Christopher Dillard 9/24/2023

BAN 525

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

In this assignment, I aimed to evaluate the forecasting capabilities of non-linear models on stock returns. Unlike in previous assignments dealing with these financial instruments, my focus is solely on data from the period directly following the onset of the Covid-19 pandemic. Altogether, the dataset contains 611 rows spanning from January 2nd, 2020, to June 3rd, 2022.

My two response variables are "RSPY" and "Sign of RSPY", where RSPY reflects daily stock returns for SPY (SPDR S&P 500 ETF Trust) and the "Sign of RSPY" variable simply tells us whether the returns in the RSPY column were positive or negative for any given day, with 0 denoting negative returns and a 1 signifying positive returns. When constructing the models, I selected all Rs and Ls as my X variables, otherwise referred to as predictor/explanatory variables. The only exception here is that "RSPY" was not entered into the models predicting the "Sign of RSPY", and vice versa, since their clear correlation between each other renders them extraneous. With this in mind, I selected 29 explanatory variables when building each model. The variable nomenclature follows a logical pattern where any Rs preceding the Stock Symbol identifies the returns, being the daily changes in the stock valuations (ie. Contemporaneous effects). On the other hand, lagged effects are represented by Ls preceding the Stock Symbols in the variable names. Lagged effects are important in the stock markets for traders since they can serve as determining factors as to whether stocks should be bought, held, or sold to incur the highest growth (profits) and minimize losses as valuations fluctuate over time.

In this assignment, I employed two different types of non-linear models in JMP: Decision Trees, called a Partition model in the software, and Random Forest, referred to as a Bootstrap Forest model.

One of the biggest differences between these two models is in how they deal with the variable selection process. For Decision Trees, variables are selected by finding the best split that minimizes the misclassification rate when the response variable is categorical in nature (the "Sign of RSPY" in our case) or minimize the Mean Standard Error (MSE) when we are dealing with a regression problem (the "RSPY" response variable). This criteria is maintained for as many splits as the model deems appropriate for the data. One potential problem here is that Decision Trees might produce too many splits, which tends to lead to overfitting. We can test for this phenomenon by evaluating the model's R-square value for the regression problem or the Misclassification rate for the classification problem.

Random Forests, like the name implies, can be considered much more computationally intensive collection of Decision Trees, where the optimal model is constructed by randomly selecting subsets of rows and cycling through random selections of predictor variables. The final model can be decided in two different ways, depending on whether the response variable lends itself to either a classification or a regression problem. In the model dealing with RSPY as the response variable, JMP will take an average of all the predicted returns. As for the "Sign of RSPY" model, the difference is that JMP will take a majority-vote of the predictions, which is either 1 (positive) and 0 (negative) for the predicted returns, due to how we should not logically expect any predictions between 0 and 1. Due to how Random Forests create randomized subsets of data, especially in terms of how different selections of variables are used in the forests, these models are much better at avoiding overfitting and, therefore, are very capable at handling new data.

Model Comparisons for RSPY:

Holdback	Predictor	Creator	RSquare	RASE	AAE	Freq
0	RSPY Predictor DT	Partition	0.6900	0.0102	0.0063	365
0	RSPY Predictor RF	Bootstrap Forest	0.9132	0.0054	0.0032	365
1	RSPY Predictor DT	Partition	0.6227	0.0047	0.0036	122
1	RSPY Predictor RF	Bootstrap Forest	0.7064	0.0041	0.0033	122
2	RSPY Predictor DT	Partition	0.4847	0.0105	0.0079	123
2	RSPY Predictor RF	Bootstrap Forest	0.6740	0.0084	0.0065	123

Based on the output from this table, I selected the Bootstrap Forest (Random Forest) as the superior model since it produced an R square value of 0.674 on the testing set (corresponding to the 2s in the Holdback column).

Non-linear models do not allow analysts to measure variable significance using parameter estimates, like in previous assignments. That being said, JMP does offer several other ways for analysts to measure variable importance, in addition to the previously seen Independent Uniform Input table from the model's Prediction Profilers. One tool that was created specifically for Random Forests in JMP is the "Column Contributions", which tells us what percentage of the daily fluctuations in the valuations of the SPY stock are explained by each predictor variable, as demonstrated in the table below:

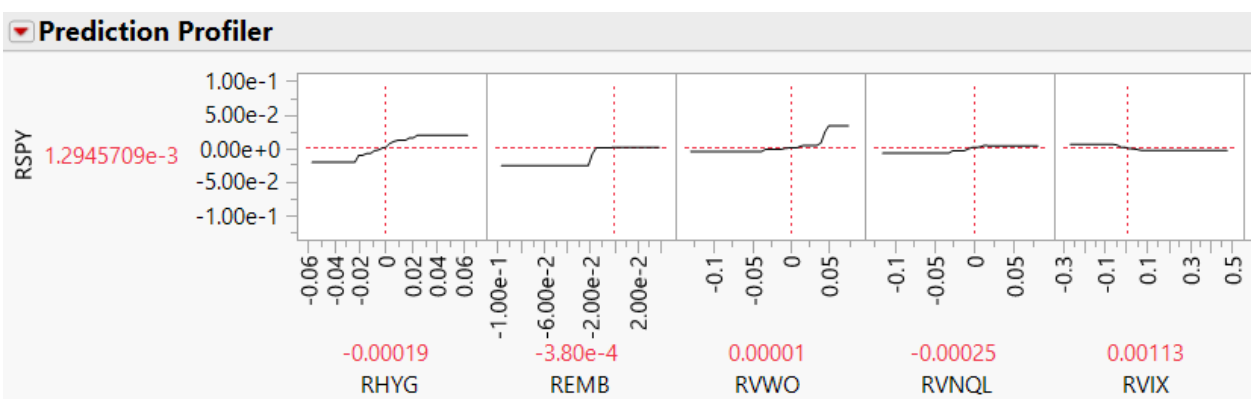
Column Contributions: Bootstrap Forest for RSPY

Term	Number of Splits	SS	Portion
RHYG	67	0.02817719	0.3629
RVWO	58	0.01441578	0.1856
REMB	32	0.01261443	0.1625
RVNQL	51	0.00779644	0.1004
RVIX	79	0.00749683	0.0965
LRHYG	29	0.00076693	0.0099
RTLH	20	0.00068206	0.0088
RUSO	19	0.00058439	0.0075
RLQD	22	0.00057117	0.0074
LRTIP	27	0.0004949	0.0064
LRLQD	28	0.00045417	0.0058
ROVX	27	0.000435	0.0056
RFXE	23	0.00030124	0.0039
LRUSO	14	0.00025269	0.0033
LRBWX	25	0.00024139	0.0031
LRTLH	19	0.00023912	0.0031
RTIP	18	0.00021785	0.0028
RIBND	27	0.00020625	0.0027
LRIBND	25	0.0001971	0.0025
RGLD	27	0.0001926	0.0025
LREMB	17	0.00019229	0.0025
LROVX	20	0.00018191	0.0023
LRSLY	28	0.00018166	0.0023
LRVIX	22	0.00017491	0.0023
LRFXE	16	0.00016267	0.0021
RBWX	17	0.00013184	0.0017
LRGLD	15	0.00010802	0.0014
LRVNQL	19	0.00009196	0.0012

Term	Number of Splits	SS	Portion
LRVWO	12	8.7948e-5	0.0011

The 5 most important variables in the model, in descending order, are: RHYG, RVWO, REMB, RVNQL, and RVIX. RHYG, being the most important, explained roughly 36% of the fluctuations in the RSPY column. RVWO and REMB accounted for a little under 20% each, while RVNQL and RVIX were around 10%. All other variables had an extremely small impact on the model's predictions, so I will only focus on the aforementioned 5 variables later in this paper in terms of what they represent.

One limitation when relying on the Column Contributions table, is that the output does not tell us the direction of each variable's impact, which would otherwise help us to see if there is a positive or negative relationship between each predictor variable and the RSPY variable. So, to evaluate these effects, we must look at the Predictions Profilers. In this model, many variables had a very minute impact on the model, so paying close attention to the slope of each graph gives an additional layer of insight into how daily changes in these variables correlate SPY's returns. In the screenshot below, I have only included the 5 most impactful variables from the previously discussed Column Contributions graph:



From the profilers, we can see that RHYG, REMB, RVWO, and RVNQL had positive slopes while RVIX had a negative slope.

Model Comparison: Bootstrap Forest for Sign of RSPY

The second Response variable in this assignment is the “Sign of RSPY”, which contains categorical data. This column takes the values from the RSPY column and represents them as either a 1, meaning stock returns are positive, and 0, denoting a devaluation of the stock. This knowledge can be used by traders to create a strategy that better suits the level of risk they are willing to take in the market. Some may follow a more straightforward approach with a predicted 1 influencing them to buy the stock and, alternatively, a 0 to sell the stock.

The first tool I used to compare the Decision Tree with the Random Forest model was the Measures of Fit tool. The output, contained in the following screenshot, only provides helpful information on each model’s Misclassification Rate. In this case, the Random Forest model was the better choice again given that it only had a Misclassification Rate of 0.13 while the Decision Tree had a value of 0.20. These measures tell us how often the models were incorrect in their prediction on the testing data.

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			.	.	.	.	.	0.1699	365
0			.	.	.	.	.	0.0274	365
1			.	.	.	.	.	0.1557	122
1			.	.	.	.	.	0.2049	122
2			.	.	.	.	.	0.2033	123
2			.	.	.	.	.	0.1301	123

To run further model comparisons on this data, I also looked into other tools in JMP to evaluate model performance like the Confusion Matrix which is uniquely helpful for evaluating the performance of classification models.

Confusion Matrix

Target Sign of RSPY

Predictor

0

Actual Sign of RSPY	Predicted Count 0	Predicted Count 1
0	105	51
1	11	198

Actual Sign of RSPY	Predicted Rate 0	Predicted Rate 1
0	0.673	0.327
1	0.053	0.947

1

Actual Sign of RSPY	Predicted Count 0	Predicted Count 1
0	36	12
1	7	67

Actual Sign of RSPY	Predicted Rate 0	Predicted Rate 1
0	0.750	0.250
1	0.095	0.905

2

Actual Sign of RSPY	Predicted Count 0	Predicted Count 1
0	43	24
1	1	55

Actual Sign of RSPY	Predicted Rate 0	Predicted Rate 1
0	0.642	0.358
1	0.018	0.982

Predictor

0

Actual Sign of RSPY	Predicted Count 0	Predicted Count 1
0	149	7
1	3	206

Actual Sign of RSPY	Predicted Rate 0	Predicted Rate 1
0	0.955	0.045
1	0.014	0.986

1

Actual Sign of RSPY	Predicted Count 0	Predicted Count 1
0	38	10
1	15	59

Actual Sign of RSPY	Predicted Rate 0	Predicted Rate 1
0	0.792	0.208
1	0.203	0.797

2

Actual Sign of RSPY	Predicted Count 0	Predicted Count 1
0	53	14
1	2	54

Actual Sign of RSPY	Predicted Rate 0	Predicted Rate 1
0	0.791	0.209
1	0.036	0.964

We can interpret the screenshot of this output by focusing on the right-hand side, where the 2s are, since these numbers represent the models performance on the unbiased, testing set. The first predictor model is the Decision Tree model. The second series of confusion matrices correspond to the Random Forest model. By looking at the Predicted Rates we can see that, despite the Random Forest having an overall lower Misclassification Rate than the Decision trees, it slightly misclassified more returns as being 1s (positive) when they were really 0s (negative) compared to the Decision Trees. That being said, the Random Forest greatly outperformed the Decision Tree in accurately predicting 0s, with at 79% versus 64%. From a practical viewpoint in the stock market, investors that went with the Decision Tree model over the Random Forest would be marginally better at accurately predicting stocks that actually realized positive returns on any given day, while at the same time they will have bought many more stocks that would only depreciate in value under the false expectations of growth, which would ultimately lead to much worse net losses in the value of their portfolio.

Column Contributions

Term	Number of Splits	G^2	Portion
RVIX	153	110.911567	0.3995
RHYG	114	38.1126024	0.1373
RVNQL	94	26.5959484	0.0958
RVWO	85	25.5303255	0.0920
REMB	34	5.6669018	0.0204

Unlike for the other Random Forest Model, JMP selected RVIX to be the most important variable in this model. Despite this difference, it is still reassuring to see that the top 5 variables are the same for both Random Forest models. All of these variables seem to make perfect sense as for why they are significant, as well as their slopes.

VIX has been referred to as the "Fear Index", being a measure of market volatility. With higher levels of perceived volatility, investors would have less confidence in the performance of their stocks and the overall health of the economy. So, the negative relationship between the RVIX variable in the model and our predictor variable should come as no surprise.

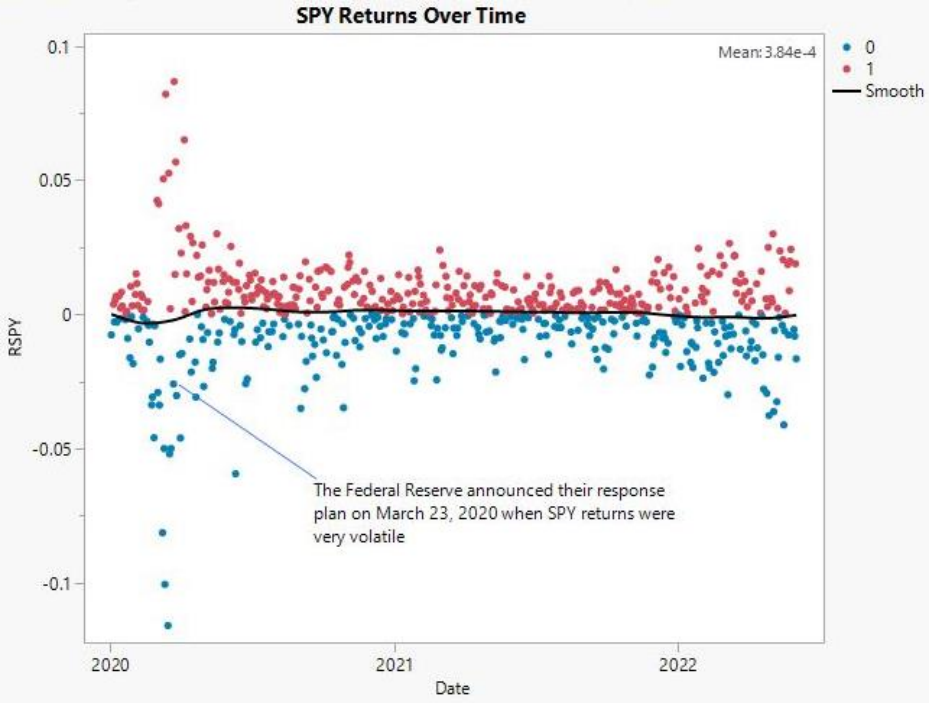
HYG is an ETF that tracks the performance of high yield corporate bonds, belonging to the colloquially called "Junk Bond" class. These are typically bought when there is an optimistic outlook on the stock market's performance since they are a riskier form of investing. Investors would be purchasing these bonds under bullish market conditions which would explain why the SPY returns would have a positive relationship from the model's prediction profilers.

VNQL is the Vanguard Real Estate Index Fund ETF, which directly relates to the US real estate market. As with other variables in this model, VNQLs performance is greatly influenced by interest rates, economic conditions, and property values.

Unlike the previously discussed variables, VWO (Vanguard's Emerging Markets Index Fund ETF) is focused on market activity outside of the US. Emerging markets are typically considered to have higher potential growth margins with the additional notion that there is also more risk associated with investments where volatility can be much more present than in developed markets. EMB is JP Morgans equivalent of an Emerging Markets Index Fund and shares many of the same characteristics and associated risks as VWO.

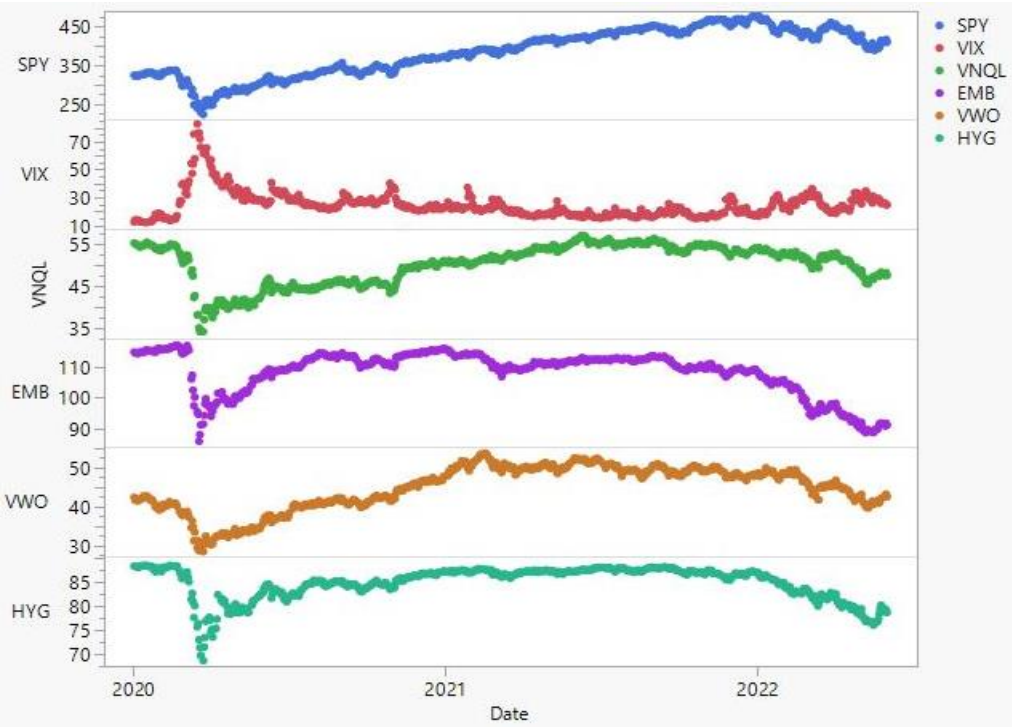
On March 23, 2020, the Federal Reserve released their response to the Covid-19 pandemic with a detailed plan to mitigate further economic turmoil that was unfolding in the US economy. Factors such as lockdowns, travel restrictions, and disruptions to trade in the movement of goods, had massive repercussions across several industries. All of these factors reasonably led to diminished consumer confidence in the market's performance which greatly impacted the values of the 5 variables that I previously mentioned and, in turn, the value of the SPY ETF.

The graph below shows how the spread (level of volatility) of the SPY returns changed throughout the dataset:



Following the Federal Reserve's intervention there is a very noticeable change in the amount of volatility to more stabilized market conditions. SPY, being an ETF that tracks the performance of the top 500 publicly traded companies in the US, shows that by mid-2020 the US economy was already considerably less volatile than in March.

For some additional context into how the predictor variables relate to SPY, I compiled a line graph with the actual values of the stocks, as opposed to the daily returns. This way, we can compare the shape of each graph and how they move together, or apart, across time.



All of these lines seemed to have moved in the same direction as SPY with the obvious exception of VIX, which appears to have a mirrored inverse relationship with SPY. So, this graph better contextualizes what I already explored using the Random Forest Model's Prediction Profilers in terms of the direction of the relationships between the predictor and response variables.