SDSC6002 Research Project: Predicting Stock Market Return

Lee Yat Shun, Pang Man Ki, Shek Wang Yuk

School of Data Science, City University of Hong Kong

1. Introduction

The expected market return has been an important research topic in asset pricing of Financial Economics. Risky assets should carry a positive expected return. Investors demand compensation when holding the risky assets as this cause uncertainty in their future wealth. Investors give up short-term liquidity to hold the risky assets as they purchase the asset with more liquid cash equivalents. In the equity market, the excess return to compensate the risk of holding a stock is called equity risk premium.

The Adaptive Market Hypothesis (AMH) from Lo, A. W [2004] suggests that equity risk premium varies. Hull and Qiao's research [2017] further show that the varying board equity market return is predictable. Upon the recent advancement in Data Science and computational power, we are interested to research whether data science techniques can contribute to predicting the future equity return. Our primary objective is to predict future stock market return with data science, then we will evaluate our methods in testing datasets and investigate whether our research can yield profiting predictions in out-of-sample settings.

1.1 Background & Objective

Being a popular investment choice, the stock market has higher potential growth and risk. Therefore, stock return prediction becomes one of the most concerning issues for investors. Many analysts and researchers have developed various methods to forecast stock market's return and seek to maximize profit in proper decision-making.

The well-known Efficient-market hypothesis (EMH) from Fama, E. F. [1970] suggests that asset prices reflect all public information, thus it is impossible to consistently earn extra return on a risk-adjusted basis. However, empirical evidence shows that some market players can generate alpha from the market through superior techniques such as advantages in accurate research. The later Adaptive Market Hypothesis (AMH) from Lo, A. W [2004] proposes that, on top of the EMH, other forms of behavioral abnormalities claimed by the EMH can co-exist in the asset market. According to the AMH, the relationship

between Risk and Reward in assets are unstable and time-varying, according to changes in exogenous conditions such as the Economy. Arbitrage and alpha can exist in the market, but the opportunities will be absorbed and exploited by investors shortly due to innovation and market adaptation. The AMH provides an important theoretical support to this project that, predicting the market equity premium and attempting to "beat the market" by generating alpha is possible.

Since 1970, with the rapid development of computer technology and data science, researchers have begun to forecast stock return by various machine learning algorithms. The general idea is to forecast the return by different predictors through machine learning model. Our research project aims to forecast an S&P 500 Index return from both technical and fundamental data. The forecast used different machine learning techniques, from feature engineering to predictive modelling. Also, we aim to design a corresponding trading strategy based on our forecasts and it can outperform the market which results in a higher Sharpe ratio compared with “buy and hold.”

Prado [2018] proposed that the advantage of Machine Learning algorithm on investing is that it can spot high dimensional patterns as easily as humans who are familiar with the 3-dimensional space. Also, Humans are slow learners which is a disadvantage in a fast-varying finance industry. Therefore, it is suitable to use machine learning tools in finance. However, he emphasizes the danger of applying Machine Learning to forecasting stock return like overfitting. Also, he characterizes some general mistakes, such as data leakage, which would make standard machine learning tools fail. In our project, we do not only aim to outperform the market, but also prefer reasonable and standard ML algorithms.

2.Data

2.1 Features and Forecasts

Neetly and Rapach [2014] utilize technical indicators to predict the excess return and compare their performance with some famous macroeconomic variables. Their results reveal that technical indicators are statistically and economically significant in-sample and out-of-sample in forecasting the monthly excess return. More importantly, they find out that combining both technical and macro factors can capture several types of information. Technical indicators perform well in detecting the typical decline in equity near peaks while macroeconomic variables perform well in detecting the rise near troughs. Therefore, we include some well-known technical and macroeconomic variables as predictors to forecast the market return.

According to Xu, Li, and Singh [2022], they find out that using the first difference of Gold to platinum ratio(ΔGP) can help to forecast market return. ΔGP is significantly having correlation with returns in many stock markets. Even after controlling the effect of some macroeconomic variables, their relationship is still robust. ΔGP has robust predictive power during different business cycle and in both short and long horizon. In our project, we add the ratio of gold to other commodities as predictor, including gold to silver and gold to copper.

Hull and Qiao [2017] has proposed a set of predictors to forecast the excess market return in their research. They believe that combining predictors with diverse characteristics could give a superior prediction. We have referenced their choice of features. Our features are diversified including valuation related, fixed income related, commodity related, Economy related, volatility related and market timing related.

We would like to describe our predictors and forecasting variables and their data source. Moreover, we have introduced new predictors in our research project. We include a wide range of features in this project, to ensure the future equity return can be predicted by different aspects of view. The features are chosen either by the support of previous research, or chosen because of its potential predictive power in future return.

Most of the data can be constructed or extracted from Bloomberg, database form Federal Reserve Economic Data, WRDS and other open data library. The raw data for features and forecasts ranged from 1989/1/3 to 2021/12/31. The raw data we used is as follows:

Features

Valuation related:

Symbol	**Description**
DP	Sum of gross dividend per share in S&P500 divided by share price.
PE	The ratio of share price to earnings per share.
BM	The ratio of share price to book value of company.
CAPE	The ratio of share price to the inflation adjustment earnings.

Yield, money market and fixed income related:

Symbol	Description
BY	the 10-year Treasury bond yield divided by the bond yield EMA
DEF	Difference between Baa yield (MOODCBAA Index) and Aaa yield (MOODCAAA Index)
TERM	Yield difference between the 10-year Treasury note and the three-month Treasury bills.
LOAN	The Net Percentage of Domestic Banks Tightening Standards (Quarterly)

Commodity related:

Symbol	Description
PCR	log of the ratio between SPY and a board commodity index (SPGSCI Index)
OIL	The difference between the log of current front oil future price and the log of the fourth future price with three-month lag.
G/P	The residual of platinum price regressed on gold price
G/S	The residual of silver price regressed on gold price
G/C	The residual of copper price regressed on gold price

Economy related:

Symbol	Description
CAY	the co-integrating residual of log consumption, assets, and wealth. Updated with the approach developed by Ren and Xie [2018]
BDI	The three-month change of BDIY Index
NOS	The log of the ratio of new orders and shipments of durable
CPI	the change in CPI over the last 12 months
M3	United States money supply including M2 plus large time deposits in banks.

Volatility and correlation related:

Symbol	Description
VRP	The difference between implied and realized volatility, estimated by the difference of CBOE volatility index (VIX) and a GARCH estimate from Yang and Zhang [2000] drift-independent volatility estimator
IC	The CBOE S&P 500 Implied Correlation Index

Market-timing related:

Symbol	Description
SIM	The proportion of the next 130 business days is between the second business day in May and the 15th business day in October
Technical Indicators	Moving average (MA), On Balance Volume (OBV) and Momentum indicator (MoM) under typical settings

Forecasts:

Symbol	Description
FutureRet (1M/3M/6M/1Y)	S&P 500 Composite Index future return in different period.

2.2 Feature Engineering

The valuation related features, technical indicators, and the commodities price ratio are transformed into PCA Price, PCA Tech and PCA Commod by principal component analysis (PCA) because of the multicollinearity issue. Hull and Qiao [2017] suggested replacing the features with the first principal component of the features to preserve most of the information while addressing the issue. See Section 3.2.1 for more details of PCA. From the correlation matrix (Figure 1), we can observe distinct groups of features with high correlation. The correlation matrix among all the features and the correlation between features and forecasts are shown in the appendix.

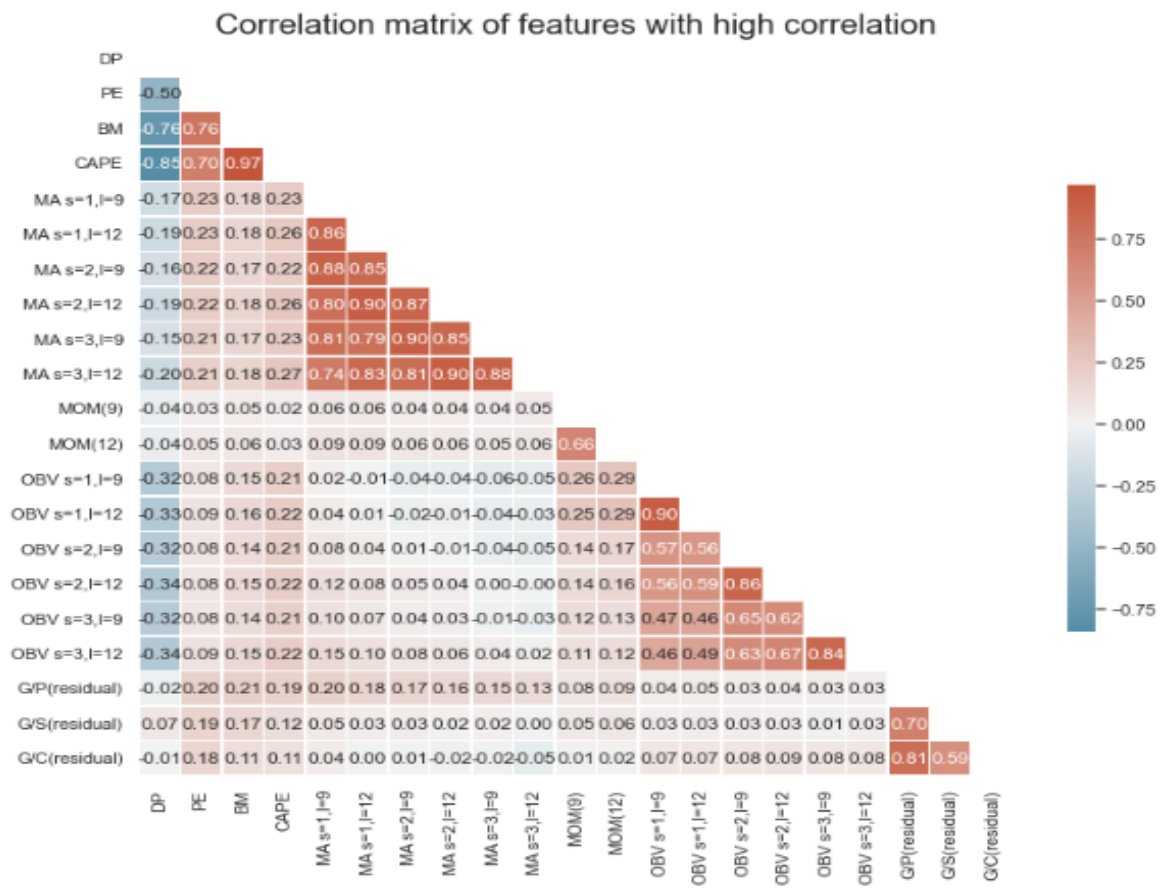


Figure 1. The Correlation Matrix of the Components of PCA Price, PCA Tech, and PCA Commod.

3 Methodology

3.1 Framework

Our research framework has referenced Hull and Qiao's research [2017]. After chosen features, they would conduct feature selection before fitting the model. Hull and Qiao perform correlation screening to select features having higher correlation with forecast. In our research, we also perform feature selection but in another way. After simple feature engineering, we first conduct principal component analysis (PCA) on features. Then we treat different principal components as features and perform feature selection by different methods such as tree-based method and lasso. Lastly, we use the principal component selected to forecast excess return.

We also referenced their design of trading strategy. They chose 130-day market return as their forecast target. For every 20 days (about 3 weeks), they would refit model by previous 10 years data to forecast

the 130-day market return, and then take the position on SPY by eight times the forecast made. The position is restricted from 1.5 to –0.5. In our project setting, for every 23 trading days, we refit the model by previous 10 years data to predict the one-month future return. We take the position on SPY by 256 times the forecast made. It ranges from 1 to –0.5. We do not change the position of the following day if the changes are not greater than 0.1 to reduce the transaction cost.

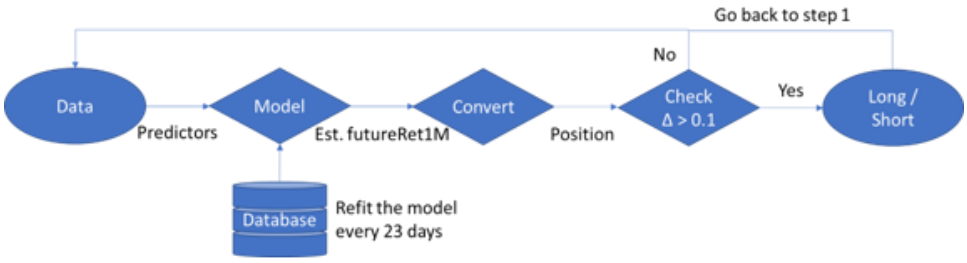


Figure 2. Flowchart of Trading Framework.

3.2 Dimensionality Reduction and Feature Selection

To simplify the model and address the overfitting issue, feature selection and dimensionality reduction on the sample set can improve the predictive ability of the unseen data by reducing the variance and help to interpret the model much easier.

3.2.1 Principal Components Analysis (PCA)

The modern idea of PCA from the book authored by Jolliffe [2002] is to reduce dimensionality while preserving essential information. The data is transformed into a new coordinate system such that the new set of variables is uncorrelated. The data must be standardized before the transformation since the result of PCA depends on the scaling of data. The first orthogonal feature is obtained by the first eigenvector v_1 of $X^T X$ which has the greatest variance, and the second orthogonal feature is obtained by the second eigenvector v_2 with the second greatest variance and is subject to the $v_2^T v_1 = 0$, and so on. Nevertheless, Principal Components (PCs) with lower variance might still have higher predictive power towards the future return than PCs with larger variance.

3.2.2 Model Based Feature Selection

To reduce the model complexity, apart from naively choosing the number of components with a cut-off variance threshold, we can measure the PCA feature importance by any estimator that generates

coefficients or a specific measure for each feature. We also adopted the following methods in our experiment to retain the PCs with higher explanatory power towards future return.

(i) LASSO Feature Selection

Reviewing the book by James et al. [2013] using the least squared method and a method to constrain the coefficient estimates and shrink the estimates towards zero. LASSO minimizes the L1 penalized RSS:

$$\min_{\beta}(y - X\beta)^2 + \lambda \sum_{i=i}^{P} |\beta_i|$$

The LASSO penalty yields sparse models and performs variable selection when λ is sufficiently large. We can drop the redundant features and keep the features that are strongly correlated with the output.

(ii) Tree-based Feature Selection

Growing a tree model splits the leaf node if the split reduces the impurity of the node. To measure the node impurity, we can measure by the impurity function [Breiman et al., 2017]:

$$i(t) = \sum_{i \in t}\big(y_i - avg(y|t)\big)^2$$

Prado [2018] proposed using tree-based estimators to measure feature importance because we can average the impurity scores across all the estimators and then pick the key features based on the Gini importance. Applying ensemble methods can reduce bias and/or variance whilst bootstrap aggregation (bagging) is a robust method to prevent overfitting because averaging reduces the variance. Therefore, we deployed the random forest to choose the prominent features before fitting them into the regression model. We will elaborate more about the theory of random forest in Section 3.3.2.

3.3 Regression Model

We selected a simpler model and a complex model for predicting the response variable in our experiment to show how the model flexibility affects the prediction result.

3.3.1 Ordinary Least Squares (OLS)

The linear model determines the best fit line for the data points and predicts the behavior of the endogenous variable. It is considered inflexible since it builds the potential linear relationship, and the

noise does not heavily influence the model. By minimizing the mean squared error, the coefficient estimates are obtained [James et al., 2013]:

$$\underset{\beta}{argmin}(y - X\beta)^2$$

3.3.2 Random Forest (RF)

The forest of trees model collects the prediction results from N decorrelated decision trees and outputs the averaged prediction [Breiman, 2001]. It utilizes bagging to reduce the variance while keeping the bias low. It draws N bootstrap samples and grows a tree for each sample. To grow a tree, it randomly selects a subset of features which reduces the correlation and splits the node by picking the best split among the variables based on the goodness of fit.

$$\hat{y} = \hat{f}_{RF}(X) = \frac{1}{N}\sum \hat{f}_{Tree}(X),\ where\ \hat{f}_{Tree}(X) = avg(y|t)\ and\ t = node\ t$$

3.4 Market Return Prediction Models

We employed 6 combinations of regression models and feature selection methods to predict the market return. We used Python and Scikit-Learn to build the models.

3.4.1 Kitchen Sink Model

It takes all the independent variables, except the components of PCA PRICE, PCA TECH, and PCA COMMOD which the components are highly correlated, and the multicollinearity may lead to an overfitting issue, to attempt to explain the variance in the dependent variable. The regression model is the OLS which is simple and easy to interpret, and it can act as a reference.

3.4.2 OLS with Tree-based Feature Selection

It combines linear regression and tree-based selection to predict the future monthly market return. The threshold to retain the feature is the average of the feature importance generated by the Random Forest. We discarded the features whose feature importance are less than the threshold.

3.4.3 OLS with PCA Features

We performed dimensionality reduction with a threshold equal to 80% explained variance ratio. We retained the features that explained most of the variance and were de-correlated. To avoid data leakage

and look-ahead bias, we transformed the training and testing data by the PCA which was only fitted by the training data and fitted into the OLS.

3.4.4 OLS with Tree-based PCA Feature Selection

Selected by the tree-based estimator, the retained PCA features had higher importance weights and were the predictors of OLS to forecast the market return. The threshold of keeping the prominent features is the mean of the impurity score from the forest.

3.4.5 OLS with LASSO PCA Feature Selection

The shrinkage method penalizes the model complexity by tuning the weight of L1 term and shrinks some of the coefficient estimates of PCA features toward zero.

3.4.6 Random Forest with PCA Features

Instead of using multivariate linear regression, random forest that combines a set of learners provides a prediction by majority voting to increase the flexibility, given the orthogonal PCA features that address the multicollinearity issue.

4. Forecasting Results

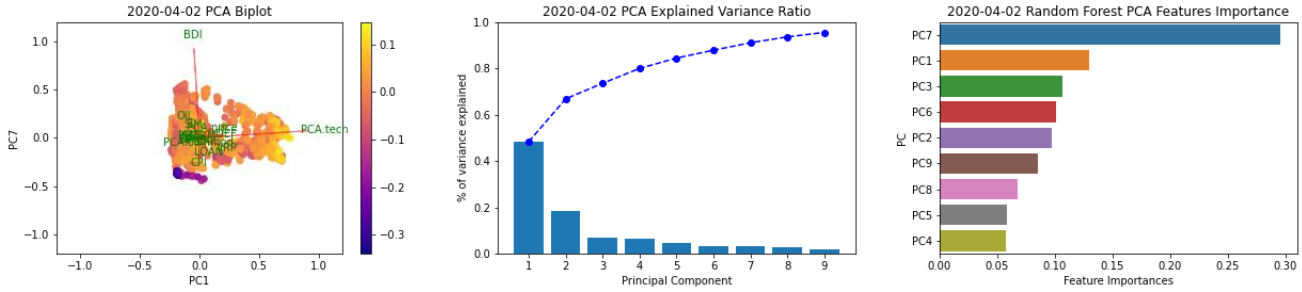
4.1 Statistical Measure

Models are trained with 10 years of data and the overlapping parts are removed to prevent data leakage, since the closest monthly data contained future information in the testing set [Prado, 2018], the model made predictions on monthly future returns for the next 23 days. To measure the performances of the forecasting models statistically, we used mean squared error (MSE) and R-squared.

Prediction Model	MSE	R-squared
Kitchen Sink Model	0.0033	-0.3655
OLS with Tree-based Feature Selection	0.0029	-0.2304
OLS with PCA Features	0.0024	-0.0062
OLS with Tree-based PCA Feature Selection	0.0023	0.0302
OLS with LASSO PCA Feature Selection	0.0024	-0.0079
Random Forest with PCA Features	0.0044	-0.8223

Table 1. Statistical Performances of Each Prediction Model.

Comparing the statistical performances of forecasting models based on the model simplicity, random forest with PCA features performed the worst among all the models, although the majority voting can reduce the variance. It showed that simpler models are preferred in this research. Overfitting issues from a complex model are likely to occur because of the low signal-to-noise ratio [Prado, 2018].



(a) (b) (c)

Figure 3. The PCA biplot (a) visualizes how the features contribute to the top 2 important PC components on 2/4/2020. On the same day, the plots (b) and (c) show the explained variance ratio of the first 9 PCs and their Gini importance respectively.

A simple model with feature selection is preferable to make predictions, as observed from our results. Using feature selection methods improves compared to predicting natively by the kitchen sink model in terms of MSE and R-squared. Besides, transforming into PCA features can address multicollinearity and overfitting issues. Orthogonalization can hugely increase performance. Nevertheless, observed in Figure 3, the first component explains almost half of the variance but the Gini importance is the second highest among first 9 PCs. PC 7 explains roughly 5% of the total variance but is ranked first among the 9 PCs. Using only PCA features that explain most of the variance might not be enough to predict the future market return effectively.

Table 1 shows that OLS with tree-based PCA feature selection has the highest predictive power for the testing set since the R-squared is the highest and MSE is the lowest. Saeys et. al [2008] believed that combining multiple weak learners can provide a more stable and better performing result. Using the ensemble feature selection method combined with PCA features in our experiment yields a more robust result in the prediction problem.

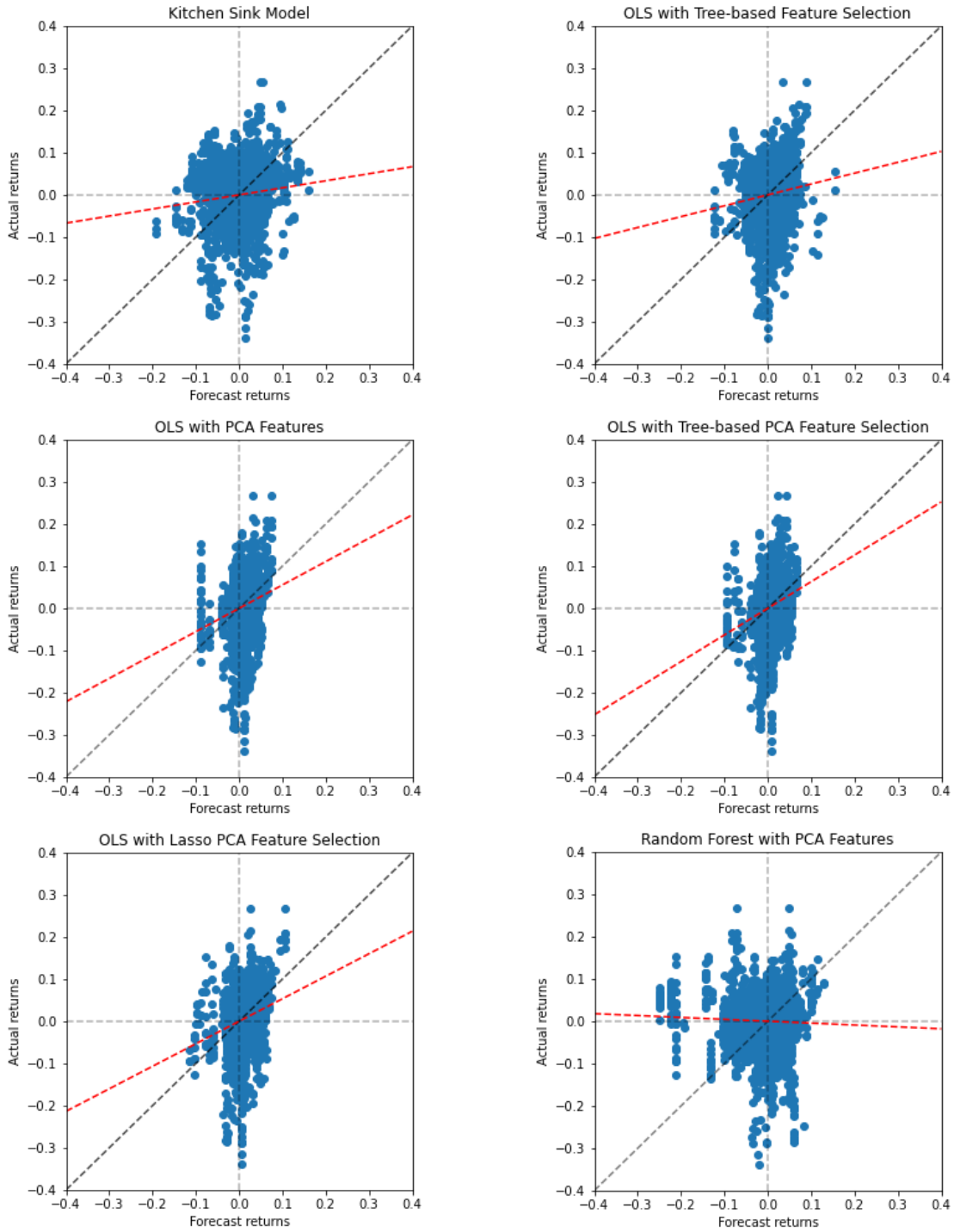


Figure 4. Actual Return v.s. Forecast Return. Note that the positive-positive region and the negative-negative region are separated by the grey dashed horizontal and vertical lines, and the dots inside the regions are predictions that models make correctly in terms of the sign. The dots would be lying on the 45-degree black-dashed line if the predictions and the actual returns are the same. Red dashed lines are the best-fit lines.

4.2 Performances of Investment Strategies

Campbell and Thompson [2008] proposed that models with R-squared statistics extremely low but positive can be economically meaningful and be converted into profitable trading strategies. OLS with tree-based PCA features selection had a R-square 3.02% which is the best performing model. We then converted them into investment strategies based on the framework we mentioned in Section 3.1. We took the T-bill 3 months interest rate to compute the annualized Sharpe ratio.

Trading Strategy	**Annualized Sharpe Ratio**	**Max. Drawdown**	**Position Changed**
Buy-and-Hold	0.35	-55.19%	1
Kitchen Sink Model	0.01	-40.86%	3304
OLS with Tree-based Feature Selection	0.28	-45.76%	3104
OLS with PCA Features	0.39	-33.72%	2305
OLS with Tree-based PCA Feature Selection	0.66	-33.72%	2294
OLS with LASSO PCA Feature Selection	0.41	-33.72%	2699
Random Forest with PCA Features	0.36	-33.20%	3104

Table 2. Performances of Investment Strategies, 4th June 2001 – 29th November 2021

The performance of all trading strategies is summarized in table 2. The OLS with Tree-based PCA feature selection yields the highest annualized Sharpe ratio of 0.66, compared with 0.35 of holding SPY. The second-best performing strategy is the OLS with LASSO PCA feature selection with a Sharpe of 0.41. Although the performance is much lower than the best strategy, it still outperforms buy-and-hold slightly. The max drawdown of these two strategies is –33.72% which is smaller than that of SPY.

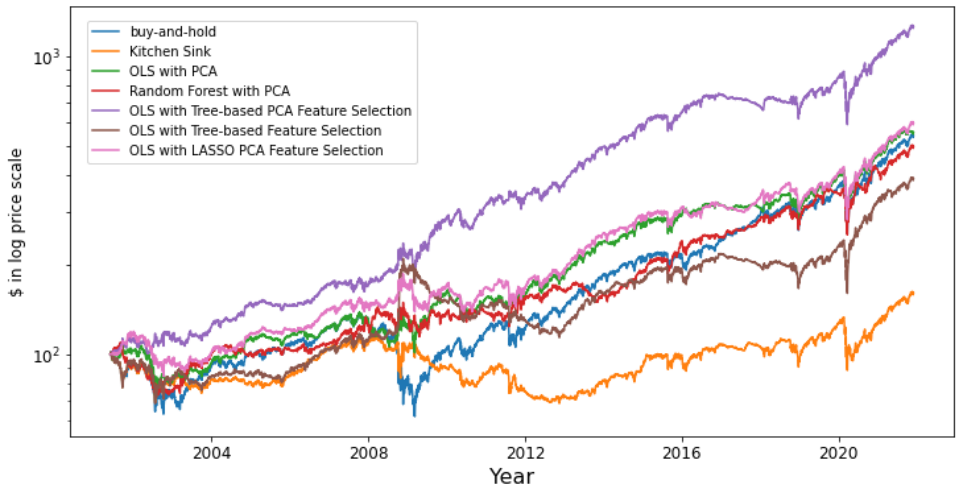


Figure 5. Wealth Accumulation of Investment Strategies, 1st June 2001 – 29th November 2021

We compare all strategies' results using the buy-and-hold strategy as the benchmark.

Firstly, it is observed that all PCA related strategies achieve a better Sharpe ratio. Except Random Forest with PCA, all of them achieve a higher return than buy- and-hold. It suggests that applying PCA for data processing and feature engineering benefits a variety of models when applying in financial return forecasting. Applying PCA may also provide a more robust prediction in testing data.

Secondly, the strategy OLS with Tree-based PCA Feature selection significantly outperforms all other strategies. Its Sharpe ratio is nearly doubled, and the back-test return is doubled compared with Buy-and-hold. It may be explained by the higher predictability of the model, which often makes more accurate predictions compared with others. We believe it is the synergy effect of PCA with tree-based feature selection since PCA with lasso feature selection does not have such a significant improvement.

Thirdly, we notice that all models with feature selection and the kitchen sink model can respond to the 2008 financial crisis and avoid a crash. The wealth accumulation of "buy-and-hold" strategy decreased dramatically in the financial crisis, while strategies with feature selection not only avoid losses and can even generate profit by short selling the market. However, all models cannot predict the market crash from COVID pandemic in 2020. Even our best strategy suffers from around 33% drawdown. Our strategies failed to predict the sudden market reaction due to the global public health issue.

5.Discussion

To test the robustness by purging the training data

To test the robustness of our strategies, we used the training data from 10 years ago to the previous 2 months ago to stretch a longer overlapping part to prevent leakage. Our previous setting used training data till 1 month ago. The result is still robust and does not change significantly. The OLS with tree-based PCA feature selection model, the benchmark model, still outperforms the SPY notably.

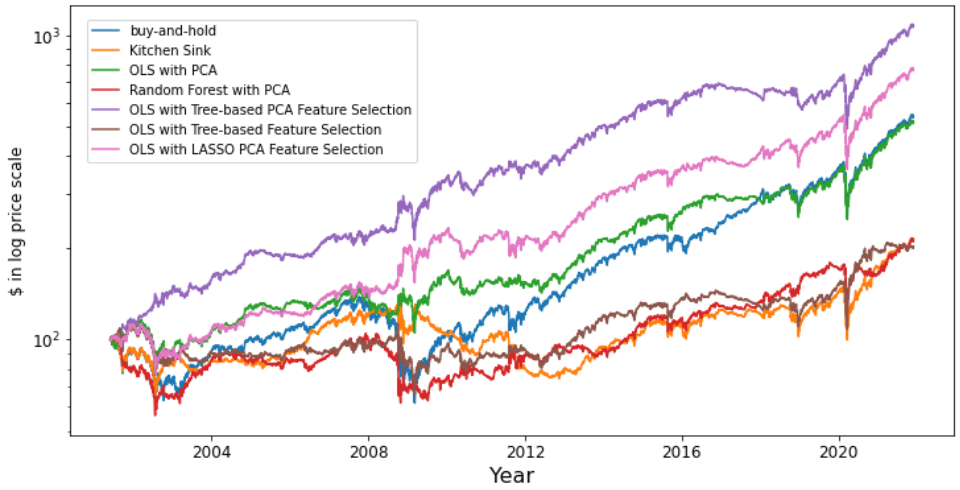


Figure 6. Wealth Accumulation of Investment Strategies, 1st June 2001 – 29th November 2021, with 2 Months Delayed

One of the potential reasons of the result of OLS with tree-based PCA feature selection changing slightly is due to the randomness induced by the random forest. Bootstrapping samples create randomness when building a tree. Our suggestion is to select a sufficiently large number to build numbers of weak learners for the majority voting. This stabilizes the result and will not introduce variance to the prediction model.

Possible reasons for outperforming the market

Here are some possible rationales why our application of PCA and Random Forest feature selection can out-perform the market:

PCA is a dimension reduction technique that can be used to reduce noise. At the same time, financial time series are full of noises and distortions. PCA seems to be a suitable candidate to denoise this dataset, proven by applying PCA alone to regression can improve testing data prediction accuracy and robustness.

One typical method of applying PCA is to retain the PCs with the largest explained variance in descending order and to keep total explained variance above a threshold. We improve the above typical setting as we observe that PCs with high variance does not imply that they also have high explanatory towards future return. Our study finds empirical evidence that PCs with smaller explained variances may have the highest feature importance within some training data. We further show that conducting feature selection in PCs improves results.

Feature selection using random forest outperforms Lasso in this study. Random forest has the advantage of using majority voting by many decision trees, while Lasso uses 1 estimator only in feature selection and the lasso coefficients introduce bias in the process. Random forest reduces overfitting by creating numbers of trees on the subset of training data. The feature selection process is more robust.

Potential issue in backtesting

There may be some potential bias even though our backtesting result seems flawless. According to the "Seven Sins of Quantitative Investing" by Luo [2014], it lists some potential issues in model building and backtesting. Below are some of them. The first is short selling. Taking a short position requires a lender and the cost of the short selling is unknown. It is hard to include in backtest. The second is look-ahead bias which the data may not be available at the time of simulation. More importantly, backtesting may be just storytelling which some random pattern is justified by making up a story ex-post. Therefore, good backtesting performance would not guarantee a high Sharpe ratio in the future. Our research project tried to avoid look-ahead bias which the data of features can be collected at simulation time. Also, our model is interpretable, and the model is not as complicated as a black box, so it is less likely to be just story telling. However, we acknowledge that the transaction cost and the issue of shorting may influence our results.

Possible further work

In our setting, we consider equity as the only asset class. In modern portfolio allocation, investors are interested to consider the correlations of equity with other asset classes, such as fixed income and commodities, and allocate in different classes to achieve diversification, a popular example is the risk parity strategy in mutual funds. As a future work, we may make use of the ability to handle high dimensional data in data science, to extend the predictions to several asset classes. This study may provide insight from a more macro view in understanding asset returns dynamics and asset allocation.

Conclusion

This research provides a comprehensive study in equity return predictability. Research in important return predictors, feature engineering with PCA and scaling, feature selection, predictive models and trading strategy are covered. Each of the above studies progressively contributes to the success of developing an outperforming investing strategy suggested by this research. We show that future equity return is predictable, and data science can make outperforming the market possible.

REFERENCES

Lo, A. W. (2004, August 31). *The adaptive markets hypothesis*. The Journal of Portfolio Management. Retrieved May 1, 2022, from https://jpm.pm-research.com/content/30/5/15

Hull, B., & Qiao, X. (2017). A practitioner's defense of return predictability. *The Journal of Portfolio Management*, *43*(3), 60–76. https://doi.org/10.3905/jpm.2017.43.3.060

Fama, E. F. (1970, May). *Efficient Capital Markets: A review of theory and empirical work - JSTOR*. Retrieved April 30, 2022, from https://www.jstor.org/stable/2325486

Prado Marcos López de. (2018). *Advances in financial machine learning*. Wiley.

Neely, C. J., Zhou, G., Rapach, D. E., & Tu, J. (2010). Forecasting the Equity Risk Premium: The role of technical indicators. https://doi.org/10.20955/wp.2010.008

Xu, D., Li, B., & Singh, T. (2022). Does gold–platinum price ratio predict stock returns? international evidence. *International Journal of Managerial Finance*. https://doi.org/10.1108/ijmf-06-2020-0328

Ren, Y., & Xie, T. (2018). Consumption, aggregate wealth and expected Stock returns: A fractional cointegration approach. *Quantitative Finance*, *18*(12), 2101–2112. https://doi.org/10.1080/14697688.2018.1459809

Yang, D., and Q. Zhang. "Drift-Independent Volatility Estimation Based on High, Low, Open, and Close Prices." Journal of Business, 73 (2000), pp. 477-492.

Kenneth R. French - data library. (n.d.). Retrieved April 30, 2022, from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Authors Peter J. Bickel and Peter Bühlmann. (1969, December 31). *[bias, variance, and] arcing classifiers*. UCB Statistics. Retrieved April 30, 2022, from https://statistics.berkeley.edu/tech-reports/460

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. Routledge.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average?. The Review of Financial Studies, 21(4), 1509-1531.

Jolliffe, I. T. (2002). Principal component analysis for special types of data (pp. 338-372). Springer New York.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Luo, Y., M. Alvarez, S. Wang, J. Jussa, A. Wang, and G. Rohal (2014): "Seven sins of quantitative investing." White paper, Deutsche Bank Markets Research, September 8

Saeys, Y., Abeel, T., & Peer, Y. V. D. (2008, September). Robust feature selection using ensemble feature selection techniques. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 313-325). Springer, Berlin, Heidelberg.

Appendix

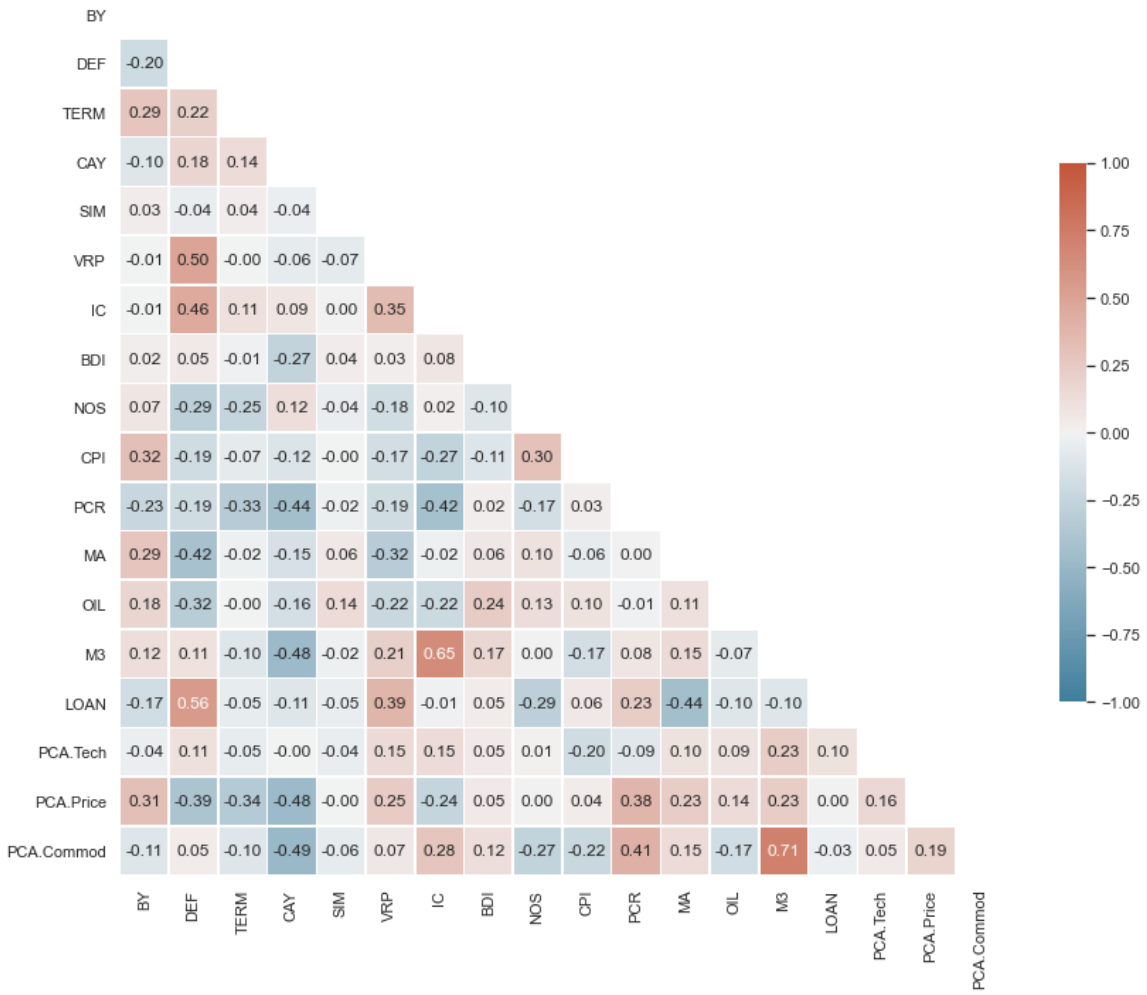


Figure 7. Correlation matrix of the predictors

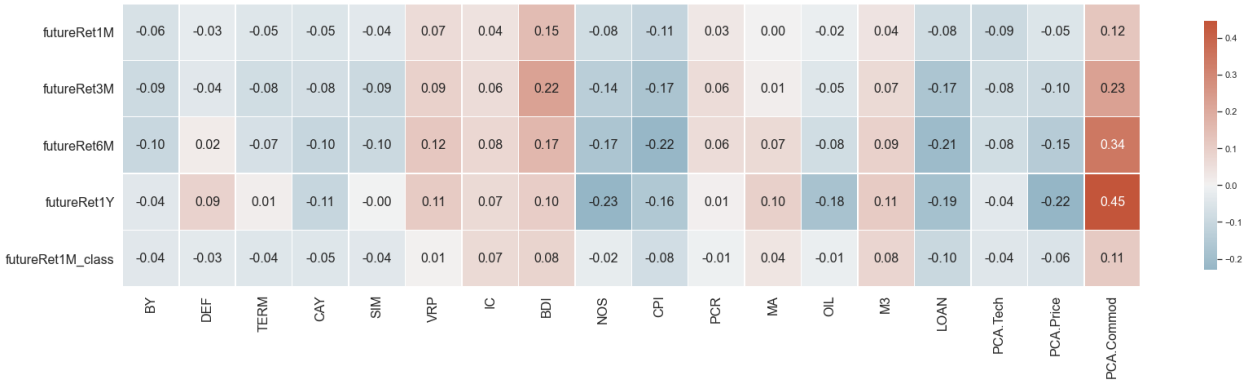


Figure 8. Correlation between predictors and features