Week 1 Summary

At the beginning of the week I was planning on using a neural network as the machine learning approach for my Capstone project. After my initial Capstone review meeting, I was asked to reconsider neural networks and look at other models. I read that neural networks, while accurate, were difficult to train and fine tune. In my research, I came across three different classification models that could work for my Capstone: Support Vector Machine (SVM), Naïve Bayes Classifier, Decision Trees (Random Forest). Each of these models has advantages and disadvantages.

Naïve Bayes is the most simple to implement as it is based entirely on probabilities. Naïve Bayes also requires less training. However, Naïve Bayes assumes that all features in objects to be classified are independent which can lead to less accurate classifications.

Random Forest is a collection of decision trees. Each tree is constructed by random subsets of training data. The final output is the mode of the classifications output by each tree. Random Forest is more accurate than Naïve Bayes but it is also more complicated.

SVM attempts to separate objects in two distinct categories with a line or curve. The goal is to come up with a function that divides the objects in different categories with the least error. SVM is highly accurate but it is also memory intensive and hard to interpret. I watched a lecture on SVM and while I could somewhat follow the theory, the math was hard to understand.

I am still undecided between Naïve Bayes and Random Forest at this point.