Assignment 6 Report: Machine Learning for Predictive Modeling

Paul Zhao

FIN 496-01

04/24/2024

Abstract

The relationship between volatility and price has formed the basis of thousands of studies into market theory, but the relationship between price and other not as popular proxies did not receive as much investigation. In addition to measuring volatility, other variables such as 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), the Volatility Index (VIX), and interest rates (10 YR Treasury Yields, Federal Funds Rate) upon the Standard and Poor's (S&P 500) daily returns. The selection of variables was carefully based upon prior studies in addition to an effort to examine possible relationships between uncommonly tested independent variables that may possibly impact S&P return data strongly. The underlying research problem was how to quantify and measure the relationships between the independent variables and the dependent variable and was done so using randomforests. Randomforests models allow for powerful classification through feature selection, quantifying how each independent variable impacted the returns of the S&P. The model incorporated lagged variables which enhanced the autoregressive randomforests framework and its ability to predict true variable values while mitigating the effects of forward-looking bias. The model returned an MSE value indicative of accurate predictions (4.003e-05) and an R^2 score of 0.71, indicating that 71% of the data variation is explained by the randomforests model (**Figure 1**). Regarding feature importance, the VIX Index (0.5124) proved to be the most influential feature which aligns with market research, suggesting that market volatility plays a significant role in our model, followed by the ICE Index (0.186). Features with higher selection values suggest that historical data on that selected market condition influences future conditions on the S&P Index. The first lagged ICE feature value suggests that recent historical data on commodities and derivates are also predictive of future movements (0.0443). US Treasury Bonds had a feature importance of 0.0366, the Sentiment Index had a feature importance of 0.028, the ICE US Dollar Index had a feature importance of 0.0267, 10 YR Treasury Yields had a feature importance of 0.0257, Gold Currency had a feature importance of 0.0217, the Euro-USD Exchange Rate had a feature importance of 0.02, and the Federal Funds Rate had a feature importance of 0.0117. The combined effect and influence of the independent variables upon predicting daily return movement of the S&P 500 was measured and allowed us to create a 100% SPX-invested

portfolio based off predicted returns from the past 10 years, comparing the portfolio with a 100% SPX-invested portfolio based on the actual cumulative returns and a traditional 60/40 benchmark portfolio.

Introduction

The motivating factor behind leveraging machine learning to generate a profitable portfolio was based on the market research and movement in predictive technology. Tracking the S&P 500 over short durations reveals 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 theorizing variables that related strongly to index prices, leading to the selection of the VIX Index, Euro-USD Exchange Rate, interest rates (10 YR Treasury Yields, Federal Funds Rate), US Treasury Bonds, and the ICE US Dollar Index to emulate the Trade Weighted US Dollar Index. The Sentiment Index was added to assess whether bearish or bullish sentiment impacts potential future movements of the S&P 500 (positive sentiment correlates with rising prices). The Intercontinental Exchange 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 added to the list of independent variables as it is considered a "safe haven" asset that investors use in times of market uncertainty. Analyzing the correlation between these nine proxies and the returns of the S&P 500 could enhance understanding of the interactions between currencies, exchange rates, interest rates, volatility, indexes, and the S&P 500. After determining the strength and predictive power of each feature, the next step defined the essential research goal of the study: leveraging the randomforests model results to predict unique S&P 500 returns on the same time-series data for a profitable portfolio purely from investing in SPX. These predictions were then compared to two other portfolios visually and numerically in terms of Alpha, Beta, Max Drawdown, Treynor Ratio, and Sharpe Ratio.

Methodology

The codebase takes in 10 years of trading day data from 01/01/2014 to 01/01/2024, a period that typically encompasses phases of expansion, peak, recession, and recovery. This

allows the model to capture and adapt to varying market conditions, enhancing its predictive accuracy and robustness across varying economic environments. The past decade also captures various regulatory and structural changes. This data allows the model to factor in the effects of such changes on market behavior and the S&P 500's movement. Additionally, it measures the significant technological advancements and changes in the sectoral composition of the market. In terms of back-testing, a ten-year period increases the statistical reliability of model results and financial metrics, minimizing anomalies.

The following libraries were used to pull in historical data, data analysis, data manipulation, mathematical operations, graphical visualizations, interest rate data (APIs), statistical modeling, statistical testing, and machine learning modeling (yfinance, pandas, numpy, matplotlib, plontnine, quandl, fredapi, statsmodels, scikit-learn).

Consumer Sentiment Index data from the same ten-year period was exported from an online source (University of Michigan) outside of the codebase. Since sentiment is only 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 heled 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), which each date entry being assigned the same sentiment index of 81.2. The subsequent months were transformed similarly until the end of December 2023, as a single entry for February 2014 became multiple training day entries during that month holding the same sentiment index of 81.6 (final_sentiment_data.csv).

Time-series data from this ten-year period on the ICE US Dollar Index, the Euro-USD Exchange Rate, Gold Currency, ICE, US Treasury Bonds, the VIX Index, and the S&P 500 was pulled using yahoo finance based on each proxy's unique ticker, calculating cumulative returns and comparing their historical movement (**Figure 2**). Two APIs (quandl & fred) were used to pull in two interest rates from the same ten-year period: 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 yahoo ticker's daily return and the interest rates of that date. This dataset was then handled properly to join in with the final_sentiment_data csv, creating a dataset organized by date and holds data on sentiment, our selected yahoo tickers, and interest rates (final_combined_data).

Random Forests was selected for the study's machine learning model. The algorithm mitigates risk of overfitting, allowing the results to be more generalizable to a higher number of features. This trait ties into its ability to also handle high dimensionality and non-linearity. Most importantly, randomforests provides built-in tools for feature importance scoring helping identify which of our selected proxies most significantly impact S&P returns. Additionally, the model allows all features to initially contribute equally to the model training, avoiding more bias towards features with larger magnitudes. Lag variables were also implemented 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. Lagging variables allow value measured from their delayed response to market stimuli. 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 the S&P 500 returns. After splitting the dataset into training and test splits, performance metrics and feature importances were then produced to generate meaningful insights.

To predict S&P returns based off our randomforests insights, an advanced implementation of a rolling window prediction model was used. This approach works well with financial time series analysis to predict future values based on past data. A window size of 100 days was chosen to optimally balance capturing recent trends with avoiding the dilution of older, less relevant data, while also maintaining computational efficiency. With a rolling window, the model continuously updates itself as new data becomes available to avoid reusing obsolete trends. The model uses a loop that starts at the first full window and progresses in steps of 5 days, simulating a typical weekly cycle effectively. Within each cycle, the model trains on the most recent 'window_size days' and generates predictions for the next trading week, mimicking

real-world scenarios when traders update their models based on the latest available data at the end of each week. A new RandomForestRegressor is then instantiated for each training cycle. That training data is then used to predict S&P 500 returns, demonstrating the model's applicability in forecasting daily market movements, creating a dataframe of predicted return data for our investment strategy.

The predicted S&P 500 cumulative return data was then compared to the actual cumulative S&P 500 return data alongside the cumulative returns of the traditional 60/40 benchmark portfolio (BM). The 60/40 BM data was pulled from yahoo finance using the stock and bonds' respective tickers (SPY, AGG) for the same ten-year period, calculating the weighted cumulative returns. The three portfolios were then compared on a time-series graph and through commonly used financial metrics generated from backtesting and benchmarking (**Figures 3 & 4**).

Results

Analysis of the predictive performance of our Random Forests model on the S&P 500 returns over a ten-year period for 2014-2024 reveals significant insights into the effectiveness of incorporated a diverse range of financial indicators as proxies for market behavior. Meaningful insights were generated after training the model with the selected set of features (market indexes, consumer sentiment, exchange rates, interest rates). The Mean Squared Error (MSE) of the mode was computed at 4.003e-05, indicating a very high accuracy in the prediction of daily S&P returns. The coefficient of determination (R^2) was measured at 0.71, suggesting that 71% of the variance in returns could be explained by the randomforests model. This level of explanation is significant and demonstrates the model's strong predictive power, especially for how large the time-series dataset is (**Figure 1**).

The feature importance analysis highlighted the VIX as the most influential factor, with an importance score of 0.5124, underscoring the critical role market volatility plays in influence

returns. The ICE Index followed with a. score of 0.186, implicating the significant impact of international trade dynamics and the US dollar's strength on the stock market. These findings align with current financial theories that suggest a strong linkage between market volatility, currency strengths, and stock market performance (**Figure 1**).

The investment analysis shows that the portfolio solely based on the predict S&P 500 returns significantly outperformed the traditional 60/40 BM over the observed ten-year period. The Predicted 100% Portfolio (P100%) generated an alpha of 0.1795, 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 0.0163, which was much lower than the A100% (1.0) and 60/40 BM (0.6003), suggesting lower volatility and risk in comparison to the market. The Maximum Drawdown for P100% was -0.2731, which again, was less than the A100% (-0.3392) 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 1.457, indicating a strong risk-adjusted performance in comparison to the other portfolios as A100% had a Sharpe Ratio of 0.545 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 return charts illustrate a clear difference in performance between the P100%, the A100%, and the 60/40 BM, particularly post-2019 which may be in part to the Coronavirus Pandemic. 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 the S&P 500's returns over a ten-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 financial data that are not readily apparent through traditional analytical means. The model's success, evidenced by its significant explanatory power and strong performance in simulated portfolio returns, provides a compelling case for the

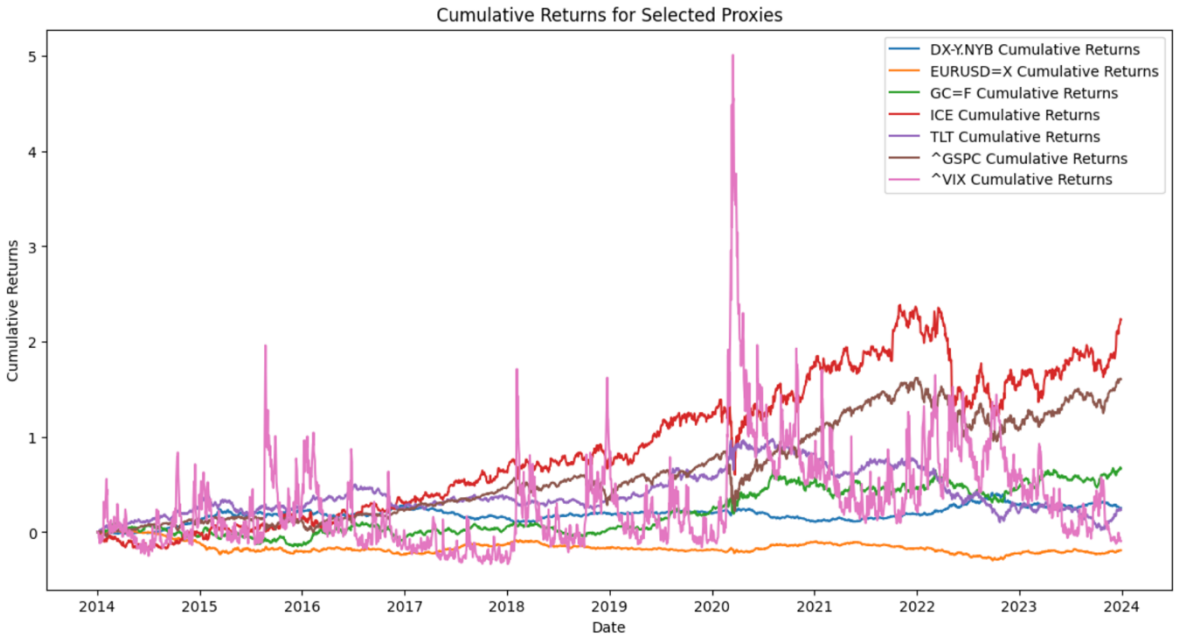
adoption of machine learning models in investment management. By achieving a robust Sharpe Ratio and Alpha in the predictive portfolio, our findings suggest that machine learning modeling can offer substantial enhancements to maximizing risk-adjusted returns compared to non-technical and conventional investment strategies.

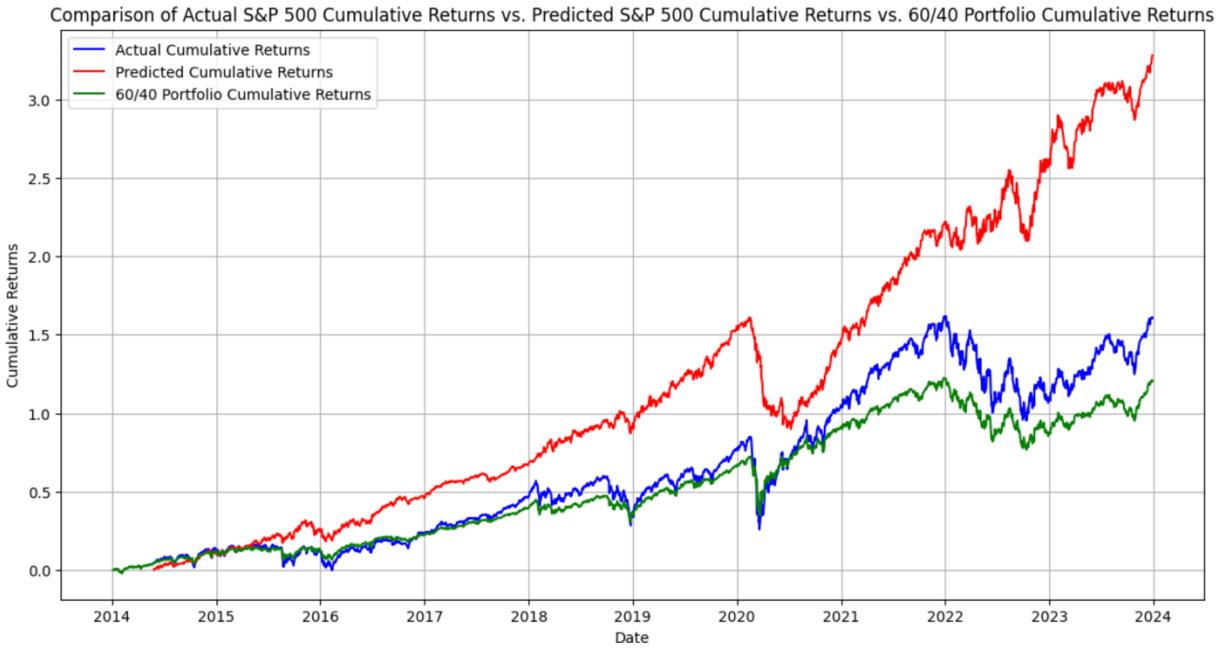
Looking ahead, the integration of real-time data processing and the exploration of additional predictive indicators represent promising areas for further research. The current randomforests model and investment strategy holds a heavy dependency on historical data. Therefore, as our current investment strategy leverages a ten-year period, future modifications could include leveraging a larger period to increase explanatory power. Other primary independent variables that can possibly provide significant insight towards future S&P 500 returns include historical put-call options ratios and Purchasing Managers' Index historical data. Additionally, analyzing historical data on P/E Ratios and inflation expectation may play a critical role in S&P 500 movement. As many other indicators exist, researching more independent variables and macroeconomic indicators could provide a more holistic view of market dynamics. Future research could also include exploring real-time trading systems that dynamically adjust to new data as timing plays a huge role in investing. Exploring other machine learning algorithms may also prove to be a better fit for our investment strategy depending on desired model specifications and computational capacities.

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 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 particular intersection of finance and technology.

Appendix

```
Mean Squared Error: 4.010114757052707e-05
R² score: 0.7133848225234216
R² score (direct method): 0.7133848225234216
                            importance
VIX Index                     0.512396
ICE                           0.186095
ICE_lag_1                     0.044227
US Treasury Bonds             0.036843
Sentiment Index               0.028038
Trade Weighted Dollar Index   0.026704
10 YR Treasury Yields         0.025727
ICE_lag_3                     0.021926
Gold Currency                 0.021743
Euro-USD Exchange Rate        0.020973
ICE_lag_2                     0.018549
VIX Index_lag_2               0.017020
VIX Index_lag_3               0.014381
VIX Index_lag_1               0.013666
Federal Funds Rate            0.011711
```





	Alpha	Beta	Max Drawdown	Sharpe Ratio
Predicted 100% Portfolio	0.1795	-0.0163	-0.2731	1.457
Actual 100% Portfolio	0.0	1.0	-0.3392	0.545
60/40 Portfolio	0.0129	0.6003	-0.2172	0.666