**Data Mining Project**

**Step 1,**

GitHubRepo: <https://github.com/nblumenfeld/Data-Mining-Project>

Dataset: <https://github.com/bgweber/StarCraftMining/blob/master/data/scmPvT_Protoss_Mid.csv>

Files turned in:

* pandas.ipynb
* sklearn.ipynb
* StarCraft.pdf - pdf of decisionTree
* starCraftProject.py - code
* team\_file.md - markdown of what each member did

The dataset we are using is a set of data collected from playing the game StarCraft: Brood War. The game is a real-time strategy game, where you play as one of three races, humans known as Terran, the high-tech aliens Protoss, or the parasitic race Zerg. The game is centered around building a base, gathering minerals, producing troops and destroying your opponent.

The data is describing choices done by a Protoss player playing against a Terran over many, many games ( the exact number is a little hard to figure since this is part of a bigger dataset published by Google DeepMind). The dataset does not describe how to play the game, where to move troops, how to move the units and similar actions. However, it describes the build order and overall strategy of the Protoss player.

The attributes used in the dataset describes an important action, such as putting down a building, starting an upgrade or producing a unit. As an example, ProtossPylon describes starting to build a pylon, while ProtossGroundArmor1 describes starting the ground armor upgrade.

The classes used are called midBuild and are FastObs, FastLegs, ReaverDrop, Carrier, FastDT, FastExpand, Unknown.  
Each describes a specific way or strategy to take for the midgame.

For step 1 we decided to build a Decision Tree using sklearn. Since our overall goal of this project is to learn what the minimal number of attributes needed and still get a satisfying accuracy we decided to use only 9 of the attributes. The attributes we choose were

ProtossPylon, ProtossSecondPylon, ProtossFirstGas, ProtossSecondGas, ProtossFirstExpansion, ProtossGateway, ProtossGroundWeapons1, ProtossShields1, ProtossCitadel

We choose these 9 based on primarily Thomas’ knowledge of the game, he has played StarCraft for about a decade and specialized in playing Protoss, and to see if we could use the data with sklearn. The idea being that if we could get this to work, we have proven the concept and can move on to use the full 52 attribute sett.

First, we used standard deviations, in this case, 1, to trim out outliers from the data we used. With a standard deviation of 1 the data used for the tree represents about 96% (980 rows) of the data, from our 9 attributes. Then we created a test sample (80% of the data) that we used to fit the decision tree, before running the remaining data (20%) through. This gave us an accuracy of 75%.

Since we used the basic version of the decision tree and changed as few of the parameters as we could for this test the decision tree is built based on Gini value. This gave us the root as ProtossFirstExpansion <= to 7105, with Gini index of 0.7521. The way decision tree is set up it also gives us a class based on what is the most likely class if we just looked at this point, which for the root was FastLegs.

The decision tree pdf (StarCraft.pdg) shows the tree, with color-coded boxes based on the class most likely at that point or the class for the leaves. The color-coding isn’t always 100% accurate.

green shades = ReaverDrops,   
white or blue = FastLegs  
yellow = FastDT  
orange = FastObs  
purple = Carrier  
turquoise = FastExpand  
pink shades = Unknown

StarCraft is an incredibly complicated game, which is why it is used by Googles DeepMind to do Artificial Intelligence studies on. So, to get a 75% accuracy from only using 9 attributes was surprising. It also proves that this is a data set that has potential, and it will be interesting to see what we can learn from using a bigger portion of the dataset.