

Random Forest and Forecasting Stock Prices after Covid-19

Introduction

According to Investopedia.com, “Left to their own devices, free-market economies tend to be volatile as a result of individual fear and greed, which emerges during periods of instability.”¹ As a result to the economic shutdowns because of the global COVID-19 pandemic in 2020, countries all over were trying to keep their national economies afloat in a very, turbulent time. Exports were stopped because countries lacked the manpower to produce them as their workers were ordered to “stay home” and Imports dried up as a result of fear that the virus would arrive attached to the imported goods.

In the United States, the Federal Reserve had to act strongly to counteract the fear of its business world as almost every worker was ordered to “stay home” without being paid. Fear of default on mortgage loans, fear of default on business loans, and worries about how consumers would purchase goods without income was prevalent. As a response, the Federal Reserve, on March 23, 2020, committed itself to further actions to “purchase Treasury securities and agency mortgage-backed securities”² so that money in the market is increased as those payments continue while investors are fearful of loan defaults. The markets continued to see stocks bought or sold which keeps money flowing. Additionally, the Federal Banks lowering of interest rates encourages businesses and consumers to borrow money to keep the economy from imploding during a heightened time of fear. Therefore, the task at hand is to look at how that economic uncertainty plus Inflation, Exchange rates, Interest rates, Stock Market sectors such as oil and energy, impacted the U.S. Stock Market.

The given data set is dated from January 2, 2020, to June 3, 2022, with 611 observations in total. So, it covers a brief period before the National shutdown in March 2020 and reflects economic recovery in that “U.S. household consumption expenditures returned to their pre-pandemic trend by the second quarter of 2021.”³ The data set includes variables such as Treasury securities, international currencies, Stock Market Volatility, Inflation measures and Commodities such as oil and energy to predict the effect on the SPDR S&P 500 ETF (SPY) as well as determine the sign of SPY stock. Returns of these variables are calculated for us by using the differences of natural log prices to give a continuously compounded return with better distributional properties and Lagged Returns of these variables are also included to see if there are any delayed effects on the SPY stock.

¹ <https://www.investopedia.com/articles/economics/08/monetary-policy-recession.asp>

² <https://www.federalreserve.gov/newsevents/pressreleases/monetary20200323a.htm>

³ [The U.S. Economic Recovery in International Context | U.S. Department of the Treasury](#)

Random Forest and Forecasting Stock Prices after Covid-19

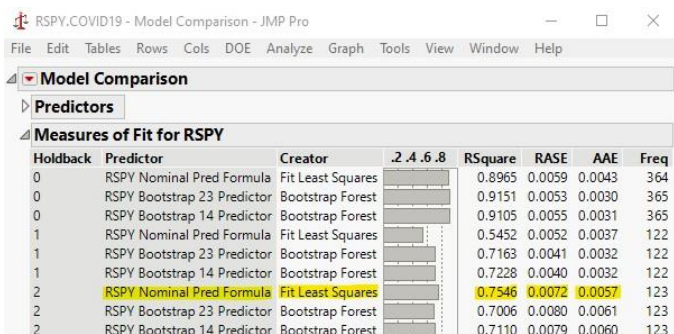
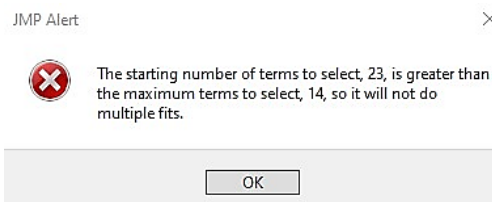
Analysis and Model Comparison

In preparation of the Data Analysis, there is a cross-validation column added to the end of the data set. The observations were broken up into a group with 60% Training Data, 20% Validation data, and 20% Testing Data, otherwise known as a 60/20/20 split. This was inserted into the data set as the “Holdback” or Validation column. SAS’s JMP software was then able to calculate statistics to be analyzed so that the most important variables could be chosen from the model to find predictive variables that would estimate how SPY stocks would act in the future. Two statistical models were compared: the Ordinary or Least Squares Method (OLS) which finds the values of the intercept and slope coefficient that minimizes the sum of the squared errors and the JMP’s Bootstrap Forest method which is based on the Random Forest (trademarked by Leo Breiman and Adele Cutler) which runs an ensemble of averaged arbitrary decision trees with a random number of predictor variables.

Additionally, included in the data set is a binary variable for the positive or negative sign depicting the direction of change applied to the SPY stock. For this nominal or categorical variable, represented by a 0 for negative direction and 1 for positive direction, the nominal logistic regression model which shows the relationship between a set of independent variables and a nominal dependent variable as well as the Bootstrap Forest method outlined above will be compared to find the best predictive model.

Part 1: Effects of the RSPY variable

A model was created using the OLS model first using the return variables and the lagged variables with the Holdback validation to compare to the RSPY value. Its results were saved to a column in the data set. Similarly, a Bootstrap Forest model was created to create a Random Forest model using the same parameters. At this time, the error message



Holdback	Predictor	Creator	.2	.4	.6	.8	RSquare	RASE	AAE	Freq
0	RSPY Nominal Pred Formula	Fit Least Squares					0.8965	0.0059	0.0043	364
0	RSPY Bootstrap 23 Predictor	Bootstrap Forest					0.9151	0.0053	0.0030	365
0	RSPY Bootstrap 14 Predictor	Bootstrap Forest					0.9105	0.0055	0.0031	365
1	RSPY Nominal Pred Formula	Fit Least Squares					0.5452	0.0052	0.0037	122
1	RSPY Bootstrap 23 Predictor	Bootstrap Forest					0.7163	0.0041	0.0032	122
1	RSPY Bootstrap 14 Predictor	Bootstrap Forest					0.7228	0.0040	0.0032	122
2	RSPY Nominal Pred Formula	Fit Least Squares					0.7546	0.0072	0.0057	123
2	RSPY Bootstrap 23 Predictor	Bootstrap Forest					0.7006	0.0080	0.0061	123
2	RSPY Bootstrap 14 Predictor	Bootstrap Forest					0.7110	0.0079	0.0060	123

on the right displayed yet it still calculated the result with default of 23 terms; it simply did not make models with random numbers of trees. Then a Bootstrap model with maximum terms set to 14 was created as the error message advised which allowed a random number of trees to be calculated. Both Bootstrap models were

Random Forest and Forecasting Stock Prices after Covid-19

compared to the OLS model using JMP's Model Comparison Tool and the additional columns added to the data table by our saved models. The OLS model performed better than both Bootstrapping methods as its R^2 value was higher and the Root Average² Error (RASE) plus the Average Absolute Error (AAE) were both lower. The default Bootstrap Forest Model without a random number of trees also outperformed the model suggested by the Error message.

Part 2: Effects on the RSPY +/- Sign

The second part of the data set comparison involved the binary variable column and the Nominal logistic Regression Model plus the Bootstrap Forest model. In JMP, when a binary variable is detected, it automatically sets the type of analyzation to the Nominal personality. The same Crossvalidation was created, and the results were saved to a column. Again, the Bootstrap Method gave an error and calculated without a random number of trees (default number was 24 trees). Then a 3rd model was created to

JMP Alert



The starting number of terms to select, 24, is greater than the maximum terms to select, 15, so it will not do multiple fits.

OK

set the maximum number of terms to 15 which allowed the random number of trees to be created. After comparing all 3 models, it was found that the default Bootstrap Forest method without the Random Number of Trees was the best model. The RASE rate was lower in the default model and both Bootstrap models had a 0.000 Misclassification rate, which is another measure of errors.

RSPY.COVID19 - Model Comparison - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help

Model Comparison

Target Sign of RSPY missing a predictor for category 0
Target Sign of RSPY missing a predictor for category 0
Target Sign of RSPY missing a predictor for category 0

Predictors

Measures of Fit for Sign of RSPY

Holdback	Creator	.2	.4	.6	.8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N
0	Fit Nominal Logistic					1.0000	1.0000	1.5e-7	0.0000	0.0000	0.0000	364
0	Bootstrap Forest					0.9858	0.9933	0.0097	0.0189	0.0095	0.0000	365
0	Bootstrap Forest					0.9792	0.9901	0.0142	0.0274	0.0138	0.0000	365
1	Fit Nominal Logistic					-0.419	-1.020	0.9509	0.2753	0.0806	0.0820	122
1	Bootstrap Forest					0.9788	0.9898	0.0142	0.0236	0.0139	0.0000	122
1	Bootstrap Forest					0.9750	0.9879	0.0167	0.0303	0.0163	0.0000	122
2	Fit Nominal Logistic					0.4782	0.6453	0.3596	0.1590	0.0272	0.0244	123
2	Bootstrap Forest					0.9853	0.9931	0.0101	0.0139	0.0100	0.0000	123
2	Bootstrap Forest					0.9724	0.9869	0.019	0.0353	0.0183	0.0000	123

Interpretation

Using the OLS Model, a deeper look was taken into the results of that model. The difference between the Cross-validation results said there was only a 11.5% difference between the Training and Test set. Looking at the Sorted Parameter Estimates on the next page, we can see that there were 3 return variables that were significant predictors of RSPY: RVIX, RHYG, and RVWO. We can see that the Volatility Index (RVIX), or "Fear Index", is the most influential in an inverse way. When the Fear Index goes up, the RSPY will probably go down. We can also see that the High-Yield

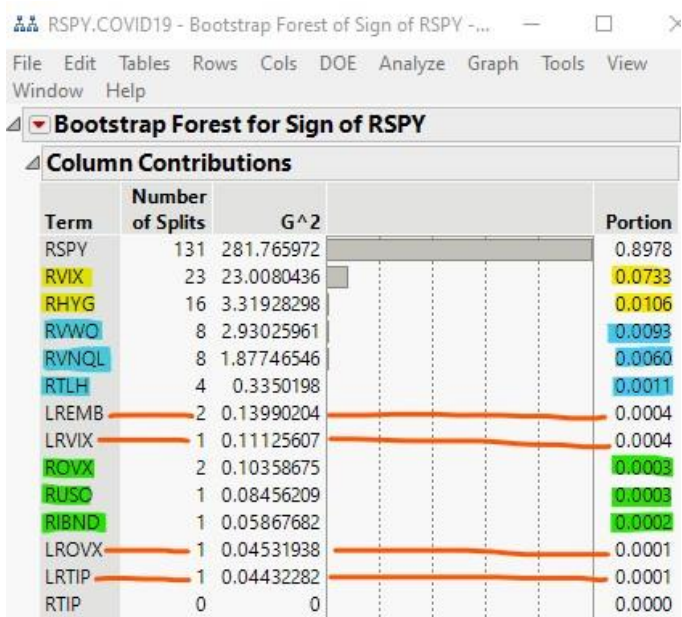
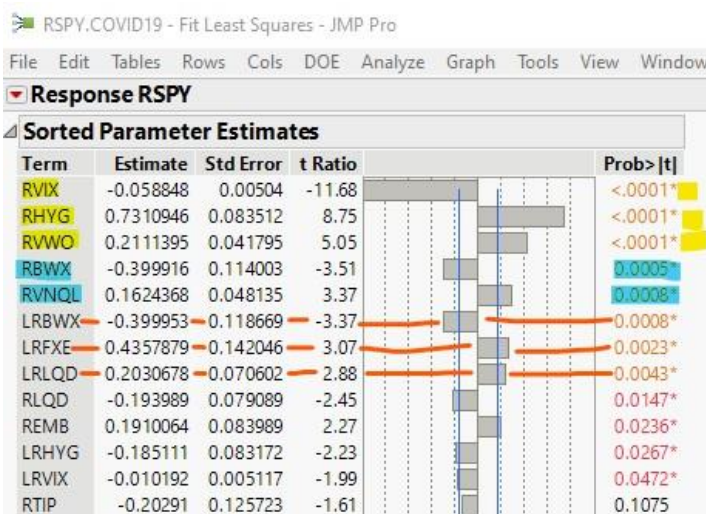
Random Forest and Forecasting Stock Prices after Covid-19

Corporate bonds (RHYG) influences RSPY in a direct relationship which means if one goes down, so does the other. Additionally, the Vanguard Emerging Market (RVWO) also represents the Interest Rates/Bonds group and similarly holds a direct relationship with RSPY.

The second group representing the less influential variables includes the BWX Technologies (RBWX), the Vanguard Real Estate Index Fund (RVNQL), and three Lagged returns for BWX

Technologies (LRBWX), the Euro Currency Rate (FXE), and the Investment grade Corporate Bonds (LRLQD). RBWX and LRBWX demonstrate an inverse relationship similar to the RVIX or “fear index” above. The fact that the return and lagged return cover the same sector suggests that the technology sector represents a longer effect than RVIX. This also means that if the SPY goes up, an investor might start selling technology stocks as it represents a tradable equation.

Looking at the results of the Sign of RSPY Bootstrap Forest Model, the Column Contributions tool is used to show us which variables influence the direction of SPY. Earlier in the exploration stage, Bootstrap models including RSPY and excluding it were compared for accuracy. There was dramatic differences in accuracy so the decision was to keep it in the final model. However, in interpretation, it is not going to affect its own sign, so the choice is to ignore it in this chart.



However, the Volatility Index (RVIX) must scream out its importance in this chart. Claiming a large 7.33 percent of the effect on the Sign of RSPY, it shows that because of its inverse relationship previously described, an investor could be convinced to buy if this index was high because the prediction for a rocky market in the S&P (RSPY). In contrast, the second most influential variable is the High-Yield U.S. Corporate bonds (RHYG) aka “junk bonds” which has a direct relationship with RSPY. So if RHYG was high, it would be a sign to sell because businesses were showing positive trends on the S&P (RSPY).

Random Forest and Forecasting Stock Prices after Covid-19

There were also 11 other variables considered to impact the sign of RSPY. These all were under 1% but it is interesting to note that 4 of them were Lagged returns: the International (Emerging market) bonds from places like Turkey, Taiwan & Malaysia (LREMB), the Volatility Index (LRVIX), the Crude Oil Volatility Index (LROVX), and the Inflation Measure (LRTIP). Along with Lagged returns were the returns for Vanguard Emerging Market (RVWO), the Vanguard Real Estate Index or VQL (RVNQL), 10-20 year U.S. Treasury Bonds (RTLH), Crude Oil Volatility Index (ROVX), US Oil Prices (RUSO), and SPDR Bloomberg International Corporate Bonds (RIBND). So that shows that Crude Oil (ROVX and LROVX) have a combined 0.04% effect on the sign of RSPY.

Conclusion

It is still fair to say that an investor should watch the fear index when predicting the S&P and look to the junk bonds raising or falling in the direction the S&P will go. The VIX will inversely predict the S&P so if the S&P is going up, the economy is more stable, and the Volatility Index will be down. Likewise, the Junk bonds in HYG will go up and these stocks will be bought a lot in a good economic climate. As a final note, if the Federal Bank feels that the volatility index is demonstrating that the economy is in a freefall, as it was in March 2020, they will use these indexes to step in and intervene to keep money flowing through the Stock Market.