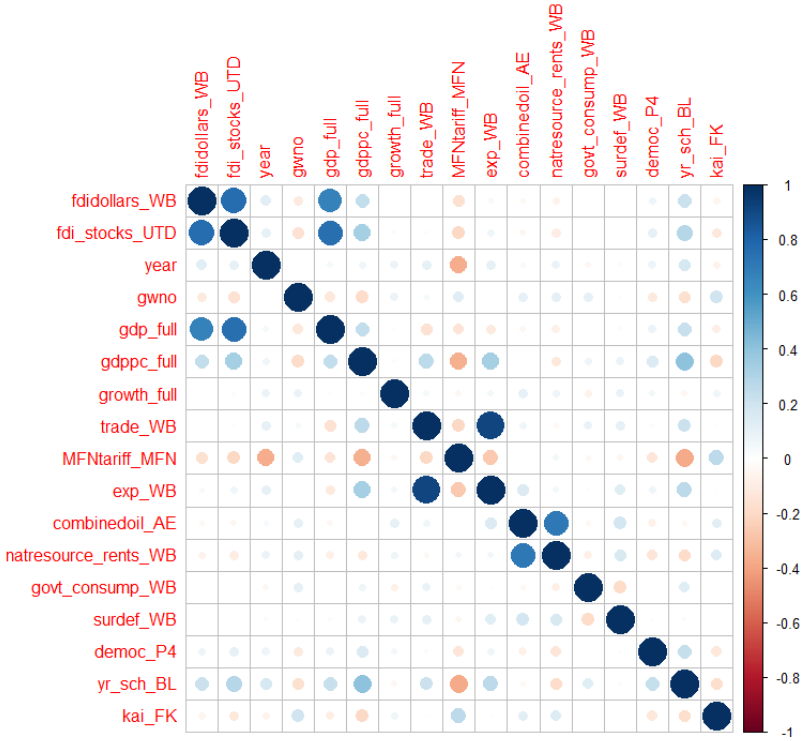
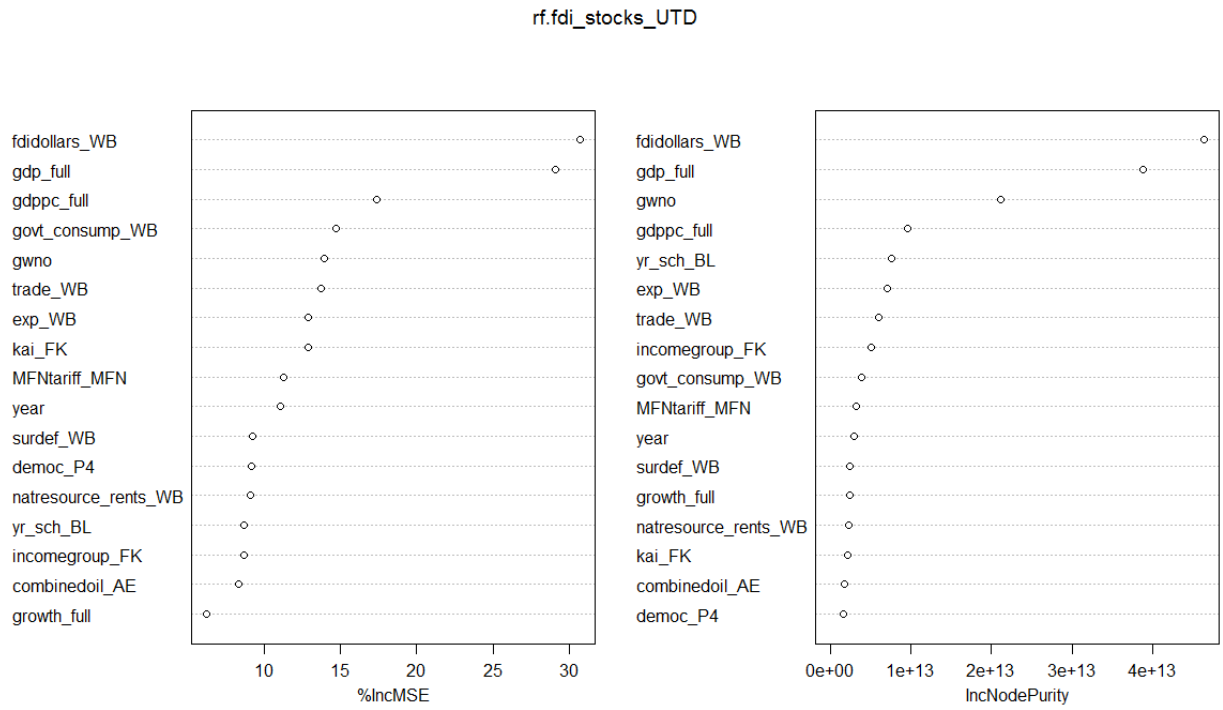
Use of Machine Learning to Find Variables of Importance

This document will explore the initial findings of the model. At this point, it is mostly a verification of the current program. We are looking to see if the program differs the intuitive relationship the variables have with the response variable. We will also note where the model might perform better than a regression, or where the output suggests a relationship that is easier viewed from this output than the usual process of model specification.

Since this program measures two highly correlated variables, unless a partial dependent plot significantly varies from one random forest to the next, I will not include it.

The first plot I will put up is the correlation matrix, a new addition to the program. There are qualifications that need to be made on the variable importance output, so the correlation matrix will be right before the variable importance plots for easy reference. Strong correlations, especially with similar groupings of variables, can overestimate the most important variable in that grouping while underestimating the subsequent variables. There is no hard cutoff of what correlation is unacceptable, but this should offer an easy way to further explore variable importance.

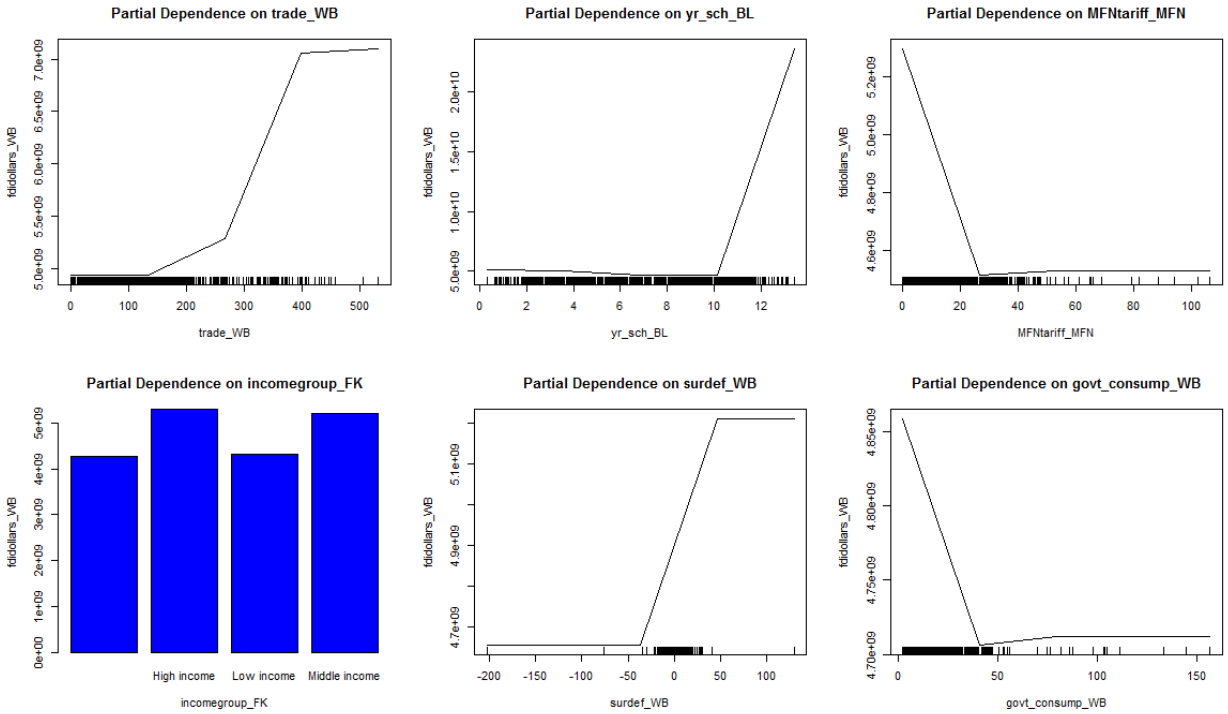




Note that the first three variables are of the highest importance, and they are also highly correlated in the previous plot. It clearly makes intuitive sense that they would be highly correlated with the response variable. Note also that there is a large gap between gdp_full and gdppc_full. As I spoke about before, the correlation between the two GDP variables likely explains the large gap.

The next most important variable in the MSE plot, gov_consump_WB, has almost no correlation with any other variables, so it must be something we examine carefully. Node purity and increasing MSE can and do differ (see documentation). Combinedoil is low in both plots; perhaps, since we have data on most countries in the world, the importance in global FDI is reduced considerably. Further, the other natural resource variables are ranked much higher, so we may have a case of reduced importance in these measures. Growth_full, a measure of GDP growth, is not highly correlated with any other variables, yet it is surprising low in importance in this model. We had discussed that the growth can be a longer term effect that possibly could be lagged for a higher variable importance. Note also that yr_sch_BLis lower than I would expect it should be. As we will see later, it has a diminishing returns effect, so it is possible that there is a section where the variable importance is high, but for the overall range, it does not affect the overall importance.

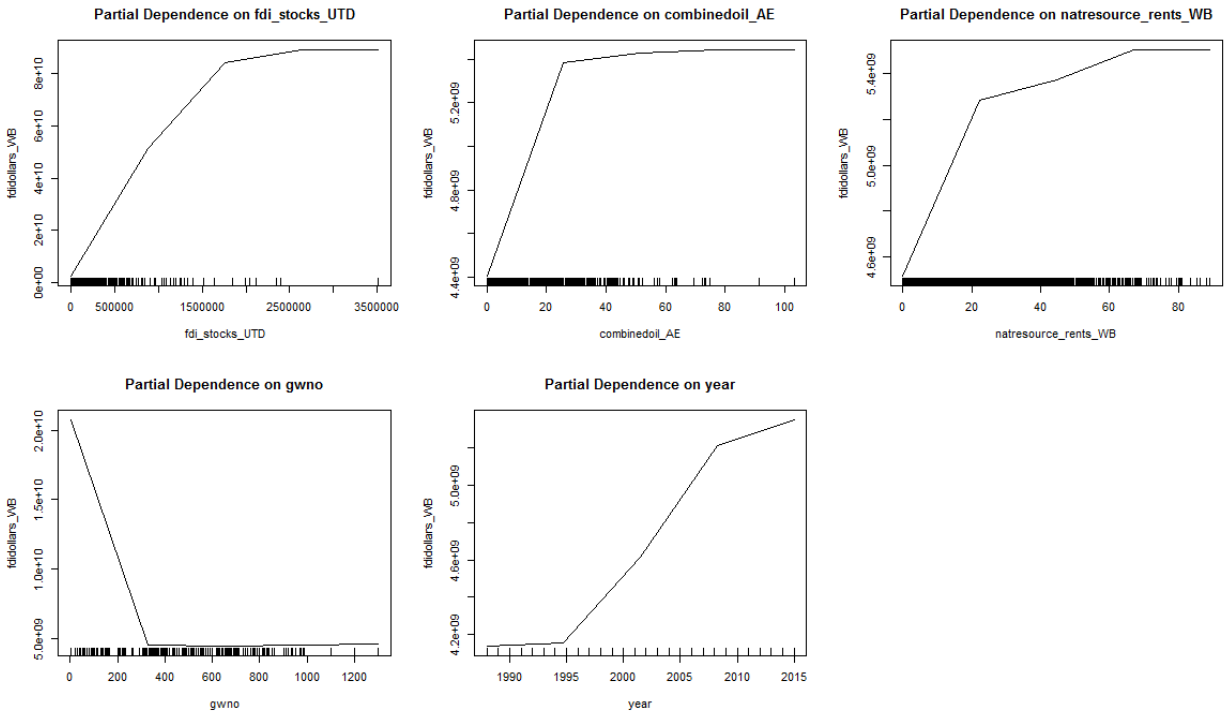
We will next look at some partial dependence plots and discuss the implications of each.



Remember that these partial plots are not affected by high correlations, so these plots will correctly estimate the relationship, even if they aren't important. I'm not going to recommend that any variables be dropped, but I will recommend transformations of variables.

The interpretations of these partial dependence plots are not always intuitive and should not be considered for explanatory power, but they can be useful. For example, the first plot shows what appears to be an exponential relationship of some sort, suggesting a transformation would make for a better fit. In the other four charts, we see a very linear relationship, but outliers are certainly present. Some sort of elimination of outliers would be the best course of action here.

I do think the year in school variable is interesting. It suggests that fewer years in school have a negative effect on FDI. It might be worth examining if countries with improving education experience a small dip in FDI until it pays off in future investment. After 10 years in school, FDI increase rapidly in a linear fashion.



This plot suggests a transformation of variables and elimination of outliers is necessary, but the partial dependence on year is intuitively worthless. While it may be possible that FDI growth increased in certain time periods, the model is relating the number of the year to the FDI (since we had to convert it from factors to integers for the random forest) and the way R computes this relationship will not reveal any true importance. Using lagged variables to check for importance is fine; looking at time as a variable is not the same as a proper time series and should absolutely not be used. However, leaving it in the dataset does not bias any other variables, so its presence does not concern us.

In summary, our correlation plot gives us information on how to properly interpret the variable importance plot. In this case, correlations were low enough that our variable importance is unlikely to be biased enough that we might exclude a variable that is actually important. Our partial plots gave us good information on possible transformations and eliminating outliers. While there are some interesting results, there are no significant deviations from intuitive understanding of these relationships. As expected, these plots hold little explanatory power, but this information is clearly useful in the first stages of building a model that can be interpreted.