vZackery Bradley

BAN 525

20220213

**Model 4 Assignment 1**

On the topic of this week’s assignment, we are tasked with developing models utilizing the Random Forest method. Random Forest is based on an algorithm-based ensemble decision tree method, which was first developed by Breiman (1996, 2001). This algorithm is considered to be generously user-friendly, as it consists of only two parameters, consisting of the number of variables in the random subset at each node, and the number of trees in the forest. Since its introduction, Random Forest methods are usually highly desired in the statistic world and its creation is considered a modern innovation. Random Forest operates by selecting distinct relevant variables for prediction, without the need to rely on any functional or distributional assumptions Random Forest functions by forming a multitude of decision trees, all of which are independent of each other. Each node in the tree is then split using the best among a subset of predictors that have been randomly chosen at that node. By calculating each tree in random fashion, the Random Forest eliminates the criticisms of decision tree methodology in which individual trees have been known to be high correlated. This functionality gives Random Forest methods immense versatility and flexibility, as it is considered to work well with both categorical and continuous variables. This method is much different from previously explored methods such as penalized regression models, as Random Forest is able to capture nonlinearities in data. Specifically for our assignment, we will be estimating models by standard decision trees following by the Random Forest for comparison. Our data set is one in which we are familiar with, as the data is in relation to stock prices after Covid-19. Our response variables consist of two distinct factors, as we will be looking at both RSPY and the sign of RSPY. Ultimately, through this analysis we are attempting to forecast stock prices. Sign of RSPY is especially helpful with this, as it is a nominal variable that reflects a “1” if stock returns are positive and “0” if otherwise stated. Given the data, there are a few points of interests that we must identify prior to our analysis. Two variables in particular, RVIX and RHYG are of the upmost interest. RVIX is a measure of expected volatility, or a “fear index” in lament terms. RHYG are high yield corporate bonds, which are most often thought of as “junk bonds.” Throughout market history, junk bonds tend to increase in hot markets, as investors tend to not foresee bankruptcy problems in the distance during this period. Assuming the general flow of the market, a strong assumption could be made that as RHYG increases, RVIX would decline, as buying junk bonds are an indication of confidence in the market.

Analysis and Market Comparison

Earlier, we briefly mentioned the vast capability of Random Forest. In addition to our brief analysis, Random Forest has many additional advantages. One advantage is that Random Forest can automatically manage missing values in the data, as well as being uniquely robust to outliers, as the method oversees them autonomously. However, Random Forest methods do come with some disadvantages as well. One disadvantage worth nothing is that Random Forest usually is a tedious process, as computing the data often takes more time as compared to other methods. Procedurally, we will be employing a cross-validation analysis. Cross-validation techniques are built on the concept of leaving out a part of the data out of the estimation process as a buffer. As the predictions of the model stop improving as the data holdout process occurs, model growth then stops, and estimates are obtained. This creates a ‘split’ in the data. These splits are then broken down into three parts in the terms of our model. Those parts include training, validation, and testing splits respectfully. Training data is the portion that is used to estimate our model. Validation data is not used in the estimate directly, but instead works in the background to determine the optimal point in which the model stops. Lastly, the test data is never actually used in the model estimation. Test data is used to represent “new observations” and assists in providing and unbiased analysis of the predictive ability of the model. With the validation method only being utilized indirectly, we choose our model that displays the best performance on the test data. Upon inspection of our model comparisons, we can see that the Random Forest model proved to be much more efficient than the latter method. This can be observed in the visual below, as we see a much stronger measure of fit for Random Forest as opposed to our decision tree. Our decision tree was measured to have an RSquare of 0.2785, combined with a RASE and AAE of 0.0081 and 0.0063 respectfully. These number paled in comparison to our Random Forest method, as it was measured to have an outstanding RSquare of 0.6605 along with a RASE of 0.0055 and an AAE of 0.0043. These numbers dictate that Random Forest is a far superior method than the decision tree on the discussion of this particular data set.

**Measures of Fit for RSPY**

| **Holdback** | **Predictor** | **Creator** |  | **RSquare** | **RASE** | **AAE** | **Freq** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | RSPY Predictor BM DT | Partition |  | 0.6710 | 0.0136 | 0.0088 | 194 |
| 0 | RSPY Predictor RF | Bootstrap Forest |  | 0.9031 | 0.0074 | 0.0041 | 194 |
| 1 | RSPY Predictor BM DT | Partition |  | 0.3683 | 0.0076 | 0.0059 | 65 |
| 1 | RSPY Predictor RF | Bootstrap Forest |  | 0.6574 | 0.0056 | 0.0043 | 65 |
| 2 | RSPY Predictor BM DT | Partition |  | 0.2785 | 0.0081 | 0.0063 | 67 |
| 2 | RSPY Predictor RF | Bootstrap Forest |  | 0.6605 | 0.0055 | 0.0043 | 67 |

As we conducted our analysis, our market assumptions proved to be of great accuracy. We evaluated this in a visual comparison of the RHYG and RVIX variables.

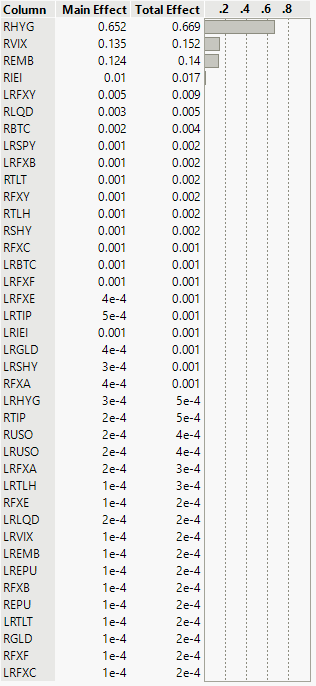
Chart, box and whisker chart

Description automatically generated

We performed an experiment by placing the RVIX variable at the lowest end of the visual spectrum, which resulted in RHYG reaching its maximum. From interpreting these results, we can see that there an extraordinarily strong inverse relationship exists between the two variables. From our study, we can validate our previous claim that as RVIX declines, RHYG should increase.

Interpretation

Upon a closer examination of our variable importance, we can see that in relation to RSPY, both RVIX and RHYG proved to be detrimentally significant. Our analysis measured RHYG to have a total effect of 0.669, while RVIX measured out to have a total effect of 0.152. Essentially speaking, these two variables proved to be strongly dominating figures in relation to RSPY. Additional measures of the latter variables can be seen in the visual below.



In addition to our previous analysis on the inverse relationship between RHYG and RVIX, we decided to assess additional measures on our third most significant variable, REMB. Interestingly, we discovered as REMB increases, both RVIX and RHYG increase as well. This can be seen as we decided to assess this by putting REMB and both the low and high spectrum in the prediction profiler. From our analysis, we can determine that the impact of REMB is not significant enough to have a detrimental effect on our models, as it is heavily overshadowed by the impacts of RHYG and RVIX. We evaluated this with additional examination conducted that resulted in RHYG displaying positive correlation in regard to REMB, while RVIX has a slight negative relationship with REMB.

REMB with a minimized prediction profiler

Chart

Description automatically generated

REMB with a maximized prediction profiler

Chart, histogram

Description automatically generated

In closing of this assignment, we can readily agree that Random Forest is a far superior method of analysis than that of the latter decision tree method. This analysis has proven to show just how vital different statistical methods effect results. In a world where accuracy is of the upmost importance, it is imperative that we understand and provide the best statistical models so that decisions can be made efficiently and appropriately.