

Daniel Sanchez

BAN 525

Dr. Cetin Ciner

Module 4: Assignment 1 - Random Forest and Forecasting Stock Prices after Covid-19

June 6th, 2021

Introduction

Alarming, unprecedented, frightening, novel. COVID-19, the deadliest pandemic in over a century, has gripped individual households, businesses large and small, as well as the global economy. Such a widespread disruption is bound to have long-lasting effects, financial and otherwise. Planning for the future in times of uncertainty requires a more focused, nuanced approach. The Federal Reserve announcement in March 2020 to shore-up financial markets with large-scale purchases of corporate bonds reflected such a shift in investment strategy.

In order to fully understand and evaluate this decision, we can examine the stock market conditions surrounding the Federal Reserve's intervention. The data set for this analysis contains stock market information between January 2nd, 2020 and April 23rd, 2021. There are 20 variables represented; features include stock prices, U.S. Treasury bond prices, as well as exchange rates. The corresponding one-week lagged values of these variables are also included.

This report will feature two juxtaposed analyses of RSPY and Sign of RSPY. These variables of interest both represent the S&P 500 index in different ways. Analyzing the predictors of RSPY will help determine whether we can predict large company stock returns based on this data set. On the other hand, analyzing the binary variable Sign of RSPY will solve a classification problem. Analyzing Sign of RSPY will more explicitly forecast a recommendation to either buy (1) or sell (0).

The analysis in this report has been performed in JMP Pro 16, using the following methods:

- 1.) Ordinary Least Square (OLS) - for RSPY
- 2.) Nominal Logistic Regression - for Sign of RSPY
- 3.) Standard Decision Tree
- 4.) Random Forest

Analysis and Model Comparison

Ordinary Least Squares(OLS): This method will be used as a benchmark for the analysis of the continuous variable, RSPY. OLS fits a model by minimizing the squared differences between observed and predicted values. Since all of the model effects are included in the final model, overfitting can become an issue.

Nominal Logistic Regression: This method will be used as a benchmark for the analysis of the binary variable, Sign of RSPY. By retaining all of the data set variables in its results, this method also tends to overfit models to the training set.

Standard Decision Tree: This method creates a tree by separating the variables at each node, according to the largest disparity. This greedy approach does not always provide an optimal solution. This method is simply included in the analysis to be compared with the Random Forest model.

Random Forest: This method creates many decision trees, using a subset of variables to enforce an element of randomness. The resulting array of uncorrelated decision trees make predictions, and the one that makes the most accurate predictions is the model that is chosen.

The cross-validation procedure applied to this analysis accounts for the time-series nature of the data set. The data is split into three sections chronologically (as opposed to being randomized). Implementing a 60-20-20 split, the first 195 rows are used to train the models, the next 66 rows are tied to validation, and the final 66 rows are set aside for testing. This forward-looking approach to cross-validation helps to ensure that the model considers the sequential nature of the data set.




Both analyses were conducted in a similar manner. The first analysis aimed to predict the returns of the S&P 500, and RSPY was selected as the response variable. The second analysis looked to predict the recommendation to buy, and Sign of RSPY was selected as the response variable (with the target level of “1”). In both analyses, all of the other variables in the data set (both contemporaneous and lagging) were included as candidate predictors.

Using JMP Pro 16, this analysis was conducted with the assigned methods. As predictions from each model were generated, they were saved and stored in new columns in

the data set. The figure below displays a comparison of how all models performed with the test data.




Model Comparisons

RSPY Model Comparison - Test Data

Creator		RSquare	RASE	AAE
Fit Least Squares		0.3624	0.0076	0.0058
Decision Tree		0.2772	0.0081	0.0064
Random Forest		0.5564	0.0064	0.0048

The ideal RSPY predictive model in this comparison has the highest RSquare value, with the lowest RASE and AAE values. Random Forest was the highest performer in this group. Its RSquare value of 0.5564 was nearly 20 percentage points higher than that of the OLS model. The RASE and AAE values (0.0064 and 0.0048, respectively) were also lower than the other models in the analysis.

Sign of RSPY Model Comparison - Test Data

Creator		Entropy RSquare	Generalized RSquare	RASE	Mean Abs Dev	Misclassification Rate	AUC
Fit Nominal Logistic		-10.20	-1e+6	0.4464	0.2037	0.1970	0.7857
Decision Tree		0.3245	0.4803	0.3707	0.2522	0.1667	0.8600
Random Forest		0.4482	0.6144	0.3377	0.2332	0.1667	0.9079

The ideal Sign of RSPY predictive model in this comparison has the lowest Misclassification Rate and the highest AUC. In this case, the Random Forest model clearly outperformed the others. Although its Misclassification Rate of 0.1667 was equal to the standard Decision Tree model, the Random Forest model had a higher AUC of 0.9079.

The Random Forest models for each of these response variables will be more thoroughly examined for the rest of this analysis.

Interpretation - RSPY

To better understand the predictive capability of the Random Forest models, we will specifically highlight model parameters, column contributions and variable importance. The RSPY Random Forest model contained 26 trees, with 31 terms sampled per split. The model shows that the following variables are the highest contributors to the prediction of RSPY: RHYG, RVIX, REMB, RIEI, and RBTC. These Column Contributions are displayed below.

Column Contributions - RSPY

Term	Number of Splits	SS		Portion
RHYG	90	0.0271429		0.4339
RVIX	82	0.01074695		0.1718
REMB	39	0.00915375		0.1463
RIEI	31	0.00290021		0.0464
RBTC	15	0.00158757		0.0254

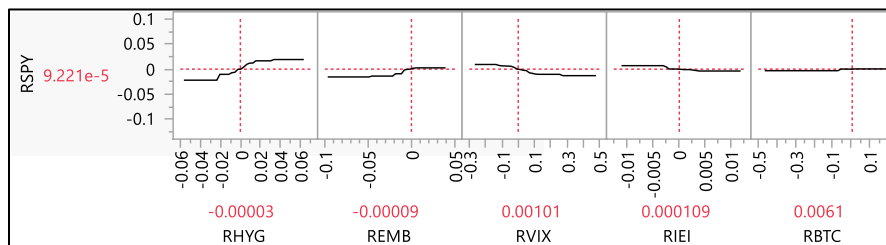
Further examining these contributing variables in terms of importance shows a slightly adjusted order. The largest total effects are as follows: RHYG (67.8%), REMB (19.1%), RVIX (11%). Although RIEI and RBTC were featured more prominently as Column Contributors, their total effects are only 1.2% each. No lagged effects had any significant impact to RSPY.

Variable Importance: Independent Uniform Inputs - RSPY

Column	Main Effect	Total Effect	
RHYG	0.656	0.678	
REMB	0.171	0.191	
RVIX	0.093	0.11	
RIEI	0.007	0.012	
RBTC	0.006	0.012	

The Random Forest prediction profiler for key predictors of RSPY is displayed below. Adjusting RHYG, REMB, and RVIX appear to have the largest impact to RSPY and other predictive variables. These adjustments do seem to impact most variables similarly, as a rising or falling tide. Changes to RIEI or RBTC do not visually appear to have a large effect on RSPY or the other variables.

Prediction Profiler - RSPY








This analysis practically reveals that, over the timespan of this data set, RHYG (High-Yield Corporate Bonds), and REMB (Emerging Markets) had a positive impact on S&P 500 returns. The RVIX (Volatility Index) had the greatest negative impact on the S&P 500. Interestingly, over this same timeframe, RBTC (Bitcoin) and RIEI (10-Year Interest Rates) show a nearly equal inverse relationship with their impact on S&P 500 returns.

Interpretation - Sign of RSPY





Turning to the Sign of RSPY Random Forest model, this contained 49 trees, with 31 terms sampled per split. This model shows that the following variables are the highest contributors to the prediction of Sign of RSPY: RVIX, RHYG, RFXC, REMB, and RFXA. No lagged effects had any significant impact to Sign of RSPY, either. The main Column Contributions are displayed below.

Column Contributions - Sign of RSPY

Term	Number of Splits	G ²		Portion
RVIX	69	60.8671597		0.4188
RHYG	58	27.8069865		0.1913
RFXC	28	10.3112243		0.0709
REMB	13	4.21512858		0.0290
RFXA	16	3.70744202		0.0255

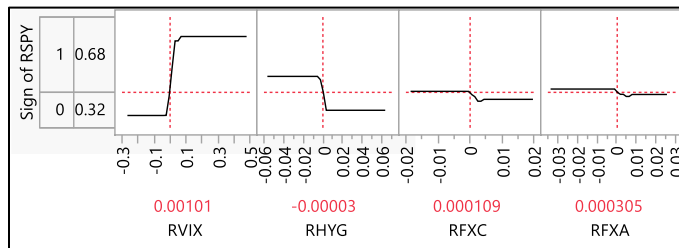
Variable importance analysis shows a clearer picture of the degree to which these top variables impacted the Sign of RSPY. Interestingly, RVIX has a very high total effect of 71.5%. RHYG comes in a distant second place, with 22.8%. RFXC is included here with a total effect of 10.9%. Finally, the total effect of RFXA is minimal, at 2.2%.

Variable Importance: Independent Uniform Inputs - Sign of RSPY Summary Report

Column	Main Effect	Total Effect	
RVIX	0.617	0.715	
RHYG	0.172	0.228	
RFXC	0.069	0.109	
RFXA	0.013	0.022	

The Random Forest prediction profiler for key predictors of Sign of RSPY is featured below. A change made to RVIX makes a dramatic change to Sign of RSPY and the other variables in the dataset. Adjusting RHYG also has a large impact to the other variables. Fluctuations in other variables appear to have very little impact on the overall picture. With RVIX at its peak, RFXC and RFXA start to exhibit their own volatility.

Prediction Profiler - Sign of RSPY



The analysis of Sign of RSPY in this data set shows that RVIX (Volatility) made the most difference in the model's predictive recommendations. As fear impacted the market, the model actually recommended buying, seizing upon the opportunity to invest at a discount. The Federal Reserve's efforts to purchase RHYG (High-Yield Corporate Bonds) appear to have improved performance of the S&P 500, motivating the model to recommend selling. The inclusion of RFXC (Canadian Dollar) and RFXA (Australian Dollar) in this model also alludes to commodities playing a minor supporting role to the impact of High-Yield Corporate Bonds.

Reference

Board of Governors of the Federal Reserve System. (2020, March 23). *Federal Reserve announces extensive new measures to support the economy* [Press release].

Retrieved from

<https://www.federalreserve.gov/newsevents/pressreleases/monetary20200323b.htm>