPTIMIZED PORTFOLIO INVESTMENT WEIGHT OF 14 CONSIDERED STOCKS

Stock	Weight	Stock	Weight
HSBA.L	0.01%	**BP_.L**	0.01%
JD_.L	0.01%	**TSCO.L**	0.01%
RIO.L	33.374%	**ULVR.L**	0.01%
VOD.L	0.01%	**AVV.L**	0.01%
SGRO.L	46.870%	**AZN.L**	19.647%
NWG.L	0.01%	**BRBY.L**	0.01%
IHG.L	0.01%	**CCH.L**	0.01%

3.2 Results of Machine Learning

Three machine learning models are selected to improve the model: Ridge regression model, Lasso, and Multi-task Elastic-Net. 80% of the data are trained in the training set, and 20% of the data are to be tested in the testing set. The whole machine learning process is done using the sklearn module in Python.

Table 2 shows the performance of these three models. In order to evaluate how well or bad the model is to the data, the score provided in the sklearn is used. The larger the score is, the better the model acts on the data.

TABLE 2. SKLEARN SCORE OF THE THREE MODELS

Model	Score
Ridge regression model	0.09164
Lasso	0.71966
Multi-task Elastic-Net	0.18368

3.2.1 Ridge regression model

The Ridge regression model is a widely used regularization method that can help to avoid model overfitting and make the coefficients smaller by imposing a penalty on the size of the coefficients. The penalty consists of a constraint α and the l2-norm of the coefficient vector to the ordinary least squares (OLS) when estimating the coefficients. Ridge can provide more realistic and reliable coefficients. However, Ridge has little helped in reducing insignificant features.

3.2.2 Lasso

Contrary to the limitations of Ridge, Lasso helps researchers reduce features with a very small coefficient by adding a penalty that consists of a constraint α and the 1-norm of the coefficient vector to the OLS. Similar to Ridge, Lasso can also prevent overfitting to some certain extent.

3.2.3 Multi-task Elastic-Net

The Multi-task Elastic-Net has both advantages of the Ridge model and Lasso. It is an elastic network model used to jointly estimate the sparse coefficients of multiple regression problems. The constraint is that all tasks have the same selected characteristics.

Table 2 shows that the Lasso model has the best performance. Thus, we decided to use the Lasso model for analysis later. The coefficients corresponding to the five factors in the original Fama French 5 factors model are shown in Table 3.

TABLE 3. LESSO MODEL COEFFICIENTS OF 5 FACTORS

Mkt-RF	SMB	HML	RMW	CMA
0.00833	-0.00076	0	0	0

Lasso model helps eliminate insignificant factors, which are the factors with coefficient 0. Besides eliminating three insignificant factors, HML, RMW, and CMA, we plan to add another factor, which represents the influence of COVID-19 on the financial market

3.3 Results of models

We minus the Financial Time Stock Exchange (FTSE) of 2019 to that of 2020 and then divide it to 1000. It can be seen as an estimate of how much the pandemic influences the financial market, so-called the factor data of the pandemic. Table 4 shows the comparison of R-squared of the Lesso model for the original five factors (Mkt-RF, SMB, HML, RWM, CMA), Lesso model for six factors, including COVID-19, and Lesso model for three factors, deleting insignificant factors mentioned above and including COVID-19.

TABLE 4. SKLEARN SCORES OF 5, 6 AND 3 FACTORS LESSO MODEL

Number of factors	R-squared
5	0.71966
6	0.73027
3	0.74052

The sklearn score after adding the factor of a pandemic is about 0.02, larger than the five factors Lasso model. And three factors Lasso model has the highest score among the three.

TABLE 5. COEFFICIENTS OF 3 FACTORS LESSO MODEL

Mkt-RF	SMB	Covid	Alpha
0.00828	-0.00293	-0.00246	0.000239

Given the fact that the coefficient of the Covid factor is significant and the sklearn score is highest, these mean that the using Lasso model to estimate the coefficients of the new three factors, Mkt-RF, SMB, and COVID-19, can best explain the stock returns and make the further prediction.

4 CONCLUSION

In this paper, 14 British stocks are selected to find a good stock price trend prediction model. The considered stocks are companies with relatively high market capitalization and relatively well-known in different industries. According to the construction of efficient frontier and calculation results of portfolio Sharpe ratio, the optimal portfolio is mainly composed of three stocks, RIO, SGRO, and AZN, with their weights accounting for 33.374%, 46.870%, and 19.647%, respectively.

Fama-French 5 factor model is established based on the optimal portfolio. When estimating the coefficients, three models of machine learning are used: the Ridge regression model, Lasso, and Multi-task Elastic-Net. Sklearn gives Lasso the highest score: 0.7197, which means compared with the other two models, the coefficient estimated by Lasso better explains the stock returns. However, among the five factors, three of them are not significant - nearly zero. Therefore, considering the initial five factors, the preliminary results of model construction are not reasonable.

Among the major events in the past year, the pandemic has had the most significant impact on the world economy. The Covid factor is calculated by subtracting FTSE 2019 from FTSE 2020 then divided by 1000. After the analysis of the 3-factor model, 5-factor model, and 6-factor model, it is easy to conclude that the best model for stock trend prediction is the three factors Lesso model. The three factors include excess market return (XSMKT), size (SMB), and COVID-19.

REFERENCES

[1] COUNCIL F S. Financial Stability Report[J]. 2020

[2] Office of National Statistics website: https://www.ons.gov.uk/economy/grossdomesticproductgdp/articles/coronavirusandtheimpactonoutputintheukeconomy/december2020

[3] Reuters Website: https://www.reuters.com/article/idUSL4N2LM276

[4] Alzazah F S, Cheng X. Recent Advances in Stock Market Prediction Using Text Mining: A Survey[J]. E-Business-Higher Education and Intelligence Applications, 2020.

[5] Gupta A, Dhingra B. Stock market prediction using hidden Markov models. In: 2012 Students Conference on Engineering and Systems. IEEE; 2012. pp. 1-4

[6] Beunza D, Stark D. From dissonance to resonance: Cognitive interdependence in quantitative finance[J]. Economy and society, 2012, 41(3): 383-417.

[7] Rehman A, Baloch Q B. Evaluating Pakistan's Mutual Fund Performance: Validating through CAPM and Fama French 3-Factor Model[J]. Journal of Managerial Sciences, 2016, 10(1).

[8] Suh D. Stock returns, risk factor loadings, and model predictions: A test of the CAPM and the Fama-French 3-factor model[M]. West Virginia University, 2009.

[9] Khan K, Zhao H, Zhang H, et al. The impact of COVID-19 pandemic on stock markets: An empirical analysis of world major stock indices[J]. The Journal of Asian Finance, Economics, and Business, 2020, 7(7): 463-474.

[10] Yahoo finance: https://finance.yahoo.com

[11] Fama French 5 factors pricing model: https://blog.quantinsti.com/fama-french-five-factor-asset-pricing-model/

[12] Hilt, Donald E., and Donald W. Seegrist. Ridge, a computer program for calculating ridge regression estimates. Vol. 236. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station, 1977.

[13] Multitask elastic net model: https://www.tutorialspoint.com/scikit_learn/scikit_learn_multi_task_elastic_net.html