**API 222: Prediction Competition**

Arjun Bisen

Harvard Kennedy School

**a) The Process**

As per usual, I began by setting up my RStudio module, including by loading all my usual libraries and the training and test data sets. I then explored the data using the usual *summary*, *dim* and *str*, functions and noticed some non-numeric and missing data points using *!complete.cases*. I then broke these down by column name to get a more granular picture of the offending predictors.

I then deleted the three columns with the with majority missing data sets (*v2a1, v18q1, rez\_esc)*.

I also removed the two ID columns as they were not going to be useful for building a model.

I found *meaneduc* and *SQBmeaned* had four missing values each. I decided to fill these missing values instead of deleting the entire columns because the test data didn’t have any missing observations in these categories. Deleting the entire columns could have severely reduced my model’s predictive power.

To minimize the impact of those missing observations, I decided to replace them with either the mean or median. To help me select which would be best, I checked the value of both and plotted both. I ended up replacing the missing values with the median amount to reduce noise from outliers and did a final check to make sure there weren’t any missing categories.

I decided to build a Random Forest model because I was working with a categorical output and wanted to produce a high level of prediction, and if I could get to it, a sense of categorical importance. I had time on my hands so could select something computationally expensive and non-parametric. I also didn’t want to assume an underlying function. I knew Random Forest would have a better prediction capacity than bagging so took a chance and selected it over of Boosting.

To proceeded, to split the training data into 80% for training and 20% for testing. I then trained the forest model and measured its performing using the F1 score. My model was performing very poorly, at around 0.002 F score. To diagnose the problem, I looked at the predicted output and noticed it was predicting continuous variables.

For my own clarity, I separated my predictors and outcome variable so I could easily convert the outcome variable into a categorical variable and my model performed much better, at around F score: 0.85.

To try and outperform this score, I tried to build a Neural Network using a few R packages: Neuralnet, and later- nnet, and caret, which I found through online forums. This took a great deal of computational power and time to try different snippets of code (even more than the Random Forest). In the end the neural networks were giving my a slightly lower predictive power than the Random Forest model so I continued using the Random Forest.

I had major difficulty applying the model to the test data provided and received an error saying the predictors were different to my training data.

I once again had to examine the datasets and eventually found that the *dependency, edjefa*, and *edjefe* variables had a mix of values, some of which didn’t make sense e.g. ‘yes’, ‘no’ and ‘8’ dependencies. I removed these columns and the prediction worked.

**b) My model**

I used a random forest model with p-5 predictors. As mentioned above, I deleted *v2a1, v18q1, rez\_esc, dependency, edjefa, & edjefe* columns as these were riddled with errors. I filled missing data points in *meaneduc* and *SQBmeaned* with the median of those predictors to minimize the impact of those missing piece of information on the model. I also classified my output variable into a categorical variable to improve accuracy.

**c) Bias and transparency**

One of the downsides of the two models I tried – Random Forest and Neural Networks, is that they have low interpretability and transparency. The model had no functional assumptions so should have resulted in low bias. I couldn’t tell the full impact of demographics on actual rates of poverty but in terms of predictive power, removing men from the model had a very slightly higher negative impact on the model’s performance than removing women, which might suggest that the model is better at predictions for men. Similarly, it had a stronger predictive power for adults over children, which is concerning for predicting childhood poverty.