ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 11 (2021) No. 4, pp. 47-64
DOI:10.33168/JSMS.2021.0403

The Relationship Between the Rate of Return and Risk in Fama-French Five-Factor Model: A Machine Learning Algorithms Approach

Bui Thanh Khoa[1,2], Pham The Son[1,2], Tran Trong Huynh[3]

[1]University of Information Technology, Ho Chi Minh City. Vietnam

[2]Viet Nam National University, Ho Chi Minh City, Vietnam

[3]FPT University, Ho Chi Minh City, Vietnam

khoadhcn@gmail.com (corresponding author); sonpt@uit.edu.vn; huynhtt4@fe.edu.vn

Abstract. This study aimed to apply Support Vector Regression (SVR), Ridge Regression (RR) and Lasso Regression (LR) algorithms to the Fama-French five-factor model, which includes market (Mkt), size (SML), value (HML), profitability (RMW) and investment (CMA), in order to explain fluctuations in expected returns of diversified portfolios. The study examined the stock market in the US from July 1963 to September 2021. The stocks are grouped into ten portfolios by industry and separated into two phases. In phase 1, this study selected the optimal parameters for the algorithms SVR, RR and LR. In phase 2, the researchers used the optimal parameters obtained in phase 1 to set up four forecasting models using four different algorithms: SVR, RR, LR and OLS (Ordinary Least Squares). The rolling window approach is used to generate forecasts. Consequently, the Lasso Regression algorithm produces the smallest average Root Mean Squared Error (RMSE); however, this difference is not statistically significant through the F-test.

Keywords: Machine learning, fama-french five-factor model, svr, ridge, lasso

1. Introduction

The Capital Asset Pricing Model (CAPM) was introduced in the early 1960s by CAPM quantifies the correlation between an asset's systematic risk and the asset's expected rate of return in a holding period. The capital asset pricing model is based on the portfolio investment theory of Markowitz (1952). The regression formula is as follows:

$R_{it} - R_{ft}$ = a_i + $(R_{mt} - R_{ft})$ + e_{it} (1), where: R_{it} is the expected return of asset i in period t, R_{mt} is the expected return of the market in period t, and R_{ft} is the risk-free rate

Since then, CAPM has become a widely used instrument in portfolio risk management. When investing in a risky asset, investors utilise the CAPM to determine the minimum rate of return. CAPM's simplicity and ease of use for estimation are two of its notable features. However, CAPM uses too many assumptions, and these assumptions are difficult to meet in practice. As a result, some criticisms about the empirical validity of the CAPM have arisen. Roll (1977) was one of the first to criticise CAPM. He argues that since it is challenging to construct a portfolio that includes all assets traded in the capital markets, the CAPM has no practical value. A few years later, Banz (1981) discovered the size effect, which indicates that some small firms, in terms of market capitalisation, seem to earn a higher average return than other large-scale ones.

In the following years, E. Fama and French (1992) studied several factors affecting return rate, including beta coefficient, size, financial leverage, P/E ratio, B/M ratio, and concluded that the size and B/M ratio have the most significant influence on the return rate of listed companies in the US. Eugene F. Fama and French (1993) proposed a three-factor model as follows:

$Rit - R\mathrm{ft} = ai + bi\ (R\mathrm{mt} - R\mathrm{ft}) + siSMBt + hi\mathrm{H}M\mathrm{L}t + eit$ (2), where: $SMBt$ is the return on the diversified small-cap portfolio minus the large-cap one, $SMBt$ is the return on a high B/M portfolio minus the low B/M one.

A test of the three-factor model carried out in the US market from 1963 to 1990 shows that the three-factor model explains better than the previously proposed CAPM (R^2 coefficient of the three-factor model is approximately 90% while CAPM is only 70%). In 2015, Fama and French continued to propose a five-factor model based on the previous three-factor model and added two factors, including profitability (RMW) and investment (CMA) (Eugene F Fama & French, 2015). The model's formula is as follows:

$Rit - R\mathrm{ft} = ai + bi\ (R\mathrm{mt} - R\mathrm{ft}) + siSMBt + hi\mathrm{H}M\mathrm{L}t + riRMWt + ciCMAt + eit$ (3), where RMWt is the return of the robust operating profitability portfolio minus the weak operating profitability portfolio, CMAt is the difference of return between the conservative investment portfolios and the aggressive investment portfolios.

Fama and French then examined model (3) in the United States from July 1963 to December 2013 and found that it was more effective than the three-factor model, with the R2 coefficient ranging between 71% and 94% for different portfolios. Non-linear expansion in asset pricing models has been the focus of research in recent years. For example, Dittmar (2002) applied the non-linear pricing kernel technique to estimate CAPM. Gogas et al. (2018) have applied support vector regression (SVR) under the framework of CAPM, Fama-French three and five-factors and concluded that SVR is more effective than the Ordinary Least Square method (OLS). Chen et al. (2019) used deep learning to estimate the individual securities' returns and proved that it is more effective than the linear factor model.

One of the limitations of the previous studies is the method of dividing the data set. Due to the characteristics of the time series, the model's parameters also change from time to time. Therefore, the rolling window method should be employed to predict time series. Additionally, the duration of the historical data should also be considered in light of the sample's representativeness. This study uses the rolling window approach with a historical data length of five years.

2. Literature review

2.1. The Fama-French Five-Factor Model

The Fama-French three-factor model describes the relationship between expected return rate, size (market capitalisation), and B/M ratio. In 1933, the three-factor model surpassed the CAPM model as it explained some components that the CAPM could not. The three-factor model is tested by time-series regression of formula (2); the estimated coefficients b_i, s_i and h_i should be statistically significant; the intercept coefficient a_i expected to be statistically insignificant for portfolios *i*.

Novy-Marx (2013); Titman *et al.* (2004) proved that the model is imperfect for explaining the volatility of expected return rate due to the lack of elements coming from the profitability and investment. This evidence is the driving force for Fama-French's research. Eugene F Fama and French (2015) have upgraded the three-factor model by adding two factors related to profit and investment. The Fama-French five-factor model is described in Formula (3). The b_i, s_i, h_i, r_i, and c_i coefficients measure the sensitivity of volatility in expected return rate toward volatility in respective factors. The intercept a_i is expected to be 0 for portfolios *i*.

The experimental results of Fama-French in the US market show that the GRS test does not support the five-factor model, but the obtained R^2 is relatively high, ranging from 71% to 94% (Gibbons *et al.*, 1989). Cakici (2015) used historical data from June 1992 to December 2014 to examine a Fama-French five-factor model in 23 developed stock markets. The author compares its performance to that of three-factor and four-factor models in explaining the returns of portfolios. Research findings indicate that the five-factor model in the North American, European, and Global markets is similar to that in the US; however, the two new factors (profit and investment) are not statistically significant in the Japan and Asia Pacific markets. Martinsa and Eid Jr (2015) assessed the Fama-French five-factor model in the Brazilian market from

January 2000 to December 2012 and found it is more useful than the three-factor model. Foye (2018) evaluated whether the Fama-French five-factor model can better describe the emerging markets equity returns than the three-factor model. The author examined a sample of 18 countries from three different regions between December 1996 and June 2016. Furthermore, this research examines the performance of the five-factor model across a variety of emerging markets. As a result, the five-factor model consistently outperforms the three-factor model in Eastern Europe and Latin America.

2.2. Support Vector Regression

Support Vector Machine (SVM) is a supervised learning algorithm that solves the data classification problem (Cortes & Vapnik, 1995). The idea of SVM is to map the original data set to a high dimensional space by the mapping Φ, which is convenient for data classification. SVM computes an optimal hyperplane (H) from the training data set. Assume that X is a matrix of independent variables and Y is a categorical variable vector $(y_i \in \{-1,1\})$. Therefore, the hyperplane is represented by the following equation: $a^T\Phi(x_k) + b = 0$. Please assume that the input data can be perfectly separable; then, by adjusting the suitable parameters, this study can transform the problem so that the shortest distance to (H) is always equal to 1 on both sides. Thus, the SVM problem is to determine the model's parameters a and b.

Considering particular observation i, if $a^T\Phi(x_i) + b \geq 1$, $y_i = 1$, on the other hand, if $a^T\Phi(x_i) + b \leq -1, y_i = -1$. An equivalent way is as follows: $y_i[a^T\Phi(x_i) + b] \geq 1 (4)$. Minimising $\|a\|$ and b under constraint (4), this study obtain the optimal parameters of the classification model. Cortes and Vapnik (1995) proposed the classification condition as $class(x_i) = sgn(a^T\Phi(x_i) + b)$. Because the assumption on the existence of a perfectly separated hyperplane (H) is unrealistic, Cortes and Vapnik (1995) proposed to add a soft margin that allows some misclassified observations. The constraint becomes: $y_i[a^T\Phi(x_i) + b] \geq 1 - \xi_i (5)$. Then, the SVM problem becomes $\min_{w,b}\left(\frac{1}{2}\|w\|^2 + C\sum_{k=1}^{N}\xi_k\right), \xi \geq 0 (6)$ with the constraint (5), where C is a hyperparameter in the classification model. Continuing to transform (5) under Wolfe (1961): $\min_{\alpha}\left(\frac{1}{2}\alpha^T M\alpha - e^T\alpha\right)$, where, the function $K(x_i, x_j) = \Phi(x_i)^T\Phi(x_j)$ is called the kernel function and $y^T\alpha = 0, 0 \leq \alpha_i \leq C, e^T = [1 \quad 2 \quad \dots \quad N]$, M is the square matrix with components $m_{ij} = y_i y_j K(x_i, x_j)$. The classification result will be based on the following equation (1):

$$class(x_i) = sgn(a^T\Phi(x_i) + b) = sgn(\textstyle\sum_{i=1}^{n} y_i\alpha_i K(x_i, x) + b) \qquad (1)$$

With the same idea as the SVM algorithm, the SVR algorithm is also implemented similarly, except that the dependent variable is a continuous variable that takes on a real value. However, according to Patel *et al.* (2015); Qu and Zhang (2016), instead of finding the hyperplane as in (5), the SVR algorithm develops a regression function $f(x, a) = a^T x + b$. A boundary ε is introduced as (2):

$$|y - f(x,w)|_\varepsilon = \begin{cases} 0, |y - f(x,w)| \leq \varepsilon \\ |y - f(x,w)| - \varepsilon, |y - f(x,w)| > \varepsilon \end{cases} \tag{2}$$

SVR method is to minimise the R by ε and $||w||^2$ in the following equation (3):

$$R = \frac{1}{2}||w||^2 + C(\sum_{i=1}^{N}|y - f(x_i,w)|_\varepsilon) \tag{3}$$

2.3. Ridge Regression

A method called Ridge regression may be used to estimate multiple-regression model coefficients in situations when the independent variables are closely linked (Hilt & Seegrist, 1977). It has a variety of applications in various areas, including econometrics, engineering and chemistry (Gruber, 2017), which was proposed the idea for the first time in 1970 (Hoerl & Kennard, 1970a, 1970b). This method was the culmination of decade-long research into the topic of Ridge analysis. In the case of linear regression models with multi-collinear independent variables, Ridge regression was developed to overcome the imprecision of least square estimators. It is possible to get a more accurate estimation of ridge parameters by creating an RR that has lower variance and mean square error than the prior least square estimators (Jolliffe, 2011).

In conventional linear regression, $n \times 1$ column vector y is projected onto the column space of the $n \times p$ design matrix X, whose columns are highly correlated. The ordinary least squares estimator of the coefficients $\beta \in R^{p\times 1}$ by which the columns are multiplied to get the orthogonal projection $X\beta$ is: $\hat{\beta} = (X^TX)^{-1}X^Ty$, where X^T is the transpose of X.

By comparison, the ridge regression estimator is $\widehat{\beta_{rıdge}} = \left(X^TX + kI_p\right)^{-1}X^Ty$, in which I_p denotes the $p \times p$ identity matrix, and *k* is small.

2.4. Lasso Regression

Lasso was independently created in 1986 in geophysics literature, building on past work that employed the penalty for fitting and penalising the coefficients. Based on Breiman's nonnegative garrote, statistician Tibshirani independently revisited and popularised (Tibshirani, 1996). Before Lasso regression, Stepwise selection was the most popular approach for selecting covariates. Only in some situations, such as when a few variables have a major impact on the result, can this method improve prediction accuracy. Other times, it may worsen the accuracy of the prediction.

To improve prediction accuracy, ridge regression was considered the most popular method at the time. Reducing the sum of the squares of regression coefficients to be smaller than a predetermined value reduces overfitting and hence decreases prediction error using a ridge regression. There is no covariate selection, hence the model's interpretability is not improved by it. Both of these objectives may be achieved using Lasso, which constrains the sum total of the absolute values of the regression coefficients to less than the fixed value, thereby eliminating certain coefficients and avoiding their impact on prediction outcomes. Ridge regression, which similarly minimises the coefficients' size, is similar to this principle; however, Ridge Regression tends to zero out a smaller number of coefficients.

Linear regression mentions the linear relationship between the independent and dependent variables. Assuming there are k independent variables x*1*, x*2*, …, x*k* and one dependent variable y, the overall regression function has the form: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon(1)$. With data consisting of N observations, the Lasso regression method is to find the estimated coefficients $\widehat{\beta_\iota}$ by solving the optimisation problem: $\min_{\beta_0,\beta}\left(\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{k} x_{\mathrm{ij}}\beta_i\right)^2\right)$ subject to $\sum_{j=1}^{k}\left|\beta_j\right| \leq C$, where C is a hyper parameter and $\beta = (\beta_1, \ldots, \beta_k)$ is the vector of regression coefficients.

Alternatively, the matrix form is $\min_{\beta_0,\beta}(\|y - \beta_0 + X\beta\|_2^2)$ subject to $\|\beta\|_1 \leq C$, where $X = [1 x_1 x_2 \ldots x_k]$; $y = [y]$ are matrices written in columns and $\|a\|_p = (\sum_{i=1}^{n}|a|^p)$ is the standard ℓ^p norm in Euclidean space. Normalising the variables, it can rewrite as $\min_{\beta}\left(\frac{1}{N}\|y - X\beta\|_2^2\right)$ subject to $\|\beta\|_1 \leq C$ or the Lagrangian form $\min_{\beta}\left(\frac{1}{N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right)$

3. Methods

Monthly data collected from French library in the US market from July 1963 to January 2021 includes return series of Mkt, HML, RMW, SMB, CMA and ten weighted portfolios grouped by industry, including NoDur, Durbl, Manuf, Enrgy, HiTec, Telcm, Shops, Hlth, Utils, Other. In addition, the risk-free rate used is the 1-month T-bill. The following table summarises the variables used in the study (Eugene F Fama & French, 2015).

The study split data into two sets: the first covers the period from 7/1963 to 12/1898, and the second covers from 1/1990 to 9/2021. The first set is the optimal parameters for the SVR, RR, and LR algorithms. This study includes some potential parameters, as shown in Table 2.

This research chose the default degree 3 for the polynomial kernel and the epsilon of 0.2 for the radial kernel. This study has a total of 25 potential models predicting for ten portfolios. These models are filtered, leaving three models corresponding to 3 algorithms. Because financial data change over time, this study uses the rolling window method with a fixed length of five years for datasets (1) and (2), as shown in Figure 1. RMSE criterion is employed to compare forecast performance where $Y_t, \widehat{Y_t}$ is the real value and predicted value, respectively.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}{n}},$$

The chosen model with the smallest RMSE is applied in forecasting the dataset (2). The forecast results compared to the regression model. Finally, this research uses F-distribution for testing the models' performances.

Table 1. Variable description

Variable	Description
Mkt	A portfolio's excess return above the market's
HML	Differential performance of high and low B/M diversified stock portfolios
RMW	Returns on a diverse portfolio of high and low profitability equities.
SMB	Return on small-cap stock holdings less return on large-cap company holdings from diverse portfolios
CMA	Low- and high-investment company stock returns may be prudent and aggressive, depending on the investor's preference.
NoDur	The weighted return rate on portfolios of non-durable goods - Food, Tobacco, Textiles, Leather, Toys
Durbl	The weighted return rate on portfolios of durable goods - Automobile, TV, Furniture, Home appliances
Manuf	The weighted return rate on portfolios of manufacturing industries - Machinery, Trucks, Aircraft, Chemicals, Oil, gas and coal extraction and byproducts
Enrgy	The weighted return rate on portfolios of energy industries – Gas, oil, gasoline
HiTec	The weighted return rate on portfolios of high-tech industries – Computers, software and electronics
Telcm	The weighted return rate on portfolios of telecommunication industries – Television, phones and data transmission services
Shops	The weighted return rate on portfolios of commercial industries – Wholesales, retails and some services (laundry, repair shops)
Hlth	The weighted return rate on portfolios of medical industries – Healthcare, medical devices and pharmaceuticals
Utils	The weighted return rate on portfolios of social utility industries
Other	The weighted return rate on portfolios of other industries – Mining, transportation, hospitality, entertainment, finance

Table 2. Potential parameters of the models.

RR	LR	SVR
k = 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4 and 0.5	λ = 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4 and 0.5	kernel function: linear, radial, poly (polynomial) cost: 0.1, 0.5, 1, 5, 10

4. Results

4.1. Descriptive statistics

The results in Table 3 showed that most portfolios have a higher average excess return than the market portfolio (Mkt), except for the Utils portfolio, because the Mkt subtracted the risk-free rate, and the other portfolios did not. On the other hand, these sub-portfolios overall risk is mostly higher than that of the market portfolio, which is consistent with the risk-reward trade-off principle. As for the T-bill rate, during the Covid-19 pandemic, FED has loosened its monetary policy to stimulate the economy. Accordingly, the basic interest rate decreased significantly, sometimes to zero, during this period.

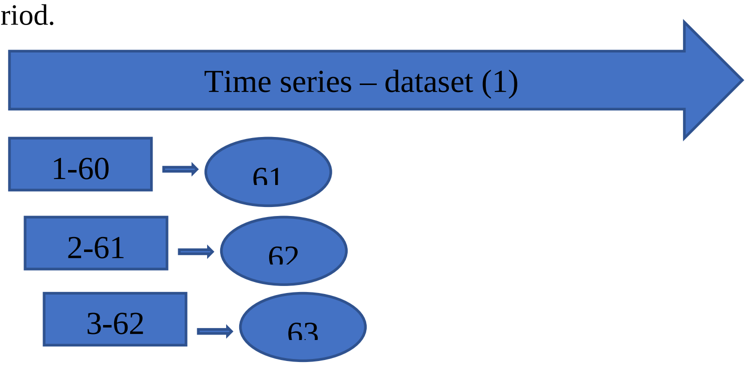

Fig. 1: The rolling window diagram.

Table 3. Descriptive statistics.

	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Mkt	0.580	4.450	-23.240	-1.950	3.415	16.100
SMB	0.236	3.035	-15.390	-1.505	2.075	18.380
HML	0.271	2.904	-14.020	-1.405	1.700	12.480
RMW	0.257	2.196	-18.760	-0.815	1.270	13.380
CMA	0.263	1.978	-6.780	-1.000	1.470	9.060
RF	0.368	0.267	0.000	0.150	0.510	1.350
NoDur	0.776	4.236	-21.290	-1.530	3.440	18.580
Durbl	0.733	6.720	-33.200	-3.080	4.185	45.000
Manuf	0.716	4.918	-27.620	-2.010	3.740	17.060
Enrgy	0.703	5.926	-34.760	-2.590	4.080	32.110
HiTec	0.825	6.340	-26.190	-2.870	4.640	20.490
Telcm	0.600	4.607	-16.490	-1.990	3.315	21.070
Shops	0.809	5.120	-28.330	-2.220	3.960	25.720
Hlth	0.819	4.802	-20.720	-2.065	3.675	29.250
Utils	0.560	4.004	-13.280	-1.690	2.980	18.570
Other	0.691	5.286	-23.870	-2.235	3.850	19.970

The location characteristics of the factors and portfolios are shown in Figure 2.

Most of the portfolios and factors distributions are symmetrical, where the Durbl and Enrgy portfolios have a wider range of values than the rest, which demonstrates an overall risk relatively higher than that of the rest. In terms of stability, the Utils portfolio is less volatile than the rest.

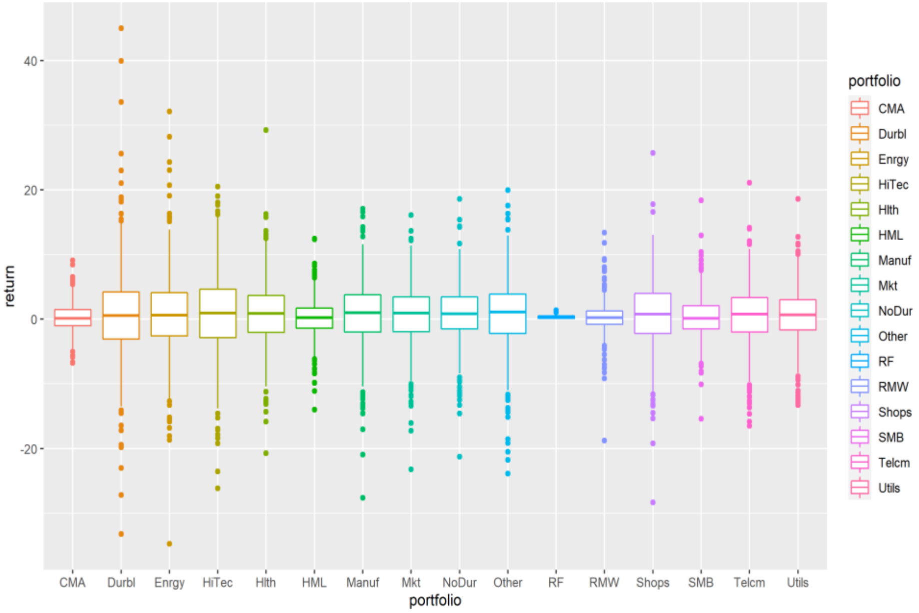

Fig. 2: Distribution of return rate of portfolios.

4.2. The correlation between the variables

Based on the correlation graph in Figure 3, the portfolios mostly correlate strongly to the explanative factors, especially the market factor (Mkt). In which, the Enrgy and Utils portfolios correlate relatively weaker than the other portfolios, in particular, Utils' correlation is statistically insignificant to the factors SMB, RMW, CMA; however, the Enrgy portfolio's correlation is statistically significant to 5 factors at 10%; the relative correlation is weak.

According to the correlation graph in Figure 3, most portfolios strongly correlate with explanatory factors, especially the market factor (Mkt). The two portfolios Enrgy and Utils, have a relatively weaker correlation than the rest. Specifically, Utils has no statistical significance with SMB, RMW, CMA; although the Enrgy portfolio has a statistically significant correlation with all five factors at 10%, the correlation is relatively weak.

In general, the correlation between factors is relatively weaker than the correlation between portfolios. Among the correlation pairs between factors, the correlation between the factor HML and CMA is the strongest (0.67) and is statistically significant at less than 1%. The remaining pairs have low correlation (with absolute values being less than 0.4).

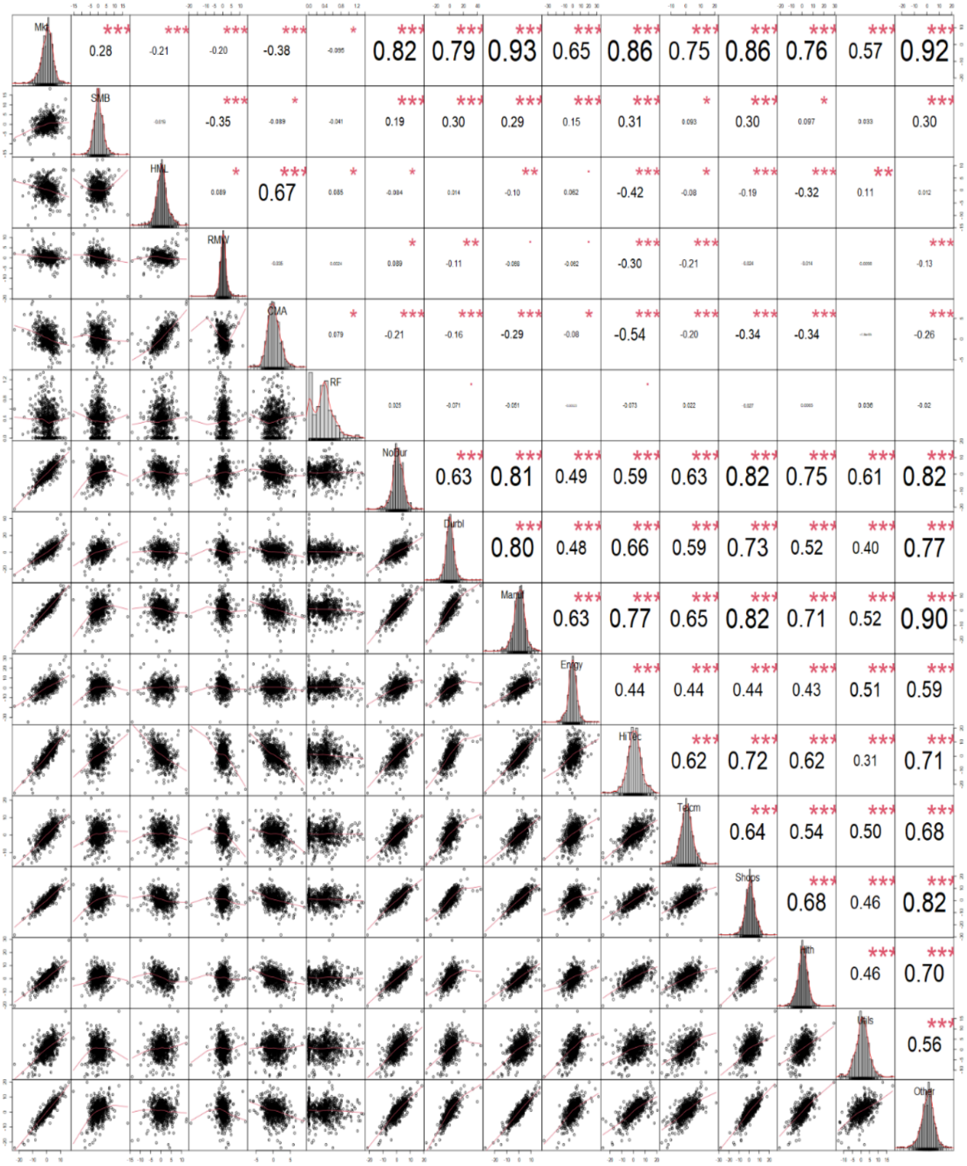

Fig. 3: Correlations between portfolios.

4.3. Forecast results of the dataset (1)

After running 25 forecasting models for ten portfolios in stage 1. Table 4 shows that the SVR algorithm with linear kernel outperforms the radial and polynomial functions. Among the input parameters, the SVR algorithm with a linear kernel and a cost of 0.5 produced the best results, with an average RMSE of 2.57 for ten portfolios. Moreover, the model has accurately predicted two portfolios, Manuf and Other, with an RMSE of less than 1.5, while the Enrgy portfolio had the highest RMSE (3.97).

Table 4. The RMSEs corresponding to the parameters of SVR models

SVR Model												
Kernel	**Cost**	**NoDur**	**Durbl**	**Manuf**	**Enrgy**	**HiTec**	**Telcm**	**Shops**	**Hlth**	**Utils**	**Other**	**average**
linear	0.1	2.04	2.85	1.56	4.01	2.61	2.95	2.8	3.11	2.65	1.39	2.6
radial	0.1	4.22	5.03	4.42	5.53	5.15	3.88	5.19	4.88	3.92	4.58	4.68
poly	0.1	5.62	7.58	6.79	5.91	8.4	5.4	7.1	6.27	3.76	6.99	6.38
linear	0.5	2	2.77	1.49	3.97	2.6	2.94	2.78	3.13	2.64	1.39	2.57
radial	0.5	3.57	4.3	3.69	5.16	4.46	3.56	4.45	4.23	3.51	3.66	4.06
poly	0.5	8.37	11.52	10.95	7.84	13.68	8.41	13.96	9.72	5.24	12.86	10.25
linear	1	2.01	2.78	1.49	3.96	2.61	2.98	2.79	3.13	2.67	1.39	2.58
radial	1	3.45	4.16	3.51	5.09	4.3	3.49	4.29	4.1	3.48	3.5	3.94
poly	1	10.05	13.81	14.2	7.81	15.52	10.63	17.17	11.7	5.97	15.5	12.24
linear	5	2.01	2.79	1.5	3.95	2.63	3.01	2.79	3.14	2.67	1.4	2.59
radial	5	3.3	4.18	3.38	5.3	4.4	3.52	4.29	4.12	3.49	3.41	3.94
poly	5	15.88	19.96	24.42	14.34	20.83	14.26	25.98	17.6	9.59	25.13	18.8
linear	10	2.01	2.78	1.5	3.95	2.63	3.01	2.8	3.14	2.67	1.4	2.59
radial	10	3.37	4.37	3.41	5.48	4.61	3.66	4.45	4.26	3.61	3.48	4.07
poly	10	20.8	24.1	29.89	15.6	23.84	15.92	31.43	22	12.89	30.56	22.71

For the RR algorithm in Table 5, the parameter k = 0.1 is most efficient with an average RMSE of only 2.21 and even lower than the SVR algorithm. RR model with k = 0.1 also predicts pretty well in two portfolios, Manuf and Other, but predicts poorly in Enrgy portfolio just like SVR model.

Table 5. The RMSEs corresponding to the parameters of RR models

RR Model											
k	**No Dur**	**Durbl**	**Manuf**	**Enrgy**	**HiTec**	**Telcm**	**Shops**	**Hlth**	**Utils**	**Other**	**average**
0.01	2.01	2.41	2.22	3.13	2.77	2.73	2.38	2.87	2.83	2.43	2.58
0.05	3.05	2.79	2.81	3.93	3.26	3.53	2.96	3.75	3.42	3.05	3.25
0.1	1.9	2	1.47	2.86	2.05	3.02	1.71	2.43	2.83	1.79	2.21
0.15	4.21	4.5	4.2	4	4.72	4.61	4.59	4.19	4.39	4.42	4.38
0.2	3.21	3.27	2.76	3.97	2.62	4.22	2.84	3.21	4.1	2.99	3.32
0.25	3.37	3.6	3.56	3.32	4.26	2.93	4.11	3.89	2.99	3.87	3.59
0.3	3	2.93	2.82	3.86	2.92	3.92	2.72	3.56	3.9	2.75	3.24
0.35	3.28	3.81	3.26	3.53	3.4	4.08	3.52	2.91	3.96	3.54	3.53
0.4	3.35	3.57	3.48	3.29	4.21	2.89	4.06	3.9	2.61	3.64	3.5
0.5	2.24	2.29	1.89	2.82	2.41	3.36	2.1	2.96	3.06	1.41	2.46

With the parameter λ = 0.1, the LR algorithm performed very well and was more efficient than SVR and RR, with the average RMSE in ten portfolios of 2.11 in Table 6. The LR model has low RMSE and predicts a low stability error for all ten portfolios. All categories have RMSE of less than three, and six out of ten portfolios have RMSE of less than 2.

Table 6. The RMSEs corresponding to the parameters of LR models

LR Model										
λ	**NoDur**	**Durbl**	**Manuf**	**Enrgy**	**HiTec**	**Telcm**	**Shops**	**Hlth**	**Utils**	**Other**	**average**
0.01	2	2.4	2.21	3.12	2.76	2.72	2.37	2.86	2.82	2.41	2.57
0.05	3.01	2.78	2.84	3.82	3.27	3.49	2.93	3.68	3.39	2.99	3.22
0.1	1.79	1.84	1.45	2.72	1.93	2.95	1.63	2.34	2.77	1.69	2.11
0.15	4.19	4.33	4.12	3.94	4.56	4.59	4.51	4.17	4.37	4.42	4.32
0.2	3.13	3.14	2.82	3.8	2.58	4.14	2.83	3.15	4.02	2.92	3.25
0.25	3.28	3.49	3.5	3.21	4.08	2.93	3.87	3.73	2.93	3.79	3.48
0.3	2.98	2.87	2.81	3.64	2.93	3.91	2.7	3.42	3.83	2.6	3.17
0.35	3.24	3.51	3.19	3.43	3.15	3.99	3.38	2.91	3.91	3.41	3.41
0.4	3.2	3.41	3.35	3.15	3.93	3.01	3.7	3.67	2.75	3.57	3.37
0.5	2.2	1.89	1.81	2.61	2.17	3.49	1.94	2.74	3.2	1.48	2.35

4.4. Forecast results of the dataset (2)

According to the forecast results in stage 1, three models are chosen for the forecast at stage 2 including: SVR with linear function, cost = 0.5, RR with k = 0.1 and LR with λ = 0.1. Stage 2 forecasting begins on 1/1990 and ends on 9/2021. At this stage, this study compare all three models with the forecasting model using the OLS method. Forecast results for ten portfolios are described in Table 7 and Figure 4.

Table 7. The RMSEs of SVR, RR, LR và OLS models

Portfolio	SVR	RR	LR	OLS
NoDur	2.241	2.199	2.184	2.216
Durbl	4.824	4.810	4.775	4.808
Manuf	1.804	1.816	1.829	1.820
Enrgy	4.808	4.908	4.861	4.936
HiTec	2.596	2.549	2.524	2.555
Telcm	3.376	3.189	3.156	3.207
Shops	2.380	2.336	2.338	2.334
Hlth	3.073	3.021	3.004	3.049
Utils	3.755	3.575	3.538	3.602
Other	1.582	1.601	1.578	1.614
average	3.044	3.000	2.979	3.014

The results in Table 7 indicated that the LR is the most effective model with an average RMSE of 2.979, while the SVR is the least effective one with an RMSE of 3.044, but the difference between these two models is relatively small. The Manuf and Other portfolios give very good RMSE results in all models; in contrast, the Enrgy and Durbl portfolios have a relatively high RMSE compared to the other portfolios. Among all the portfolios and models, the LR model most accurately predicts the Other portfolio with RMSE = 1.578, while the OLS model predicts the least accurately the Enrgy portfolio with the highest error of 4.936.

The accuracy of the models strongly correlates with each other and fluctuates across portfolios. In each portfolio, the RMSE was not significantly different. As Figure 4, all four forecasting models are pretty similar, especially the outliers.

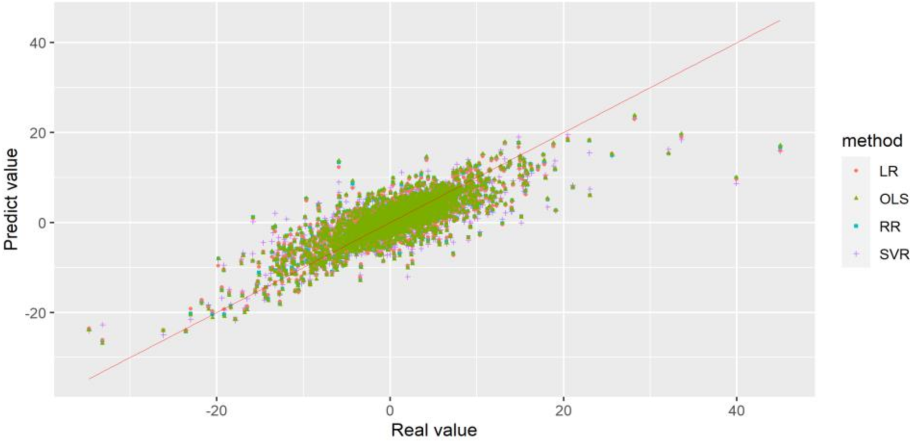

Fig. 4: Predict value v. real value.

4.5. F-Test results:

To evaluate the difference in the performance of the models, this study calculates the deviation of predictions from the real each method's value. Each model will have 321 forecasts for each portfolio, which means each model has 3210 forecasts, and in total, this study has 12840 forecasts for all four models. This study uses the null hypothesis that no difference exists between the four methods. The result of the ANOVA (Analysis of Variance) is presented in Table 8.

Table 8. One-Way ANOVA

Analysis of Variance Table

Response: test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
method	3	3	1.0237	0.1981	0.8977
Residuals	12836	66323	5.1669		

As shown in Table 9, the P-value = 0.8977 > 5%, so this study fail to reject the null hypothesis, which means that there is no difference in forecast efficiency of the models (with 5% significance). Thus, although the LR model has the lowest average RMSE, this difference is not statistically significant.

5. Discussion and conclusion

5.1. Discussion

Overfitting is a constant issue in machine learning forecast models. This study can point out this issue by comparing the forecast results from datasets (1) and (2) in Tables 5, 6, 7. 8. While the best mean errors of SVR, RR, and LR models are 2.57, 2.21, and 2.11 for dataset (1), they rise to nearly 3 for all three models when predicting for dataset (2). Again, this demonstrates that parameters change with time and that the rolling window technique is more suitable than the k-fold cross-validation method.

The explanatory factors in the Fama-French five-factor model have been effective in explaining fluctuations in the expected return of the portfolios. Except for the market factor, the correlation between factors is relatively low (Figure 3), especially the relatively strong correlation between factors and portfolio returns. All portfolios with a greater correlation to the factors forecast more accurately than those with low correlations. For example, the Manuf and Other portfolios have relatively high correlation coefficients with market factors (0.93 and 0.92, respectively), resulting in both portfolios performing very well in all four forecasting models.

One element directly affecting the RMSE of the models that this study needs to consider the outliers. The outliers appear in both the explanatory factor and the portfolio's return (Figure 2). The portfolios with a high degree of volatility, such as Enrgy and Durbl, have a significantly high RMSE compared to the rest. Moreover, some portfolios with relatively lower volatility, such as Manuf and Other, provide better RMSE results from all forecasting models. The SVR algorithm produces very different forecast results with the choice of the kernel function. The Fama-French five-factor model generates empirical values of the coefficient R2 ranging from 71% to 94%, demonstrating that the linear function is the most suitable choice (Eugene F Fama & French, 2015); therefore, the SVR model performs better with the linear kernel as a consequence.

Because the RR and LR models are dependent on the coefficients k and λ, the optimal parameter values are mostly determined by testing different values on the training data. When comparing the forecast performance of RR and LR models, this study discovers that LR is more accurate than RR in both dataset (1) and dataset (2); this result is consistent with previous research of Roy et al. (2015). The prediction results in the dataset (1) and (2) further show that when the suitable parameters are applied, the difference in efficiency between the two models RR and LR, is not considerable (Madhuri et al., 2019; Manasa et al., 2020; Yu & Wu, 2016; Zhang et al., 2019). The final results in Table 6-7 affirmed that, although LR has the lowest average RMSE, this difference is not statistically significant. Thus, there is no difference in the models' effectiveness when compared to the research results of Venkatasubbu and Ganesh (2019)

5.2. Conclusion

This study has performed forecasting return rates of diversified portfolios under the Fama-French five-factor framework. The SVR, RR, LR and OLS algorithms are used. The LR model was effective for dataset (1) with the lowest mean errors for ten portfolios. The prediction error of the SVR algorithm varies significantly depending on the kernel function used, with the linear kernel being the most efficient one.

The correlations between explanatory factors, portfolios and outliers have a significant effect on the model's error. More precisely, the portfolios with a strong correlation to the factors will have low RMSE and vice versa. As a result, the RMSE fluctuates differently for different portfolios. Finally, whereas the LR model produces the slightest error, this difference is not statistically significant. This finding implies no difference in the efficiency of all four algorithms used in the research data.

The Fama-French five-factor model is an excellent predictor of changes in expected returns of diversified portfolios. The model quantifies the linear relationship between risk and expected return. From a Machine Learning perspective, this study can forecast the portfolio returns with controlled errors by estimating the optimal input parameters. Therefore, Machine Learning should be considered as an alternative to the traditional econometric methods. The rolling window method should be considered for time series with changing characteristic parameters instead of other methods such as k-fold cross-validation to increase model reliability and avoid overfitting. Lasso regression should be considered as an alternative to the OLS. However, caution should be taken when making statistical inferences about the estimated coefficients.

The study has not considered the factors affecting the errors of the forecasting models, such as outliers, normality of the distribution. Moreover, the study is limited to only four algorithms. Further research should analyse the prediction error factors and apply various Machine Learning algorithms to choose the ideal forecasting model

Acknowledgement

This research is funded by the University of Information Technology, Vietnam National University Ho Chi Minh City under grant number D1-2021-33.

References

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of financial economics*, 9(1), 3-18.

Cakici, N. (2015). The Fama-French five-factor model: International evidence. *Fordham University*

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

Dittmar, R. F. (2002). Non-linear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns. *The journal of finance*, 57(1), 369-403.

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

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1), 3-56.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics,* 116(1), 1-22.

Foye, J. (2018). A comprehensive test of the Fama-French five-factor model in emerging markets. *Emerging Markets Review*, 37, 199-222.

Gibbons, M. R., Ross, S. A., & Shanken, J. (1989). A Test of the Efficiency of a Given Portfolio. *Econometrica*, 57(5), 1121-1152.

Gogas, P., Papadimitriou, T., & Karagkiozis, D. (2018). The Fama 3 and Fama 5 factor models under a machine learning framework.

Gruber, M. H. (2017). Improving efficiency by shrinkage: the James-Stein and ridge regression estimators. *CRC Press*, Boca Raton, Florida.

Hilt, D. E., & Seegrist, D. W. (1977). Ridge, a computer program for calculating ridge regression estimates (236). *Dept. of Agriculture, Forest Service, Northeastern Forest Experiment Station, Upper Darby, PA.*

Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1), 69-82.

Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Jolliffe, I. (2011). Principal component analysis. *Springer*, London

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The journal of finance*, Vol. 20, 4, 587-615.

Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019). House price prediction using regression techniques: a comparative study. *Paper presented at the 2019 International Conference on Smart Structures and Systems (ICSSS)*.

Manasa, J., Gupta, R., & Narahari, N. (2020). Machine learning based predicting house prices using regression techniques. *Paper presented at the 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*.

Markowitz, H. (1952). Portfolio Selection. *Journal of Finance*, 7(1), 77–91.

Martinsa, C. C., & Eid Jr, W. (2015). Pricing assets with Fama and French 5–Factor Model: a Brazilian market novelty. *XV Encontro Brasileiro de Finanças*.

Mossin, J. (1966). Equilibrium in a capital asset market, Econometrica. *Journal of the econometric society*, 768-783.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of financial economics*, 108(1), 1-28.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1), 259-268.

Qu, H., & Zhang, Y. (2016). A new kernel of support vector regression for forecasting high-frequency stock returns. *Mathematical Problems in Engineering,* 2016.

Roll, R. (1977). A critique of the asset pricing theory's tests Part I: On past and potential testability of the theory. *Journal of financial economics*, 4(2), 129-176.

Roy, S. S., Mittal, D., Basu, A., & Abraham, A. (2015). Stock market forecasting using LASSO linear regression model. *Paper presented at the Afro-European Conference for Industrial Advancement*.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3), 425-442.

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

Titman, S., Wei, K. J., & Xie, F. (2004). Capital investments and stock returns. *Journal of Financial and Quantitative Analysis*, 39(4), 677-700.

Treynor, J. L. (1961). Market value, time, and risk.

Venkatasubbu, P., & Ganesh, M. (2019). Used Cars Price Prediction using Supervised Learning Techniques. *International Journal of Engineering and Advanced Technology (IJEAT)*, 9(1), S3, 216-223.

Wolfe, P. (1961). A duality theorem for non-linear programming. *Quarterly of applied mathematics*, 19(3), 239-244.

Yu, H., & Wu, J. (2016). Real estate price prediction with regression and classification. *Retrieved from Stanford*.

Zhang, Y., Ma, F., & Wang, Y. (2019). Forecasting crude oil prices with a large set of predictors: Can LASSO select powerful predictors? *Journal of empirical finance,* 54, 97-117.