Machine Learning for Predictive Modeling

Paul Zhao

Abstract

This research paper investigates the relationship between nearly a dozen proxies and the daily returns of companies on the Standard & Poor's 500 list. From this investigation, we followed a chain of procedures where we can re-predict historical data of S&P 500 daily returns by leveraging machine learning practices to produce a profitable investment portfolio, hence encapsulating our machine learning forecasting quantitative strategy. While there has been extensive research on the relationship between volatility and price, this study investigates the impact of other economics factors on daily returns: consumer sentiment (University of Michigan Sentiment Index), the ICE US Dollar Index (DX-Y.NYB), the exchange rate between the Euro and US Dollar (EURUSD=X), gold currency (GC=F), the Intercontinental Exchange Stock (ICE), US Treasury Bonds (TLT), volatility (VIX), Daily Put/Call Volume ratios of the S&P 500, Daily P/E ratios of the S&P 500, and interest rates (10 YR Treasury Yields, Federal Funds Rate). The use of educational journal articles and personal research kicked off this quantitative investment strategy: examining possible relationships between uncommonly tested independent variables and S&P 500 companies' returns. The underlying research problem was how to quantify and measure the relationships between the dependent and independent variables. This inspired the development of our random forests algorithm which allows for powerful classification through feature selection, quantifying how each independent variable impacts the returns of a selected company's value. This model incorporates lagged variables which enhances the autoregressive random forests framework and its ability to predict true variable values while mitigating the effects of forward-looking bias while removing the non-lagged version. As each ticker provides distinct historical data, the random forests model returns unique statistical metrics to evaluate model performance (for the sake of this paper, we will assume the BlackRock ticker, 'BLK', it is important to keep in mind these metrics reflect BLK, not the entire S&P). The following metrics act as strong indicators as to whether an investor should allocate capital towards the company they input into the model. The BLK model returned a mean squared error (MSE) value of 0.00024, a root mean squared error (RMSE) value of 0.01551, a mean absolute error (MAE) of 0.0104, and a R^2 score of 0.1509. Our MSE represents error as the difference between actual and predicted returns, meaning that a lower MSE indicates better performance. Our RMSE indicates that the average prediction error is around 1.55% in terms of returns and our MAE indicates that the average prediction error is around 1.05% in terms of returns. If these

three metrics are low in value, we can deduce that user_ticker company (BlackRock) may be a lucrative option to invest in. Our coefficient of determination indicates how well our model explains the variability of the response data around its mean. In BlackRock's context, approximately 15.09% of the variability in the company's returns is explained by the model, which is relatively low. This suggests that there is a significant amount of variability not explained by the model and there are other proxies not included in the model that may influence BlackRock's returns more heavily. Regarding feature importances, US Treasury Bonds proved to be the most influential feature (0.1415), suggesting that low-risk and interest-bearing securities play a significant role in BLK returns, followed by ICE_lag_1 (0.1076). Features with higher selection values suggest that historical data on the respective feature influences future conditions of the selected company more heavily. The ICE US Dollar Index feature value suggests that the strength/weakness of the US dollar impacts BlackRock's returns, possibly through the company's influence on global financial markets (0.1056). PUT_CALL_VOLUME_RATIO measures market sentiment and investor activity and suggests varying levels of hedging and market speculation impacts BlackRock's returns (0.0937). Following put/call option volume ratio, the P/E ratio shows moderate importance and suggests that valuation levels of equities play a role in predicting BLK returns (0.0669). As we can see clearly, features such as Federal Funds Rate, Euro-USD Exchange Rate, and Sentiment Index lay towards the bottom of the feature importances (0.0302, 0.0472, 0.0505). The combined influence of the collected proxies upon predicting daily return movement was measured through feature importance and would create a new portfolio based off tailored predicted returns from the past 10 years (**Figure 4**). This re-predicted portfolio was compared to a portfolio with its actual returns (100% BLK 01/2014 – 01/2024), a 100% SPX-invested portfolio, and a traditional 60/40 benchmark portfolio.

Introduction

The motivating factor behind leveraging machine learning to generate a profitable portfolio was based on market research and movement in predictive technology. Tracking the S&P 500 over short durations of time revealed that variations in daily returns tend to be minimal, which reflects the index's generally stable nature. Consideration towards the selected proxies were based off prior research that theorized variables which related strongly to index prices, leading to the final selection of the VIX Index, Eur-USD Exchange Rate, interest rates (10 YR Treasury Yields, Federal Funds Rate), S&P 500 Daily P/E Ratios, S&P 500 Daily Volume Put/Call Ratios, US Treasury Bonds, and the ICE US Dollar Index to emulate the Trade Weighted US Dollar Index. Proxies selected from personal hypotheses include the University of Michigan Sentiment Index, the Intercontinental Exchange Index (ICE), and gold currency. Sentiment data was used to assess whether bearish or bullish sentiment impacts potential future movements of the S&P 500 (positive sentiment correlates with rising prices). The ICE Index was added to assess whether performance trading volume and variation in financial instruments impacts future movements of the S&P 500. Lastly, gold currency was used as a feature since it is considered a "safe haven" asset that investors use in times of market uncertainty. Analyzing the correlation between these 11 proxies and the returns of S&P 500 companies could enhance understanding of the interactions between currencies, sentiment, exchange rates, interest rates, volatility, indexes, and the S&P 500 overall. After determining the strength and predictive power of each feature, the next step defined the essential research goal of the study: leveraging our random forests model results to predict unique historical returns on the same time-series data for a profitable portfolio purely from investing in a single company.

Methodology

The codebase takes in 10 years of trading data from 01/01/2014 to 01/01/2024, a period that captures phases of expansion, recession, and recovery. This allows the model to adapt to varying market conditions, enhancing its predictive accuracy and robustness across varying economic environments. It also measures the significant technological advancements and changes in the sectoral composition of the S&P 500. Through back-testing, a 10-year period also increases the statistical reliability of model results and financial metrics, minimizing anomalies.

Sentiment Index data was exported from the University of Michigan's Business School database online. Since sentiment is released monthly, the dataset was adjusted in a manner that added all 252 trading dates of the year for ten years and the sentiment index was constant for that entire month. For instance, the dataset's first original entry held the date (month of January 2014) with a listed sentiment index of 81.2 This entry was then converted to multiple entries (number of trading days in January 2014) where each date entry was assigned the same sentiment index. The subsequent months were transformed similarly. Afterwards, the sentiment dataset was joined with the Daily Put/Call Volume Ratio and Daily P/E Ratio data of the S&P 500 from Bloomberg Terminals. Similarly, these two variables were cleaned to handle all trading days of the same 10-year range. This lead to finalSentiment.csv, our three-variable dataset.

Time-series return data on the ICE US Dollar Index, the Euro-USD Exchange Rate, Gold Currency, ICE, US Treasury Bonds, the VIX Index, and the inputted ticker company (BLK) was pulled using yahoo finance based on each proxy's unique ticker. Their cumulative returns were compared over the 10-year span as time-series data (**Figure 1**). Two APIs (quandl & fred) were used to pull in two interest rates: US Treasury Yields and Federal Fund Rates. These daily rates were then combined with the daily returns of the ticker data to create a single dataset (combined_financial_data.csv) that held each ticker's daily return and the interest rates of its corresponding date. This dataset was then joined with finalSentiment csv, creating a dataset that is organized by date and holds data on sentiment, two S&P 500 ratios, our selected yahoo historical stock data, and interest rates (final_combined_data).

The next step was to implement the final_combined_data into a random forests model. Random forests modeling mitigates the risk of overfitting, allowing the results to be more generalizable to a higher number of features. This trait also ties into its ability to handle high dimensionality and non-linearity. Most importantly, random forests modeling provides built-in

tools for feature importance scoring, which helps properly rank our proxies. Additionally, this predictive modeling allows all features to initially contribute equally to the model training, avoiding more bias towards features with larger magnitudes. Lag variables were also used to prevent forward-looking bias and to help capture extra value out of the model as past values can be unhelpful when predicting immediate future movements. As such, the non-lagged version of the lagged variables (VIX & ICE) was removed from our final portfolio generation. Lag periods were selected based on autocorrelation, which helped determine the optimal delay in days that maximizes the explanatory power of the selected proxies on daily returns. Implementing the concept of lagged variables and lag periods was a crucial step during the back-testing process. After splitting the dataset into training and test splits, performance metrics and feature importances were then produced to generate meaningful insights.

A rolling window prediction model was then used to predict user_ticker's new daily returns based on our feature importance. Rolling window models work smoothly with financial time-series analytics to predict future values based on past data. A window size of 100 days was chosen to optimally balance and capture recent trends while avoiding the dilution of older, less relevant data and maintaining computational efficiency. The rolling window model continuously updated itself as new data became available to avoid reusing obsolete trends. The model used two for-loops that began at the first window and progressed in five-day increments, simulating a typical trading week. Within each cycle, the model would train on the most recent 'window_size days' and generated predictions for the next trading week, mimicking real-world scenarios when traders update their models based on latest available data at the end of each week.

A new RandomForestRegressor was then instantiated for each training increment. That training data was then used to predict user_ticker's returns, demonstrating the model's applicability in forecasting daily market movements. The predicted returns were then saved as a dataframe for further analysis and support of our investment strategy. It is important to highlight that modifying the window size days and rolling periods are crucial during the back-testing process to optimize model performance and accuracy. The predicted BlackRock data was then compared to an actual 100% BlackRock portfolio, a 100% SPX portfolio, and a traditional 60/40 BM (SPY & AGG) portfolio using ten-year historical data pulled through yahoo finance. The four portfolios were then compared on a time-series graph and through commonly used financial metrics generated from backtesting and benchmarking (**Figures 2 & 3**).

Results

Analysis of our Random Forests model's predictive performance on user_ticker's returns over a ten-year period reveals significant insights into the effectiveness of incorporating a diverse range of financial indicators as proxies for market behavior. The Mean Squared Error of the mode was computed at 0.00029, suggesting that the squared differences between the predicted and actual cumulative returns are small. Our Root Mean Squared Error was 0.0169, meaning that the average deviation of predicted cumulative returns from the actual values is around 1.69%. The Mean Absolute Error was 0.0115, indicating that, on average, the model's predictions deviate from the actual cumulative returns by about 1.15%. These three values (MSE, RMSE, MAE) being low in value represent an accurate model with reliably generated predictions. The model had an R^2 score of 0.0348, indicating that around 3.48% of the variance in cumulative returns is explained by our model. This score is relatively low, suggesting that there is a significant portion of variance the model does not explain (**Figure 5**).

The investment analysis shows that the newly generated portfolio outperforms our three comparative portfolios (100% Actual user_ticker returns, 100% SPX, 60/40 BM). The Predicted 100% Portfolio (P100%) generated an alpha of 0.188, which was substantially higher than the Actual 100% Portfolio (A100%) and the 60/40 BM, indicating that the model added returns over a standard benchmark. The Beta of the P100% was -0.0003, which was much lower than A100% (1.25) and 60/40 BM (0.6003), suggesting lower volatility and risk in comparison to the market. The Maximum Drawdown for P100% was -0.2868, which again, was less than the A100% (-0.439) but greater than the 60/40 BM (-0.2172), indicating that the 60/40 BM would have the best preservation of capital during downturns out of the three portfolios. The Sharpe Ratio for the P100% was 1.456, indicating a strong risk-adjusted performance in comparison to the other portfolios as A100% had a Sharpe Ratio of 0.532 and the 60/40 BM had a ratio of 0.666. A higher Sharpe Ratio in this case implies that the returns are not only higher but are achieved with a commendable control over volatility (**Figure 3**).

The cumulative returns chart illustrates a reasonable difference in performance between P100%, A100%, 100% SPX, and the 60/40 BM, particularly post-2019 which may be in part to the Coronavirus Pandemic (**Figure 2**). The ML generated portfolio not only showed resilience during market dips but also capitalized efficiently on market upswings, underscoring the efficacy of our systematic investment strategy.

Conclusion

This research provides a comprehensive overview of a user_ticker's (BLK) returns over a 10-year period using a sophisticated machine learning approach with randomforests and rolling window models. The study highlights the potential of machine learning and its capabilities in uncovering complex patterns when handling time-series data of financial indicators that are not readily apparent through traditional analytical means. The analysis included a comprehensive evaluation of our model's performance metrics and comparison against traditional investment benchmark portfolios. The model's success, evidenced by its performance metrics and positive simulated portfolio returns, provides a compelling case for the adoption of machine learning models in investment management. From our robust Sharpe Ratio and Alpha generated in the predicted portfolio, our findings suggest that modern predictive modeling can offer substantial enhancements to maximizing risk-adjusted returns compared to non-technical and conventional investment strategies.

The random forests model demonstrated its potential in generating accurate predictions through low MSE, RMSE, and MAE values (0.00029, 0.0169, 0.0115). Lower values indicate that the model's predictions are close to the actual values, reflecting its reliability. The model's R^2 of 0.0348, however, suggests that there are future steps to be done to help further explain the variance in the model.

The feature importance analysis revealed that US Treasury Bonds were the most influential predictor of user_ticker's returns. These values may vary between inputted tickers but are hypothesized to share a similar pattern. The lags (VIX & ICE) removed significant forward-looking bias and data leakage to ensure robust performance. The lagged ICE Index and US Dollar Index followed US Treasury Bonds in feature value, emphasizing the significance of macroeconomic indicators and currency strength in BlackRock's returns. When comparing P100% with A100%, the 100% SPX, and 60/40 BM portfolio, our predicted user_ticker portfolio outperformed in terms of Alpha and Sharpe Ratio, indicating higher risk-adjusted returns.

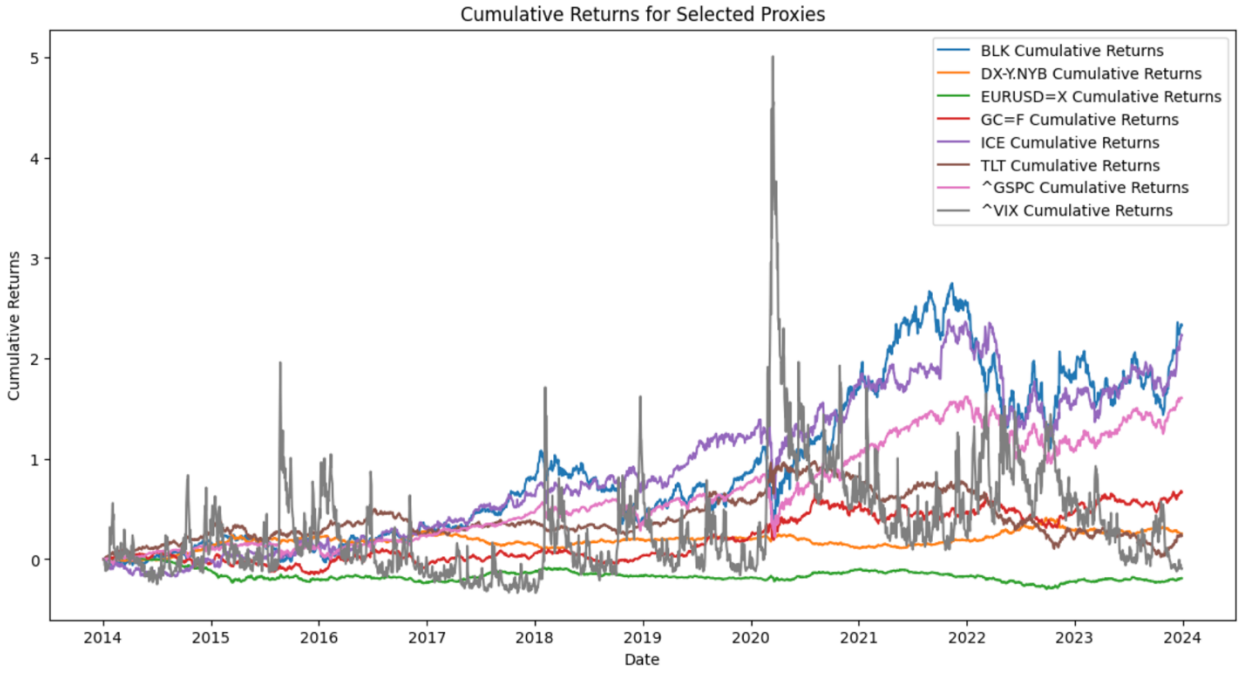
Despite these promising results, the study acknowledges multiple areas for improvement and future steps. The relatively low coefficient of determination indicates a failure to accurately measure variance in the model. Future research steps could explore expanding more features, model tuning, increasing robustness in back-testing, and implement dynamic feature selection. With an increased number of meaningful financial indicators that can be dynamically selected

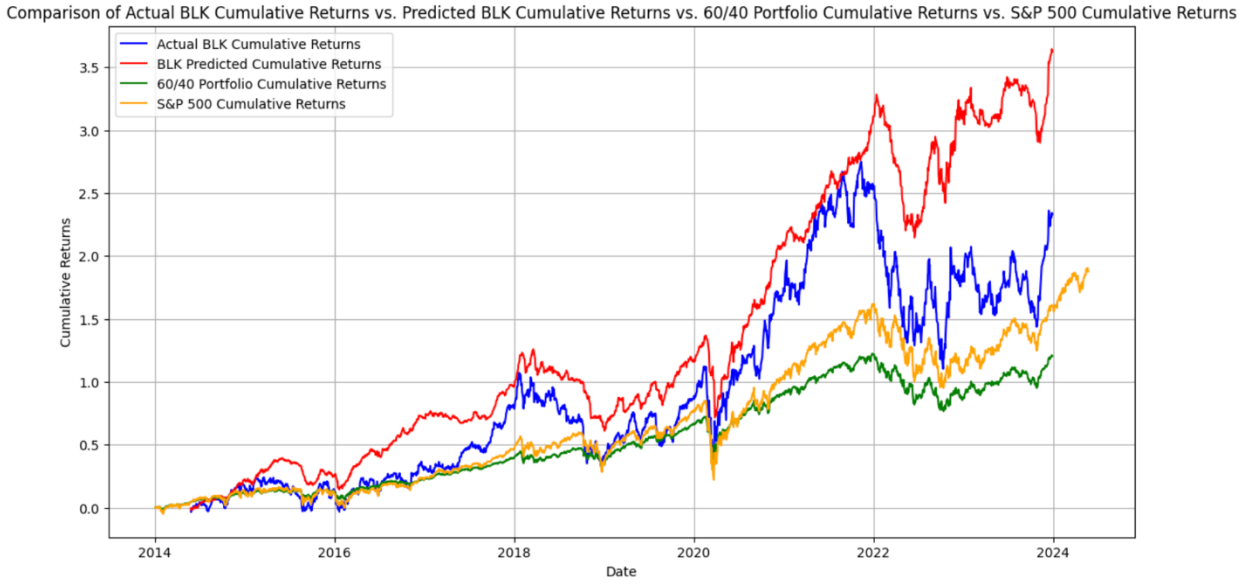
based on relevance and adaptive market conditions, the model is sure to improve in effective performance. Other primary independent variables that could provide significant insight towards future returns include historical put-call options ratios and PMI data. Diversifying the proxies may also provide a more holistic view of market dynamics and improve possible insights. Utilizing optimization and alternative machine learning models (Gradient Boosting, Neural Networks) may provide meaningful insights as well, depending on an investor's desired model specifications and computational capacities.

Looking ahead, the integration of real-time data processing and the exploration of additional predictive indicators represent promising areas for further research. The current randomforests model and investment strategy holds a heavy dependency on historical data. Therefore, as our current investment strategy leverages a ten-year period, future modifications could also include leveraging a larger period to increase explanatory power. Future research could also include exploring real-time trading systems that dynamically adjust to new data as timing plays a huge role in investing.

The application of Random Forests in this research not only supports its viability as a powerful tool for financial forecasting, but also sets a precedent for future studies to explore innovative data-driven approaches in finance. As financial markets continue to change, the integration of predictive technology will likely play an increasingly pivotal role in shaping and supporting effective investment strategies. This study ultimately lays the groundwork for future explorations and innovations in this intersection of finance and technology.

Appendix





	Alpha	Beta	Max Drawdown	Sharpe Ratio
Predicted 100% BLK Portfolio	0.1883206332904426	-0.00033293324973444725	-0.2867504966676964	1.4555521702232026
Actual 100% BLK Portfolio	0.02133501805412459	1.2557344797564165	-0.4390170367766011	0.5326123448506745
100% S&P Portfolio	0.0	1.0	-0.339249600026533	0.5951844324251704
60/40 Portfolio	0.012951912135374888	0.6002565599672507	-0.21716992743777722	0.6656418290371332

```
Mean Squared Error: 0.0002411984899007239
Root Mean Squared Error: 0.01553056630972367
Mean Absolute Error: 0.010446844239619323
R² score: 0.14898974284008715
R² score (direct method): 0.14898974284008715
                               importance
US Treasury Bonds                0.141063
ICE_lag_1                        0.107358
ICE U.S. Dollar Index            0.106340
PUT_CALL_VOLUME_RATIO_CUR_DAY    0.093809
BEST_PE_RATIO                    0.065897
Gold Currency                    0.056077
ICE_lag_2                        0.055780
10 YR Treasury Yields            0.055112
VIX Index_lag_2                  0.053021
Sentiment Index                  0.050388
Euro-USD Exchange Rate           0.048141
VIX Index_lag_3                  0.047170
VIX Index_lag_1                  0.046359
ICE_lag_3                        0.043820
Federal Funds Rate               0.029666
```

```
BLK - Mean Squared Error: 0.000286542747694831
BLK - Root Mean Squared Error: 0.016927573591475863
BLK - Mean Absolute Error: 0.011538480899406343
BLK - R² score: 0.03483663963734729
```