Topic: Microsoft Malware Prediction
Name: Mohamed Al-Rasbi

I joined a Kaggle competition hosted by Microsoft. Given big training and testing datasets (about 4GB each), the goal is to predict a Windows machine's probability of getting infected by malware. I am really interested in working on this problem because I was always interested to apply what I know about data science and machine learning on security-related issues.

The dataset has about 83 columns. Each column describes a unique machine. What I have done so far is analyzing each feature and its relationship to the response (which is the column named HasDetections in the dataset). I am using Pandas to read and calculate: a. the total number of unique values for each feature, b. the number of unique values for each feature that have the value HasDetections equals 1. By doing that I can see the percentages of all unique values being hit by malware. This will help me set weights and start implementing, or actually applying, machine learning algorithms that are already implemented by scikit-learn or other packages.

The submissions for this competition are due in one month from now. I will do my best to work on this problem and come up with a good prediction score. I am not sure what I am going to work on after the competition finishes, however, I might work on another project if that is possible.