**Random Forest and Sign of Gold Returns**

**Susan Wiggins**

**Cases in Business Analytics BAN 525**

**Professor: Dr. Cetin Ciner**



**Table of Content:**

Introduction……………………………………………………………………………………………………………………………………..…3

Analysis and Model Comparison………………………………………………………………………………………………………...4-5

Interpretation…………………………………………………………………………………………………………………………………….5-7

References…………………………………………………………………………………………………………………………………………8

Appendix A **(Ordinary Logistic Regression)**……………………………………………………………………………………….9-14

Appendix B **(Random Forest)**…………………………………………………………………………………………………………….15

Appendix C **(Model Comparison)**………………………………………………………………………………………………………16

Appendix D **(Column Contribution Importance )**……………………………………………………………………………..16-18

Appendix E **(Bootstrap Profilers)**………………………………………………………………………………………………………19-22

Appendix F **(Final Variable Importance)**…………………………………………………………………………………………..23-29

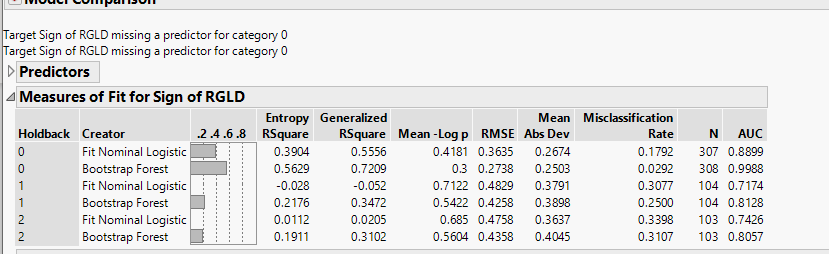
**Introduction**

Many, many years gold became an interest of the mass population and continues until today. This case study covers whether we can predict the direction of gold price changes. The dataset consists of variables from currency, bond, and stock markets. The data is weekly and consisted in the time series of March 18, 2009 through March 6, 2019. The dataset was prepared based on the following criterion: raw prices were converted into percentages, a difference of a natural log was established, continuous compound returns, returns are noted with R labeling, and L denotes lagged instances. Lagged attempts to focus on whether last prices change continue to impact gold price movements. The dependent variable being measured in this study is Sign of Gold Returns (RGLD) and the variable is classified as nominal meaning it takes a value of “1” if RGLD > 0 and “0” Otherwise. The predicator candidates range from RFXB to LRIYR and leave out the RGLD in the X-factor partition. Two methods will be addressed in this case nominal logistic regression and random forest. Random forest is the newest method and deals with nonlinear linkage. This algorithm-based decision tree method can derive both continuous and categorical data. Dr. Cetin Ciner points out that, “the random forest method has inherent advantages as it selects random variables for predication without relying on any functional od distribution assumptions. Furthermore, different from the fully linear penalized regression models discussed before, it can capture nonlinearities in data” (13). This machine language overall formulates many decision trees, which are independent of one another. The final step in the random forest involves an average calculating the predication needed. Decision trees are removed in the process narrowing down the exact idea of tree correlation.

**Analysis and Model Comparison**

The dataset employed in this case study is large and sometimes modeling can be affected when Big data is taken into account. Earlier methods discussed upon have been associated with linear characteristics and these methods have an advantage to easier interpretation. However, linear methods do not predict well when considering nonlinear linkages. On the other hand, random forest will tackle nonlinear linkages with a breeze and noted earlier can handle continuous and categorical data. In order to reach the best results in the analysis one must depend on cross validation which will decrease the ability to interpret random noise. In cross validation, data is held out and the predication criterion is based on this interpretation. Estimates are then built on the data after the completion of each modeling affect occurs. In the case predication of return of gold prices changes, the cross validation is established on 60/20/20 split of the data. Therefore, sixty percent of training data is used in estimation to row 309, twenty percent used for validation in order to stop the modeling process between row 310 and 413, and twenty percent used for testing for unbiased analysis in predication ability of the model from row 414 to end. The final interpretation of the model will be based on the results of the test data due to the basis of new observations.

Since the Sign of Gold Returns case dealt with time series data, the predictions are included at the end of the analysis making validation and test dead last in the column. JMP allows the user to set up the training, validation, and testing based on the labeling technique of 0, 1, 2. Zero is associated with training, ones are associated with validation, and twos are indicated by testing data. Next, modeling techniques of nominal logistic regression and random forest are initiated in the process to predict the direction of gold price changes. First, nominal logistic regression was performed on variables listed from RFXB to LRIYR while removing the RGLD dependent variable from the list. The results of the nominal logistic regression are listed in appendix A and the results of the random forest are listed within appendix B. After comparison of the modeling techniques was performed, I was able to deduce that the random forest was the best modeling technique in the prediction of the direction of gold prices. The results discovered in comparison modeling are in the table below.



Furthermore, Random Forest is the modeling technique that reveals the best unbiased results for testing. RSquare was the highest yield at 31.02% and Area Under the Curve (AUC) is reported at .8057 making it the greater for accuracy in the modeling technique. In addition, the Misclassification Rate signifies random forest being the best modeling based on the lowest value of .3107. Therefore, Random Forest explains 31.02% of variations in the direction of gold prices. In addition, the AUC scores clearly are much higher than Nominal Logistic Regression model as well as lower scores for Misclassification Rate are verified.

**Interpretation**

Since random forest is the chosen model a look at column contributors is analyzed in the interpretation of the data resulting in the Swiss franc (RFXF) as the largest variable to predict the direction of gold prices. RFXF explains only 9.8% and is followed by inflation proxy (RTIP) at 5.8%. Gold often thought of as real money is thought to be inflation proof. Therefore, if increase pressure in the economy then gold prices will increase. The third variable in the column contribution is related to the material sector (RXLB) for stock market at 3.9% and fourth is Canadian dollar (RFXC) for exchange rate at 3.3%. The Swiss franc is the greatest factor to explain the direction of gold prices. Also, Inflation rate helps gold prices move upward due to the decreasing value of the dollar. Next, the materials sector fluctuation of gold prices deals with the conditions in the economy and asset evaluations. The fourth column contributor has a low aspect of effecting direction of price of gold also which is the Canadian dollar. The Canadian dollar deals heavy with the exporting of commodities.

Next, the model prediction profiler reveals the direction of the variables in this case study. The first variable that was evaluated to have the greatest weight of importance in column contributors was Swiss franc (RFXF) and the profiler shows that an increase in RFXF leads to an increase in price of the sign of gold being positive. In addition, the second variable in the study for significance was inflation proxy (RTIP). Inflation proxy also increases when Sign of RGLD is greater than zero whereby leading to price increase in gold. The next profiler studied is related to material sector (RXLB) and has the same pattern to increase the direction of the price of gold. The Canadian dollar (RFXC) is the final variable researched for the answer to this case study question and the deduction is when increasing RFXC the price of gold also will rise.

In conclusion, the selected model to predict gold price direction was Random Forest. The highest variable for fluctuations in gold price is related to the Swiss franc of exchange rate group at 44 percent and has a positive relationship. The second most important variable associated with gold price is inflation proxy and has a positive relationship on gold prices. Inflation proxy total effect is rated at 19 percent. The largest variable associated with direction of gold price is the Swiss franc. These two variables mark 63 percent of prediction of the direction of gold prices. The column contribution revealed that among the four variable (RFXF, RTIP, RXLB, RFXC) a small contribution was made below 10 percent. Finally, through measuring the Swiss franc and the inflation proxy you can understand the direction of the gold prices in the market.

**Reference**

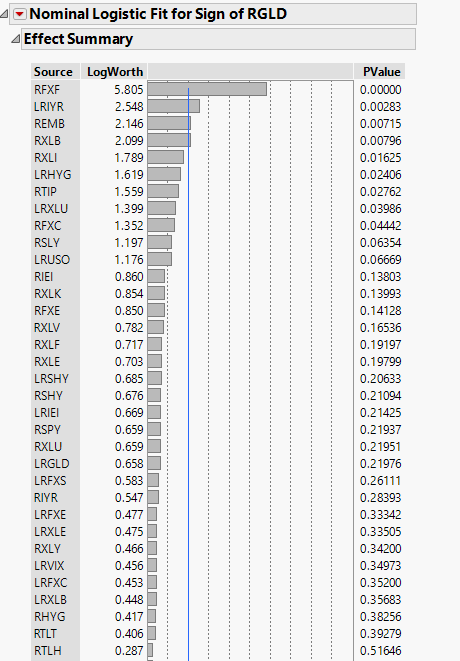
Ciner, Cetin. August 2019. Module 1. “Cross Validation and Predicting Gold Prices.” Slides 1-53. Retrieved from: <https://uncw.instructure.com/courses/24252/pages/module-1-lecture-materials-cross-validation-stepwise-regression-and-risk-factors-for-gold-prices?module_item_id=524495>

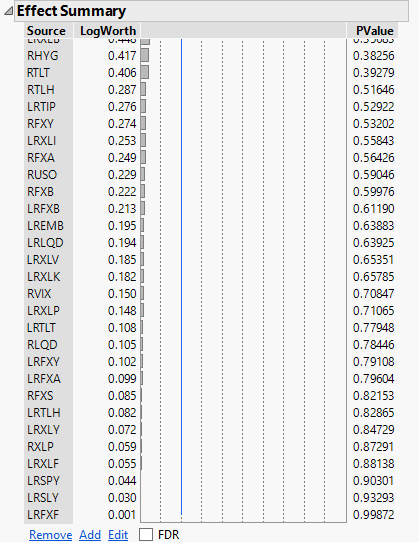
Ciner, Cetin. September 2019. Module 4. “Random Forest and Determinates of Stock Returns.” Slides 1-29. Retrieved from: <https://uncw.instructure.com/courses/24252/pages/module-4-lecture-materials-random-forest-and-determinants-of-stock-returns?module_item_id=524512>

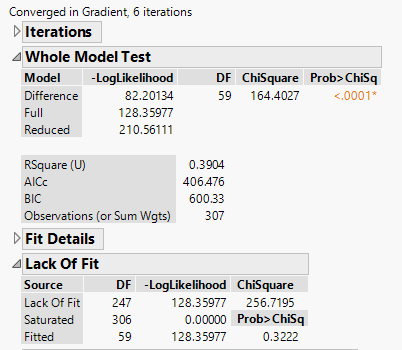
Web. Picture. <https://www.google.com/search?q=pictures+of+gold&tbm=isch&source=hp&sa=X&ved=2ahUKEwiJgMKY8d_kAhVhc98KHbL2CucQ7Al6BAgHEBs&biw=1366&bih=657#imgrc=ewATy5c2wTI_zM:>

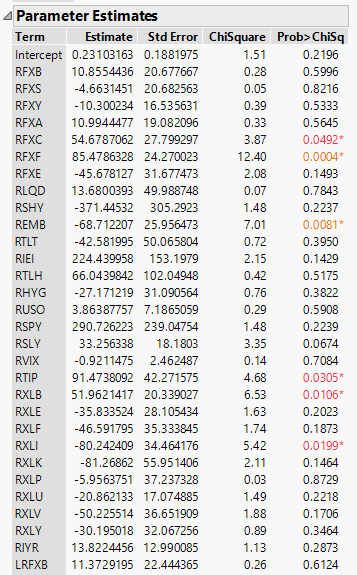
**Appendix A**

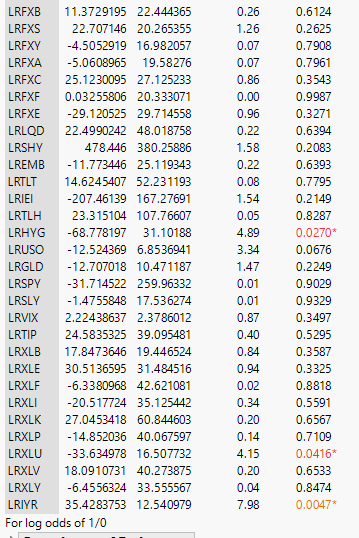
**Nominal Logistic Regression**

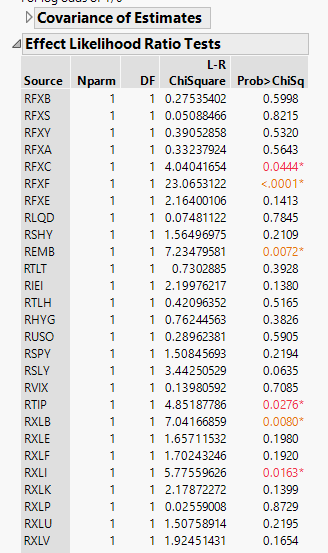


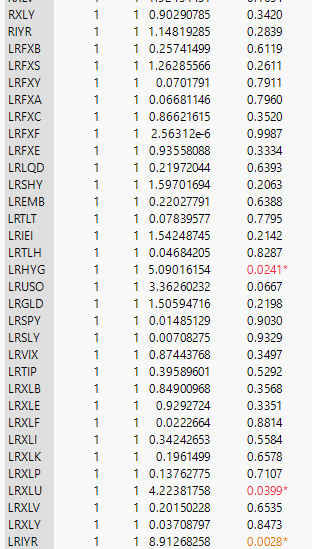






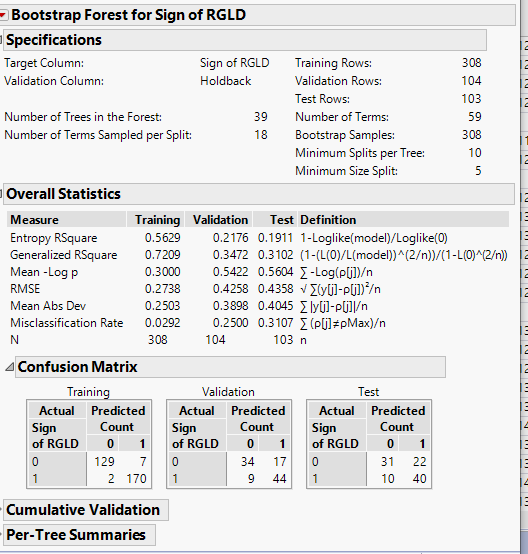






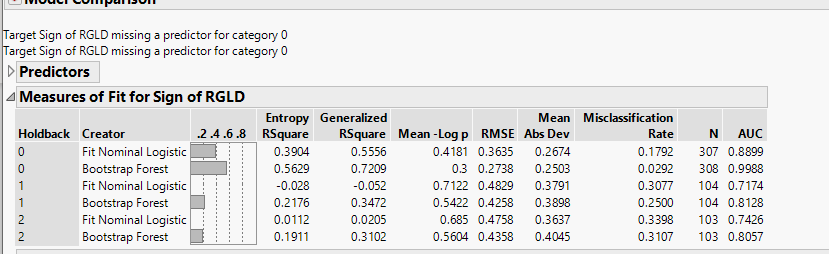
**Appendix B**

**Bootstrap Forest**



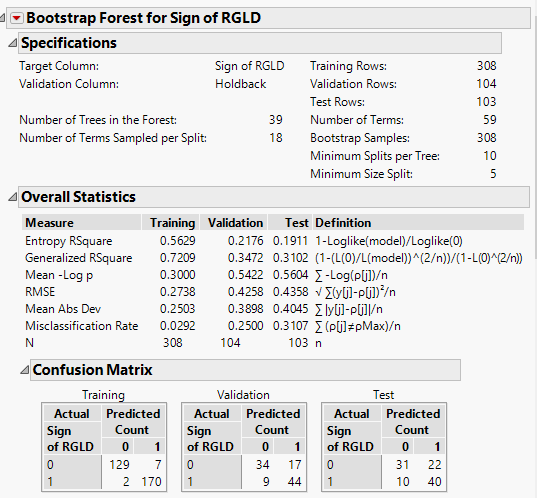
**Appendix C**

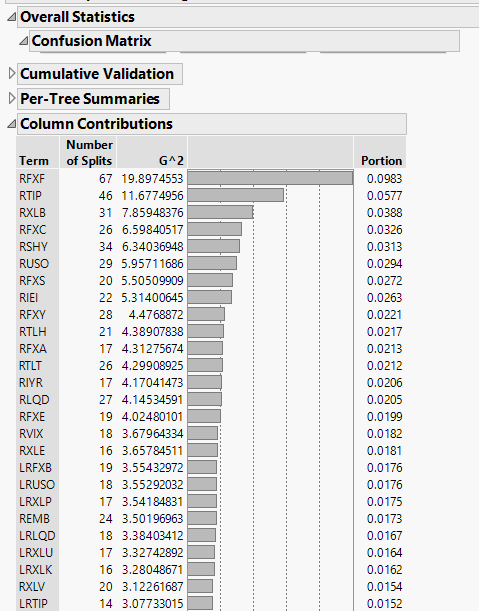
**Model Comparison**

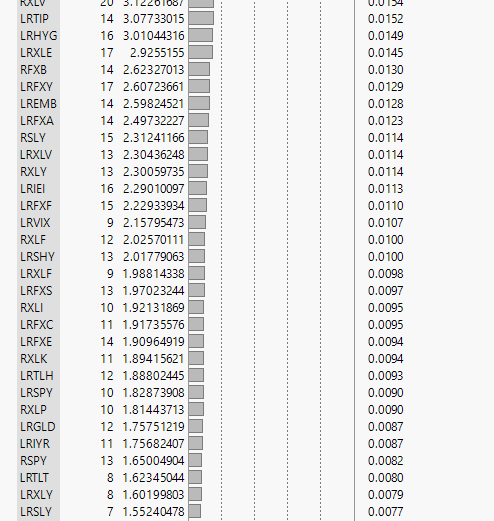


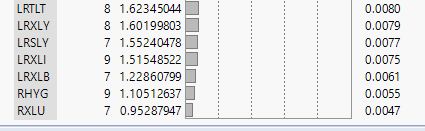
**Appendix D**

**Column Contribution Importance**



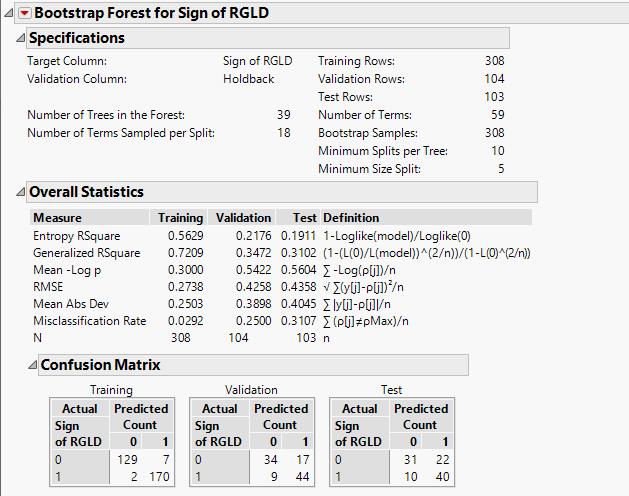


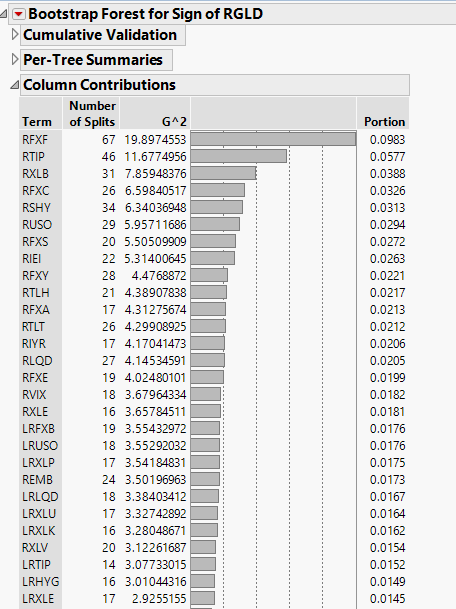


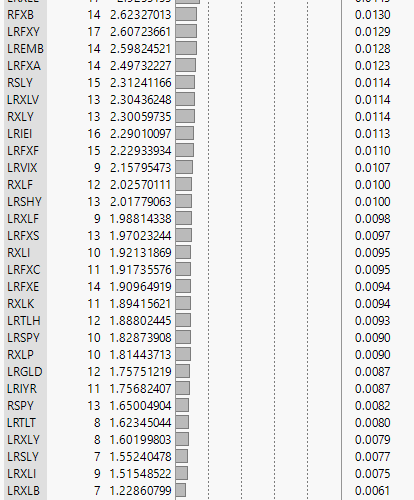


**Appendix E**

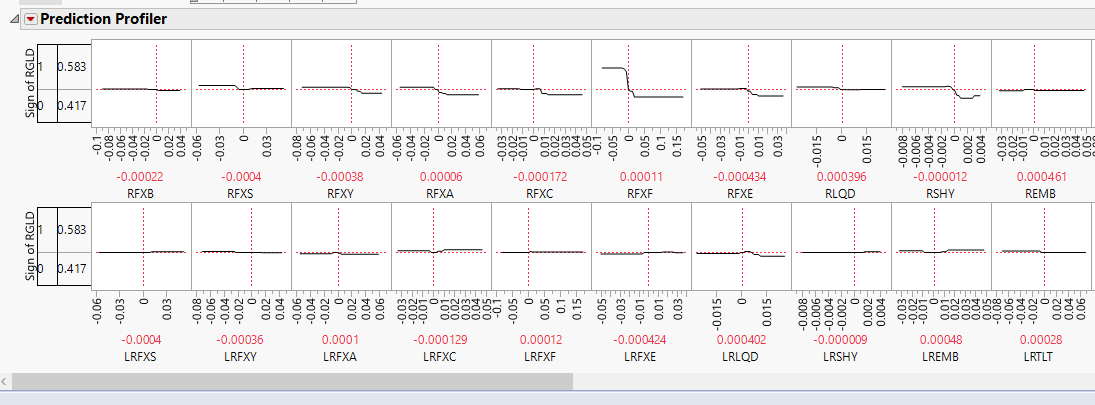
**BootStrap Profilers**

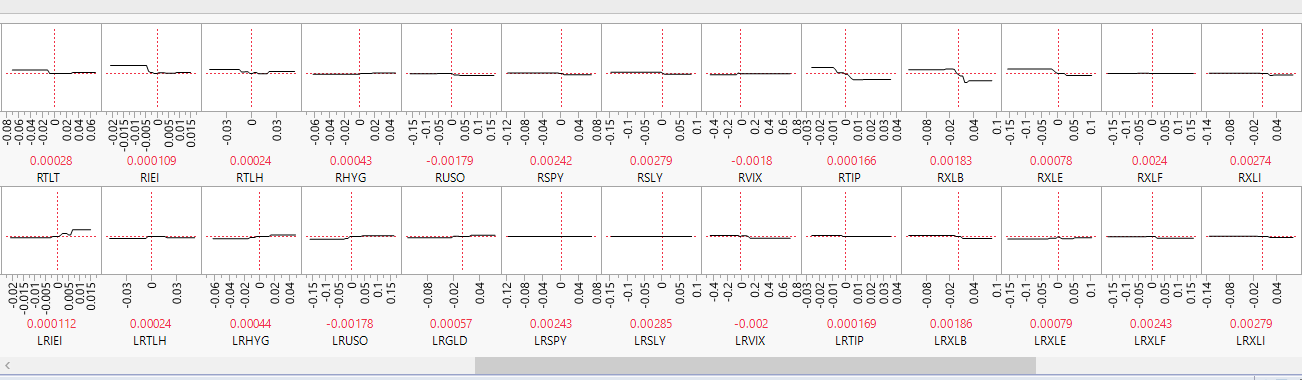


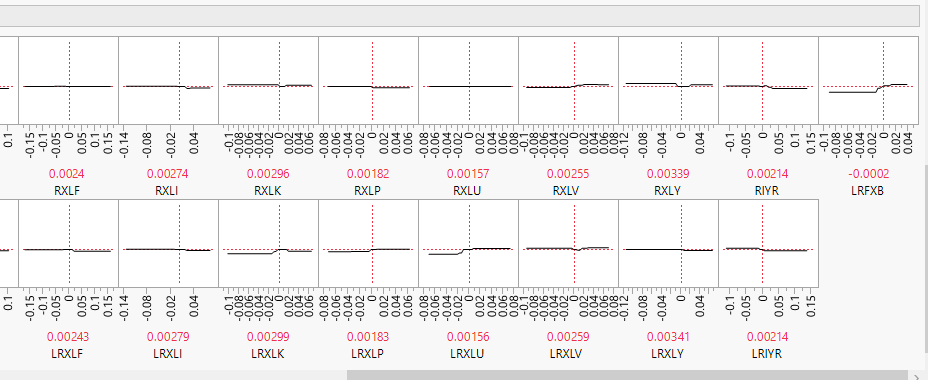












**Appendix F**

**Final Variable Importance**

