Gil Graybill – BAN 525
Assignment 4 – Random Forest and Sign of Gold Returns
Professor Cetin Ciner
June 7, 2020

Introduction

Previously we have been creating various models that will help predict our response variable. The *direction* that the response variable moves in with respect to the input variables has not been a concern to us. For instance, in Module 1, we determined that changes in the Canadian dollar, the Swiss Franc, and inflationary pressures were all positive influencers on the price of gold. Hence these variables had a *high correlation* with the price of gold.

Are these variables all in boats, riding the same waves as gold? Are some of these variables, and perhaps others, actually causing the waves? Is it possible to determine what direction gold will move in the future? If you could create a model that could look at the change in commodity prices in one week to determine the direction of gold (or any other commodity) the following week, that could be an amazing feat. Money is made on the margins all the time. If a company could look at the data to determine if they should buy/invest this week or next, that could save/make money for them on a constant basis. It's also obvious that this use of "Directional Forecasting" could be used by traders to make money on seemingly small changes in the prices of commodities. Can we look into the future and forecast the direction of the price of gold?

The dataset we are working with consists of 10 years of weekly prices for 6 different currency exchange rates, 7 different interest rates for bonds of various terms, 12 stock market indices, and 4 miscellaneous measures (See Appendix A). Each of these variables has been normalized with a percentage change from week to week. In addition, we have created a "lag" variable so we can compare last week's changes for a given variable against this week's change in silver prices. Most importantly we have calculated a variable "Sign of Gold", that is "0" if the price decreased last week and "1" if it increased. In all, there are 60 different variables we will be testing to see if they have an influence on the "Sign of Gold".

We will be using two types of analysis on this dataset:

(1) Nominal Logistic Regression: This type of logistic regression model calculates the best fit through the data points that minimizes the squared differences between the observed values and the fitted values, and uses the calculated value to further determine the value of the nominal (or categorical) response variable.

(2) Random Forest: With this type of modeling, many decision trees are created randomly, and each tree makes a prediction about the observation. The “wisdom of the crowd” takes over, and the result that gets the most “votes” is the one that is ultimately chosen. The idea is that many uncorrelated models will outperform any one individual model.

Analysis and Model Comparison

If the relationship between the response variable and the potential dependent variables is linear, Nominal Logistic Regression will help find those relationships. One disadvantage is that if there are many variables that are highly correlated, the model may create random errors. It tries so hard to include everything that the model will become inaccurate and be prone to overfitting. We will be using this as our “simple” benchmark analysis.

The Random Forest creates many decision trees and uses an “ensemble” approach to come up with its prediction for each observation. It has been proven to be a very good modeling approach. A disadvantage is that with the Random Forest there is no resulting equation or visual representation to look at – how that final result is determined is a black box to the user. It is computationally heavy, so for large datasets or small computers this could be a drawback.

For cross-validation we employed a 60-20-20 split for training-validation-test split. This gave us 308 rows of training data, 104 rows of validation data, and 103 rows of test data. Because this is time series data, we did not split up the data randomly, but instead used the first 60% of the data (time-wise) for training, the next 20% for validation, and the final 20% for testing. This makes sense since we are trying to build a model that could potentially be used to predict future direction of gold prices. With time series data there is always the risk that an event or phenomena will alter the data from a certain date forward.

"Sign of Gold" was chosen as the response variable and the rest of the variables were tested as dependent variables using JMP Pro 15. The calculated value for the response variable, the probability that gold would go up, was determined and recorded and the resulting formulas determined using both methods.

The best way to compare results for a logistical regression model is the check the "Area Under Curve", or AUC. The value consists of the ratio of "True Positives" vs. the ratio of "True Negatives", plotted against all possible cutoff values for success. The higher the AUC, the better the model is. Because we are using cross-validation, we'll be looking at the results for the Test data. The results are as follows:

Cross-Validation	Model	AUC
Training	Nominal Log. Regr.	0.8899
Training	Random Forest	0.9988
Validation	Nominal Log. Regr.	0.7174
Validation	Random Forest	0.8128
Test	Nominal Log. Regr.	0.7426
Test	Random Forest	***0.8057***

Random Forest outperformed Nominal Logistic Regression for all validation levels. Against the Test data, the AUC came back with a value of 0.8057, which is considered excellent (An AUC of 0.5 would be no better than a coin flip).

Interpretation

With Random Forest there is no formula or visual representation of the resulting model.

With this exercise we are not only calculating what factors determine *the direction* that gold is going to move in, but if those factors are different than those which *determine the price* of gold, which was done as a class exercise in Module 1.

When running variable importance against these variables, we get the following:

	Sign of Gold (Mod 4)		Gold Price Predictor (Mod 1)	
Variable	Main Effect	Total Effect	Main Effect	Total Effect
RFXF	0.374	0.448	0.583	0.608
RTIP	0.125	0.193	0.18	0.205
RIEI	0.02	0.056	NA	NA
RFXC	0.014	0.047	0.165	0.19

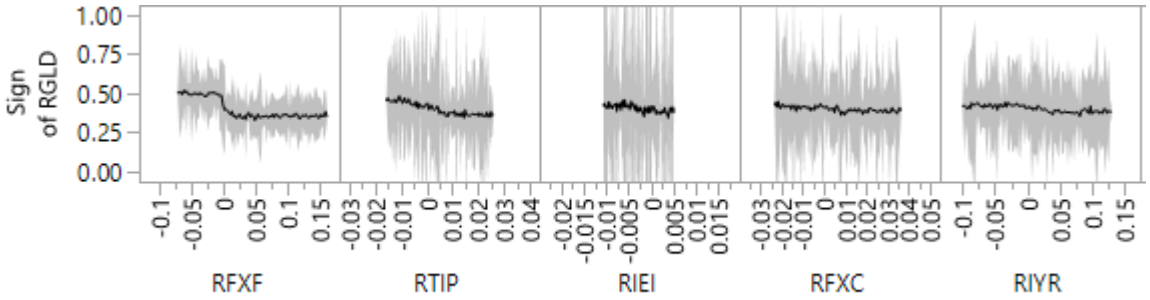
RFXF is the Swiss Franc, which accounts for 45% of the effect on the model. When investors want stability, they look to Swiss Francs. For being a predictor, RFXF has a 61% effect on the model. Even though RFXF is the main predictor for the direction of gold, it does not carry as much weight as it did as a predictor of price.

RTIP represents inflation, which accounts for 19% of the effect of the model. As a predictor of price, RTIP accounted for 20% of the effect, so RTIP appears the affect the value and direction similarly.

RIEI represents 3 to 7 year treasury bonds. These are considered a fairly stable investment overall, but with a term shorter than the more popular long term treasury bonds. They represent 5.6% of the total effect of the model for the "Sign of Gold", whereas they were considered insignificant as a gold price predictor. While this is not a large value, it is still noteworthy that it is only significant to determine the direction rather than the price. Perhaps investors are seeking short term stability when they buy the bonds (similar to gold), or they are the first thing they sell when they feel the stock market is ready for an upturn.

RFXC is Canadian dollars and has a 4.7% total effect on the model. This is not a lot, but considering that as a predictor of the price of gold, RFXC has a 19% total effect on that model. That could indicate that while the Canadian dollar does reflect developments in commodities, it is too correlated with the price of gold to be of any predictive value for the direction that gold is heading in.

Running the profiler and looking at the graphs show how “Sign of Gold” is influenced by the most significant variables. Small changes in RFXF seem to have as much influence as large changes. RTIP also appears to have a significant influence.



Appendix A: Normalized Variables used in analysis

FXB: British pound
FXS: Swedish krona
FXY: Japanese yen
FXA: Australian dollar
FXC: Canadian dollar
FXF: Swiss franc
LQD- Investment grade corporate bonds
SHY- 1 to 3 year US treasury bonds
EMB- International (Emerging market) bonds
TLT- Long term (20+) US treasury bonds
IEI- 3 to 7 year US treasury bonds
TLH- 10 to 20 year US treasury bonds
HYG- High yield (risky) US corporate bonds
SPY: S&P 500 (large company) index
SLY: S&P 600 (small company) index
XLB: Material sector index
XLE: Energy sector index
XLF: Financial services sector index
XLI: Industrials sector index
XLK: Technology sector index
XLP: Consumer staples sector index
XLU: Utilities sector index
XLV: Healthcare sector index
XLY: Consumer discretionary sector index
IYR: Real estate sector index
USO: oil prices
VIX: stock market volatility
TIP: inflation measure
GLD: Gold prices