Alec Engel
June 4, 2021

Introduction

The question at hand here is to determine what if any financial market variables play a role in the prediction of our return on US stock index (RSPY) and the sign of our return on US stock index (Sign of RSPY) during the Covid-19 pandemic era. We want to run several cross-validation models to see what variables can be strongly correlated with our response variables. These cross-validation methods will allow us to split our dataset into training, validation, and test portions. The split allows for better predictions of future outcomes with reduction in potential overfitting and by removing unnecessary model noise amongst highly correlated independent variables. We will then take a look our variables using ordinary least squares, nominal logistic regression, and random forest model structures.

Predictor variables consist of both continually compounded ETF weekly returns and one week lag prices for different currencies, bonds, crypto currencies and stock markets. Each variable at hand is converted into percentage returns and then turned into natural log prices and one week lag prices. The most important variables in this data set are the response variables of RSPY and sign of RSPY; along with predictor variable returns of high yield US corporate bonds (RHYG), International bonds (REMB), market volatility (RVIX) and Canadian fx rates (RFXC). As we will elaborate later, the fear index and junk bonds will have highly unique information as it pertains to the correlation with Covid-19 era US stock index returns.

In order to narrow down our predictor variables, we ran several statistical models and judged our outcomes. Statistical methods include cross-validation tactics of Ordinary Least Squares (OLS), nominal logistic regression, and random forests. The latter method of random forest produces extremely powerful and telling models that create a solid narrative when dealing with nonlinear data. Random forest uses the uncorrelated random method of tree sampling to select relative variables while attempting to reduce any overfitting or distributional assumptions. The average of each tree is then used to build a powerful predictive model that typically outperforms its competition when dealing with future (test) data tendencies.

Analysis and Model Comparison

For the RSPY response variable, we first used OLS which involves the inclusion of all variables. While using all variables, there are often noisy and unnecessary data variables that are accounted for. For the sign of RSPY response variable, we first used nominal logistic regression with "1" representing positive stock returns and "0" representing all other stock return outcomes. We quickly found that although assessing penalties on our variables, logistic regression often times leads to overfitting or misuse of highly correlated variables. Our random forest model performed much better while taking a nonlinear approach versus the linear and more biased logistic regression method. To get a clearer picture and better predictive abilities, we take more stock in our random forest models. Random forest will eliminate tree correlation risk and emphasize the best variables to use for future model predictability.

The time series dataset also includes a cross-validation column which has been renamed "Holdback". This splits our data into training (60%), validation (20%) and testing (20%) sets to allow our model to gain predictive attributes and score our R squared values and misclassification rates for how we believe future data will look. After cross-validation, we created predictor formulas with each of the above three methods described. Each method was then added as a dataset column so that we can easily perform model comparisons.

Model comparison results can be seen below. First with the zero "Holdback" notations, we see all four models having moderately high performance. All four are above 0.9 R squared (RSPY) or have an extremely high AUC (for sign of RSPY) with very small route average standard error and low misclassification rates. These however are our training sub sets of data and we want to focus more towards our validation and testing metrics. Moving to one and two "Holdback" notations, we see RSPY R squared values have dipped a considerable amount but the clear winner is our RSPY random forest model which explains nearly 56% on our unbiased test data (R squared of 0.56). When looking at the sign of RSPY, we are focused on a high area under the curve (AUC) and a low misclassification rate. We can clearly see over a 10% increase of AUC between regression model and random forest as well as a lower classification rate. Based on our sign of RSPY test results, we will go with our random forest model having an AUC of nearly 91% and a misclassification rate of only 0.16. Although I will dig into each model to investigate strongest column contribution variables, I want to choose the strongest models here for both dependent variables of random forest as my predominant models.

Model Comparison

Predictors

Measures of Fit for RSPY

Holdback	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
0	Pred Formula RSPY OLS	Fit Least Squares		0.9234	0.0066	0.0051	193
0	RSPY Predictor RF	Bootstrap Forest		0.9052	0.0073	0.0040	194
1	Pred Formula RSPY OLS	Fit Least Squares		0.4043	0.0074	0.0057	65
1	RSPY Predictor RF	Bootstrap Forest		0.6412	0.0057	0.0046	65
2	Pred Formula RSPY OLS	Fit Least Squares		0.3597	0.0076	0.0058	67
2	RSPY Predictor RF	Bootstrap Forest		0.5592	0.0063	0.0047	67

Model Comparison

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	AUC
0	Fit Nominal Logistic		1.0000	1.0000	2e-7	0.0000	0.0000	0.0000	193	1.0000
0	Bootstrap Forest		0.7825	0.8807	0.147	0.1835	0.1242	0.0103	194	0.9992
1	Fit Nominal Logistic		-9.528	-7e+5	7.2665	0.4641	0.2157	0.2154	65	0.8062
1	Bootstrap Forest		0.4596	0.6276	0.373	0.3497	0.2410	0.2154	65	0.9162
2	Fit Nominal Logistic		-10.07	-1e+6	7.5218	0.4430	0.2007	0.1940	67	0.7875
2	Bootstrap Forest		0.4543	0.6199	0.3709	0.3351	0.2301	0.1642	67	0.9093

Interpretation

Per below, we can see that four column contribution variables were used to maximize our overall model performance on test data. Each unbiased random forest tree model was created and combined to highlight these key variables of importance. The parameter estimates highlighted in our model are RHYG, RVIX, REMB and RFXC. We can also see that each of these four significant parameter estimates used account for 75% of our model predictions.

Bootstrap Forest for RSPY

Prediction Profiler

Variable Importance: Independent Uniform Inputs

Column Contributions

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
RTLH	18	0.00155029	0.0248
RLQD	19	0.00114415	0.0183
LRSHY	17	0.00106336	0.0170

Bootstrap Forest for Sign of RSPY

Prediction Profiler

Variable Importance: Independent Uniform Inputs

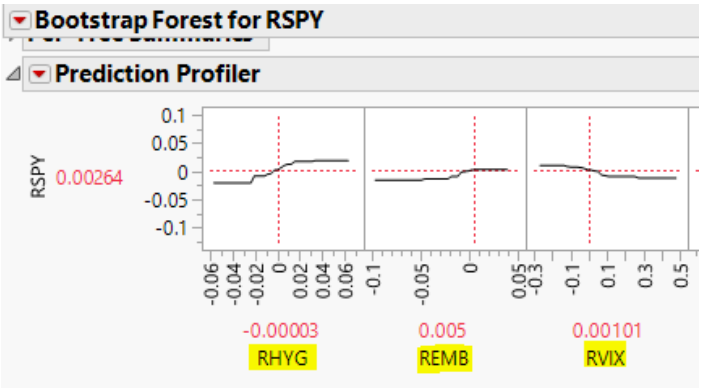
Column Contributions

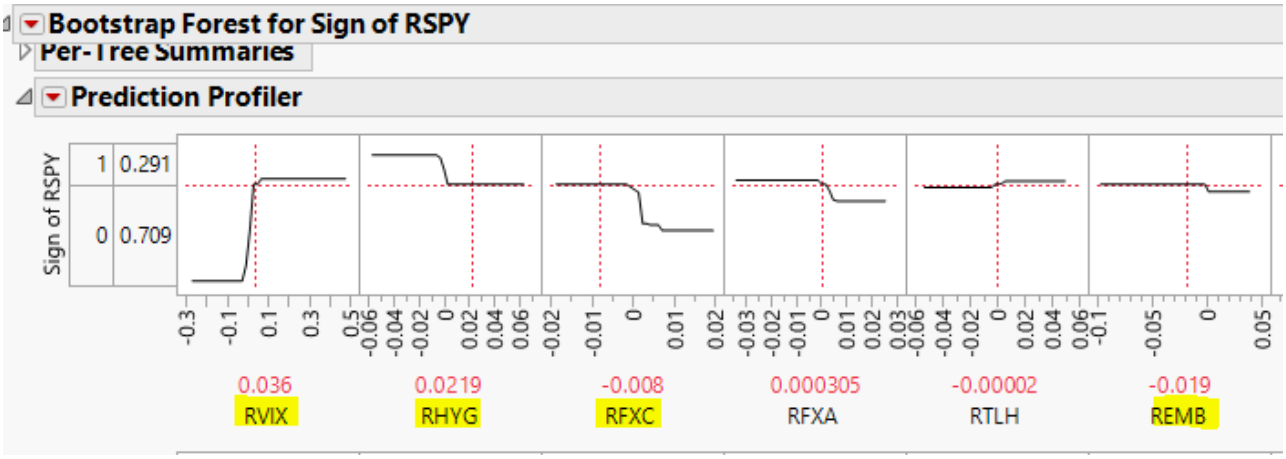
Term	Number of Splits	G^2	Portion
RVIX	69	60.8671597	0.4188
RHYG	58	27.8069865	0.1913
RFXC	28	10.3112243	0.0709
REMB	13	4.21512858	0.0290
RFXA	16	3.70744202	0.0255
RTLH	18	3.35507828	0.0231
RUSO	20	2.75257857	0.0189
LRSPY	13	2.72933712	0.0188

RHGY and RVIX will be our most important parameter estimates as they show both high estimate numbers as well as low standard errors. For RSPY, we see that as junk bond purchases increase with brokers feeling good about the markets, stock returns look to simultaneously rise. The opposite can be said for RSPY's correlation with VIX. Market volatility and this "fear index" shows that as volatility increases, we get lower stock returns. Although nominal, we see the same results for the sign of RSPY. As junk bond investments increase, we see a stronger confidence in positive stock market returns and as VIX volatility increases, we predict the opposite.

We can also see a positive correlation with confidence in International bond investments (REMB) and stock market increases for RSPY but we seem less confident with this variable prediction when using the sign of RSPY. With the sign of RSPY, REMB increases as we become less and less confident in our prediction of US stock market positive returns. This may again show volatility and minimal correlation between the US stock market and emerging market economy trends. The Canadian dollar fx rate also looks to cause an unfavorable view on positive US returns as its rate increases.

RHYG is our strongest contributor and variable for predicting RSPY while RVIX is our strongest contributor and variable for predicting sign of RSPY. Each variable shows strong ties with our US market performance and ability to accurately predict positive returns. Some consideration can be given to REMB with just over 14% variation regarding its slightly positive relationship and RFXC regarding its slightly negative relationship with US stock returns. But again, RHYG and RVIX are our strongest predictor variables as their relationships explain over 60% of variations in the US stock market.





Following the Federal Reserve's March 20, 2020 announcement regarding their bond market intervention, I am observing much greater variation in day-to-day peaks and valleys. Each day, we see greater returns of our two key variables both in positive and negative directions. There was obviously much more action in the US stock market after this announcement and we can see a noticeable increase in market volatility and more bond market movement. With the Fed's actions we can see junk bond purchasing decreasing and our "fear index" of volatility increasing which both correlate and contribute to our US stock market price declines.