My first Data Science competition

My “Microsoft Professional Program Data Science” certification cumulated with the Capstone project in the form of a competition to predict the Prevalence of Undernourishment, aka PoU.

I am happy to have finished in [5th position](https://datasciencecapstone.org/competitions/9/predicting-chronic-hunger/leaderboard/) out of 441 participants and this post is a summary of what I learnt during the competition. (I have included a link to the GitHub repo at the end of the article.)

The competition format was standard. The mission was to predict the PoU of a dataset without knowing the actual PoU. You are provided with training data (features and the PoU value) in order to train a prediction model. Your score was determined by the Root Mean Squared Error (RMSE) of your predicted values vs the actual values. Lowest score wins.

Continue reading on my personal blog…

If you are after the details and my code, you can find it at the GitHub repo. My insights from this competition are:

* It is easy to see why Python, Pandas, SciKit Learn, and Jupyter are so popular for Data Science. It is a fully functional ecosystem with a simple yet very powerful feature set.
* Getting to a prediction model is easy, getting to a good prediction model is difficult. It took me two weeks of toil to get from the middle of the pack to the top tier.
* First breakthrough came after deciding to replace missing data with a linear regression estimate by country, rather than simply deleting or substituting with mean or median.
  + The adage that 80% of Data Science is in the feature preparation was completely true for me.
  + This relatively simple change improved my RMSE to below 10 which was still way behind the leaders though.
  + From a conversation with a fellow competitor post completion it turns out this is isn’t obvious that whatever changes you make here also needs to be made to the test data.
* The second breakthrough came when I realised that stratifying by country when selecting training data was wrong. Removing this step brought my score to 7.7688

The GitHub repo contains the full details.

The data was provided by the Food and Agriculture Organization of the United Nations (FAO) and consists of various socioeconomic statistics by country.

The FAO defines PoU as:

*An estimation of the probability that a randomly chosen individual in the population regularly consumes less food than his dietary energy required to live a healthy active life.*

PoU is a statistical model and represents an estimation of how likely individuals are to suffer from chronic hunger. It is a complex metric and as such difficult to calculate, hence the desire to estimate PoU from more readily available statistics.