CAPSTONE PROJECT

Forecasting Stock Returns Using Machine Learning Methods

Heng Li
M13457579

Abstract

Forecasting stock return is an important topic in the finance industry. However, the stock market has high volatility which makes the price movements hard to be predicted. Eugene Fama and Kenneth French introduced the Fama-French three-factor model in their research paper *Common Risk Factors in the Returns on Stocks and Bonds* (1993). The traditional Fama-French three-factor model applied the conventional multiple linear regression model, which is still powerful in evaluating stocks and comparing investment results when stocks are held for different periods. However, in recent years, machine learning methods are taking advantage of calculating speed and forecast accuracy. Therefore, in this project, we will evaluate the model performance for both traditional linear models and machine learning models.

In this project, we applied multiple linear regression, univariate linear regression, random forest, XGBoost, and Artificial Neural Network, models. All models selected Market Excess Return (Mkt.RF) as the most important factor, followed by SMB then HML. However, machine learning methods are not able to outperform linear models in terms of output accuracy. We will in the end, briefly discuss the possible reasons and project limitations.

1. Introduction

Predicting stock returns is a popular yet difficult topic in the industry. Traditionally, Eugene F. Fama and Kenneth R. French had discovered a linear model -- Three-Factor Model in their paper "Common Risk Factors in the Returns on Stocks and Bonds" (1993) which was commonly used in asset pricing and stock valuation. However, with more and more machine learning tools become available in recent years, we can apply these machine learning methods to further analyze the Three-Factor Model.

In this project, I will try to apply Random Forest, Artificial Neural Network (ANN), and XGBoost on the traditional Fama-French three-factor model dataset. Generally, the machine learning method will have higher prediction accuracy. However, it may not be the truth for datasets that contain a limited number of variables, which we will discuss at the end of this article.

2. Data

We used the dataset directly from Professor French's data library. This dataset sperate all Center for Research in Security Prices (CRSP) firms that listed on the NYSE, AMEX, or NASDAQ into six value-weight portfolios formed on size and book-to-market. We selected two portfolios for analysis—small-growth portfolio(contains stocks that in the lower 30^{th} BE/ME percentile, and have market equity lower than the market median), big-value portfolio(contains stocks that in the upper 30^{th} BE/ME percentile, and have market equity higher than the market median). Table 1 shows the summary statistics for these two portfolios.

Table 1

Summary Statistics of Small Market Cap Portfolio and Large Market Cap Portfolio

This table contains the summary statistics of both small-growth portfolio and large-value portfolio from 2001 to 2020.

Summary Statistics					
Small-growth Portfolio					
Factors	Min	Medium	Max	Mean	S.D.
Expected excess return	-0.175	0.069	0.314	0.062	0.124
Market excess return	-0.155	0.041	0.122	0.020	0.074
SMB	-0.062	-0.002	0.072	0.005	0.033
HML	-0.324	-0.006	0.181	-0.003	0.091
Large-value Portfolio					
Factors	Min	Medium	Max	Mean	S.D.
Expected excess return	-0.310	0.048	0.191	0.025	0.115
Market excess return	-0.155	0.041	0.122	0.020	0.074
SMB	-0.062	-0.002	0.072	0.005	0.033
HML	-0.324	0.006	0.181	-0.003	0.091

From the table above, we conclude that the small-growth portfolio has a larger average expected excess return and higher variance between each year. Also, the average expected excess return of the large-value portfolio is very close to the market excess return.

Figure 1 and Figure 2 show the paired correlation matrix plot between each variable. As we can see, the response variable-- expected excess return(value_minus_return) has positive correlations with all 3 factors. Moreover, for the small-growth portfolio, expected excess return has a stronger correlation with SMB than the market excess return. But for large-value portfolio, expected excess return has a stronger correlation with market excess return than SMB. The fact indicates that the large-value portfolio may have a larger impact on the overall stock market.

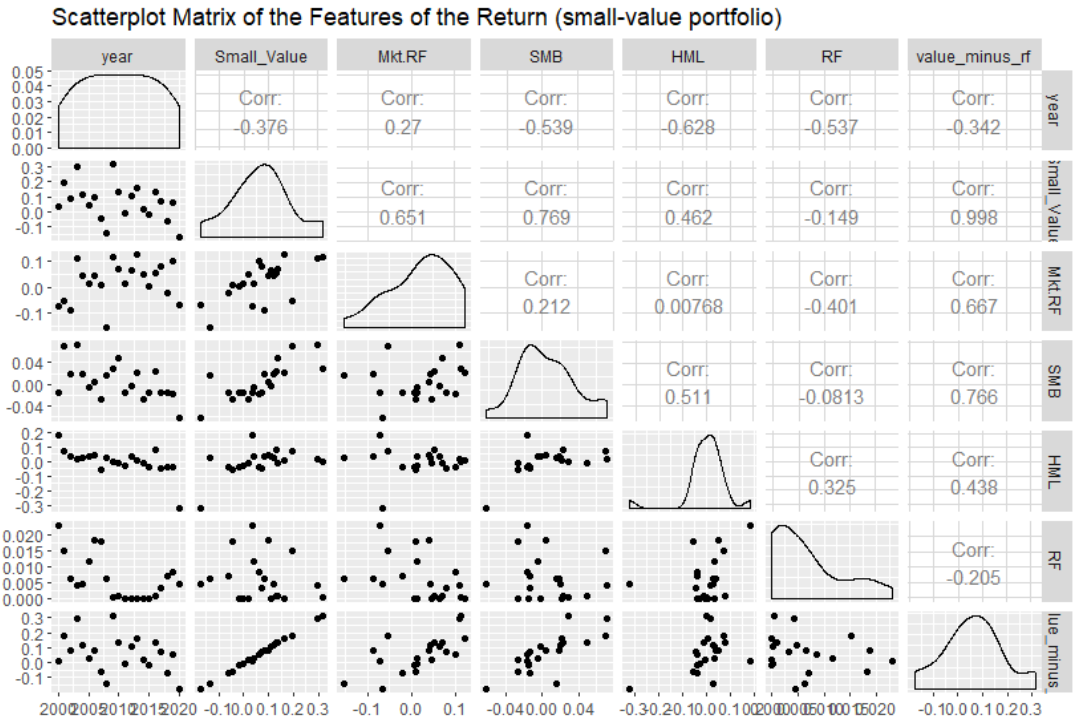

Figure 1 Paired correlation matrix for the small-growth portfolio

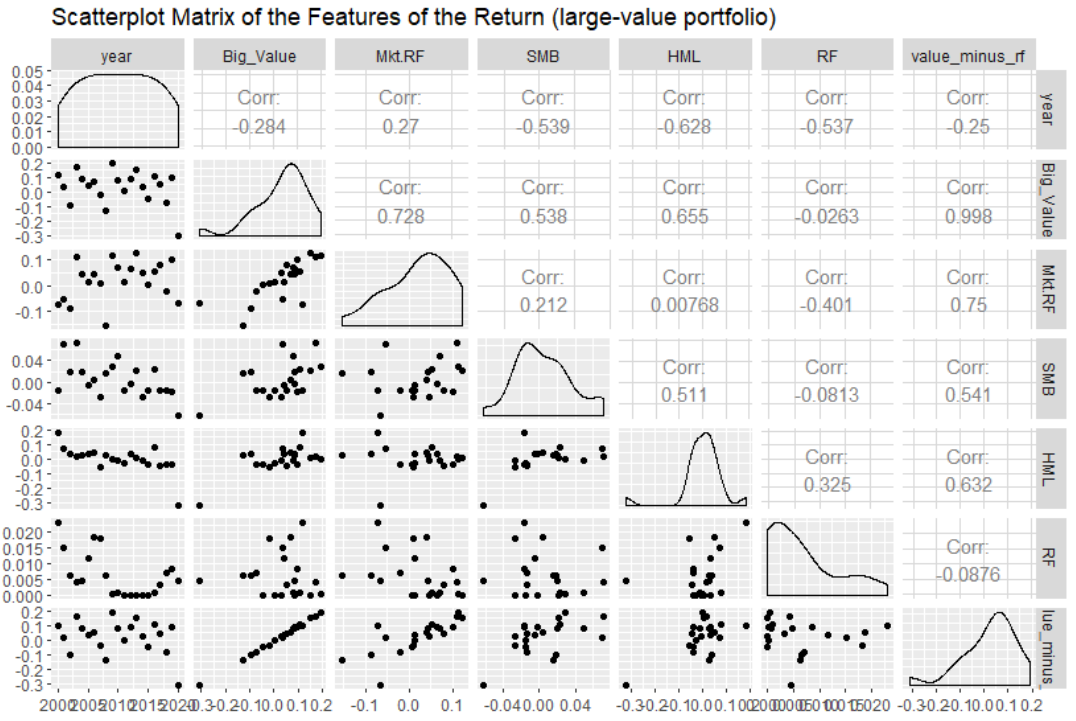

Figure 2 Paired correlation matrix for large-value portfolio

3. Models

We forecast expected excess returns on four models: linear regression, random forest, XGBoost, and artificial neural network. Since the traditional Fama-French model uses linear regression with ordinary least squares (OLS) estimation, we will use it as the benchmark for model comparison.

Estimation Procedure

We applied recursive estimation for these 5 models, which we used 10-year rolling windows. For each rolling window, we split the data into a 10-year training sample and a 1-year testing sample. We build models in the training sample and evaluate their out-of-sample predictions in the testing sample. Then we move forward by one year and repeat the process. The detailed procedure is as follows:

1. Use the first 10 years (2001 – 2010) as the training sample to build models. For the ML models, the sample is further divided into training and validation subsamples for tuning hyperparameters.
2. Evaluate the model's performance in the following year (2011).
3. Moving forward by one year, use the second 10 years (2002 – 2011) to train models and test the model's performance in the following year (2012).
4. Repeat this process until the end of the sample period.

At the end of the procedure, our "combined" testing period covers the 10 years from 2001 to 2020.

Results and Model Comparison

Table 2 summarizes the testing RMSE for each year and each model.

Table 2

Testing RMSE of Each Model for Small-growth Portfolio and Large-value portfolio

Table 2 reports the RMSE of each model for small growth portfolio and large-value portfolio. LM represents the Multiple Linear Model, ULM represents the Univariate Linear Model, RF represents Random Forests, XGB represents XGBoost, ANN represents Artificial Neural Networks.

Testing RMSE of Small-growth Portfolio and Large-value portfolio											
Small-growth Portfolio						Large-value portfolio					
year	**LM**	**ULM**	**RF**	**XGB**	**ANN**	**year**	**LM**	**ULM**	**RF**	**XGB**	**ANN**
2011	0.009	0.071	0.046	0.004	0.289	2011	0.018	0.001	0.010	0.002	0.199
2012	0.011	0.015	0.048	0.106	0.190	2012	0.002	0.039	0.094	0.082	0.322
2013	0.051	0.025	0.018	0.153	0.371	2013	0.004	0.076	0.161	0.139	0.440
2014	0.002	0.028	0.051	0.081	0.360	2014	0.009	0.015	0.037	0.061	0.232
2015	0.008	0.047	0.006	0.098	0.475	2015	0.025	0.039	0.036	0.099	0.170
2016	0.023	0.027	0.027	0.119	0.379	2016	0.007	0.037	0.115	0.091	0.358
2017	0.004	0.016	0.020	0.084	0.307	2017	0.008	0.024	0.056	0.062	0.260
2018	0.007	0.076	0.078	0.143	0.640	2018	0.026	0.067	0.074	0.143	0.181
2019	0.049	0.012	0.012	0.091	0.367	2019	0.011	0.055	0.096	0.076	0.324
2020	0.136	0.026	0.167	0.256	1.020	2020	0.009	0.173	0.302	0.377	0.545
mean	0.030	0.034	0.047	0.113	0.440	**mean**	0.012	0.053	0.098	0.113	0.303
S.D.	0.041	0.023	0.048	0.065	0.236	**S.D.**	0.008	0.048	0.084	0.101	0.121

The results show that the multiple linear regression has the best performance in terms of prediction accuracy in both portfolios. Univariate regression model comes to the close second, both traditional linear methods have better results than machine learning methods. On the machine learning side, random forests and XGBoost return similar results in terms of prediction accuracy and output variance. However, the ANN model does not deliver satisfying results. One possible reason is that we did not add RNN or LSTM layer to the regular neural networks. That may leave our regular model lack of adaptability to the time-sensitive data.

3.1 Linear Regression

We performed linear regression models on two platforms. First, we performed a multiple linear regression model with all three factors included in the model. Then, we also performed a forecast combination based on the univariate regressions. The univariate regressions take one predictor at a time and then we calculate the average of the forecasts from all the univariate regressions as the final forecast.

Table 2 illustrates the multiple regression coefficient estimates of two portfolios. We can observe that market excess return (Mkt.RF) is the most statistically significant and practically impactful factor in the multiple linear regression model. SMB factor is not statistically significant from 2011 to 2020. The same verdicts can be applied to both portfolios.

Table 2

Coefficient Estimates of Multiple Linear Regression Model

This table contains all the coefficient estimates of the multiple linear regression models we have made. 2011-2020 are the years that we made predictions.

Multiple Linear Regression Coefficient Estimates									
Small-growth Portfolio					Large-value Portfolio				
	α	Mkt.RF	SMB	HML		α	Mkt.RF	SMB	HML
2011	0.025	1.126***	2.203	0.611**	2011	0.001	1.222***	0.262	0.931**
2012	0.027	1.147***	2.065	0.747	2012	0.001	1.155***	0.572	0.435
2013	0.026	1.166***	1.968	0.647	2013	0.001	1.172***	0.498	0.394
2014	0.010	1.333***	1.755	0.455	2014	0.004	1.106***	0.564	0.436
2015	0.010	1.330***	1.734	0.459	2015	0.004	1.104***	0.699	0.421
2016	0.011	1.326***	1.719	0.429	2016	0.001	1.115***	0.723	0.493
2017	0.010	1.318***	1.725	0.326	2017	0.000	1.113***	0.744	0.448
2018	0.010	1.318***	1.701	0.371*	2018	0.001	1.114***	0.681	0.555*
2019	0.011	1.320***	1.653	0.339*	2019	-0.002	1.139***	0.705	0.621*
2020	0.005	1.339***	1.728	0.373*	2020	-0.018	1.399***	0.451	0.563*
mean	0.015	1.272	1.924	0.577	mean	-0.001	1.638	0.590	0.530
S.D.	0.008	0.088	0.186	0.144	S.D.	0.006	0.091	0.153	0.159

Table 3 shows the univariate regression coefficient estimates of small-growth portfolio and large value portfolio. Like the multiple linear regression model, the Mkt.RF is still the most important

factor in the univariate regression model for both portfolios. Meanwhile, we found SMB replaces HML becomes the second significant factor in the univariate linear regression model.

Table 3

Coefficient Estimates of Univariate Linear Regression Model

This table contains all the coefficient estimates of the multiple linear regression models we have made. 2011-2020 are the years that we made predictions. Both α and coefficients of 3 factors are presented separately.

Univariate Linear Regression Model Coefficient Estimates													
Small-growth Portfolio							Large-value Portfolio						
	α	Mkt.RF	α	SMB	α	HML		α	Mkt.RF	α	SMB	α	HML
2011	0.09	1.189**	0.032	2.948*	0.1	-0.214	2011	0.037	0.975**	0.019	1.017	0.035	0.162
2012	0.079	1.193*	0.029	3.008*	0.082	0.811	2012	0.019	1.136***	0.003	1.35	0.033	-0.15
2013	0.057	1.353**	0.039	3.238*	0.08	0.571	2013	0.012	1.203***	0.013	1.704	0.037	0.001
2014	0.03	1.581***	0.045	3.292*	0.089	0.619	2014	0.013	1.178***	0.033	1.855	0.058	0.377
2015	0.019	1.435***	0.052	2.798	0.066	0.55	2015	0.009	1.136***	0.039	1.597	0.046	0.331
2016	0.011	1.450***	0.049	2.863	0.057	0.689	2016	0.001	1.159***	0.032	1.745	0.037	0.537
2017	0.014	1.465***	0.051	2.901	0.063	0.869	2017	0.003	1.181***	0.034	1.86	0.041	0.708
2018	0.01	1.421***	0.055	2.738	0.069	0.838	2018	-0.001	1.146***	0.037	1.751	0.047	0.693
2019	0.014	1.427***	0.052	2.768	0.067	0.785	2019	-0.002	1.179***	0.031	1.904	0.042	0.786
2020	-0.027	1.861**	0.077	3.166**	0.095	1.328	2020	-0.036	1.629***	0.056	2.078*	0.071	1.125
mean	0.03	1.438	0.048	2.972	0.077	0.685	mean	0.005	1.192	0.029	1.686	0.045	0.458
S.d	0.036	0.191	0.014	0.199	0.015	0.386	S.D	0.019	0.166	0.015	0.305	0.012	0.389

3.2 Random Forest

Random Forest is a popular modification of the "bagging" procedure. It consists of a large number of individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction and the class with the most votes becomes the model's prediction.

Tuning

In the perimeters tuning phrase, we used a grid search method to find out the best model which returns the lowest training RMSE. Both small-growth portfolio and large-value portfolio select 300 trees. The random forest model for small-growth portfolio select the number of variables for

splitting at each node equals to 1, and the model for large-value portfolio selects the number of variables for splitting at each node equals to 3.

Predictor Importance

For random forests, predictor importance is determined by "Increase in Node Purity". The output value represents the sum of reduced RSS over all the splits for that variable, averaged over all trees. The results shown in Figure 2 indicate that all three factors have similar importance, Mkt.RF is slightly more important than the other two for the small-growth portfolio. However, for large-value portfolio (second graph in Figure 2), the Mkt.RF is significantly more important than SMB and HML.

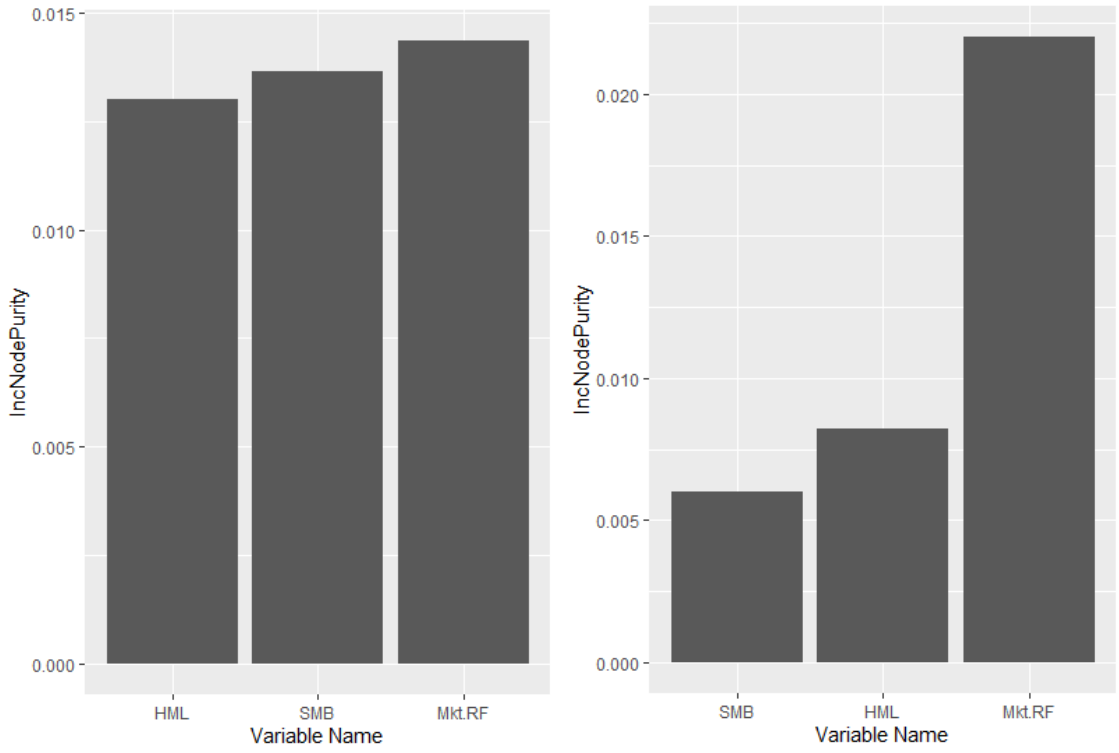

Figure 2 Predictor importance plot for small-growth portfolio and large-value portfolio

3.3 XGboost

XGBoost stands for Extreme Gradient Boosting. It is also an ensemble method based on the Gradient Boosting framework. XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements. Thus, XGBoost usually yields superior results, using fewer computing resources, in shorter running time.

Tuning

We applied a customized grid search method for hyperparameters tuning. The setup ideology is similar to what we did for random forests. Both small-growth portfolio and large-value portfolio select gamma equal to 0. The learning rate for the small-growth portfolio is set to 0.1, the learning rate for the large-value portfolio is set to 0.3. The model for the small-growth portfolio sets the maximum depth to 4, the model for large-value portfolio sets the maximum depth to 5.

Predictor Importance

We used the gain scores as the measurement of the predictor importance. The features are arranged according to the descending value of gain score resulting in the most important feature to be displayed at the top. Features which has similar gain scores will be grouped into one cluster. In Figure 4 and Figure 5 show the variable importance graph of two portfolios.

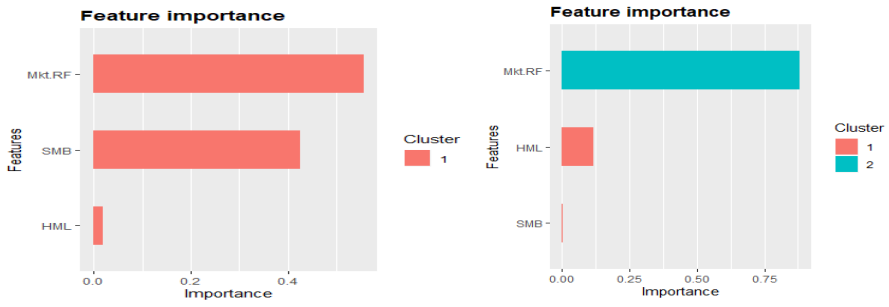

Figure 4 Predictor importance plot for small-growth portfolio and large-value portfolio

Mkt.RF is the most important factor for both models. However, SMB has a similar importance to the Mkt.RF for small-growth portfolio model. In the model for large-value portfolio, Mkt.RF is significantly more important than the other two factors. Therefore, Mkt.RF is classified into a separate cluster.

3.4 Artificial Neural Network

Finally, we performed the Artificial Neural Network model. Artificial Neural Networks (ANN) are multi-layer fully connected neural nets. They consist of an input layer, multiple hidden layers, and an output layer. Every node in one layer is connected to every other node in the next layer. ANN is commonly used in model complex patters and prediction problems.

Tuning

We manually tuned the hyperparameters for the ANN model, by setting the number of hidden layers, the number of neurons per layer, and the activation function. Thus, we constructed 4 combinations for each portfolio. Combination 1 and combination 2 use logistic activation function, combination 3 and 4 use the Tahn activation function. Combination 1 and 3 use one hidden layer, combination 2 and 4 use 2 hidden layers; the first hidden layer has 4 neurons, the second hidden layer has one neuron. Figure 5 shows the model comparison on testing RMSE between different tunings.

The results show that the small-growth portfolio prefers the Tahn function with 1 hidden layer (combination 3), the large-value portfolio prefers logistic function with 1 hidden layer (combination 1).

Discussion

As the results indicated, linear models have significantly better performance than machine learning methods. Linear models have lower average RMSE and lower variance across the testing period. However, we cannot perorate that linear models are a promising tool for asset pricing or stock return forecasting. There are still some limitations in our model.

First, the dataset we used only contains a limited number of factors. We are doing research based on the traditional Fama-French three-factor data, which as the name indicates, only contains three variables(factors). That ultimately limited the performance of machine learning methods which are more appropriate for complex datasets. In the future study, we may apply this same testing structure on the five-factor model or ten-factor model, we may come up with a more complete conclusion then.

Second, the ANN model in our project is not perfectly constructed. The original ANN model may not be the ideal method for time-sensitive data. We tried to add RNN or LSTM layer to the ANN model to help our neural networks adapt to the time changes. However, we failed to cohere these three factors with the model, which converted the original model to a time-series model. This is not our goal. In the future, we will try to add RNN or LSTM layer to the current neural networks with the original three (or five, ten) factors still playing a crucial role.

Reference

1. Fama and French, 1993, Common Risk Factors in the Returns on Stocks and Bonds, *Journal of Financial Economics*
2. Han, Yufeng, Ai He, David Rapach, and Guofu Zhou, 2018, Firm Characteristics and Expected Stock Returns, working paper.