William Hammond
Harry Longwell

Summary:

Due to difficulties acquiring data, we needed to change our project a bit. The problem we are currently trying to solve is to see whether or not it is possible to use measures of market health in order to predict a median market value for an area. Currently, we have trained and tested our model using data sets on the state and county level. We've already encountered and dealt with a number of issues that were mostly caused by our general inexperience with neural networks.

The biggest issue we've faced made it look like our model was much better than it was. On our initial run we had forgotten to normalize our data before putting it into the model. This caused massive over fitting. The weight decay of the model was around 10^6 and At first I wasn't sure about what this value was so when I checked the accuracy and the worse point was only 10% away from the expected I was ready to call our mode a success. After taking the time to look at some of the metrics that R's neural net library supplies it was obvious our model was awful. Every single one of the expected values was the same. We normalized each attribute by the z-score and reran the model. This time the value of the objective function was more reasonable and the model actually trained for multiple iterations. This is where we found our error to be incredible large. This is where we're at more or less in terms of the model.

As far as our feature selection goes, we haven't done too much other than the basics. We have been using PCA analysis to try and determine what features are the most useful, and have excluded some including features involving foreclosures. After removing these features our error was marginally improved. Because our error is so large and because of the general importance of feature selection, we plan on looking more closely into feature analysis as opposed to model tuning in order to further improve our error.

Plan:

We are going to exhaust all possible features available using different feature selection methods taught in class in order to find the best set. We plan on using

1) k-Fold Cross Validation
2) Hold-One-Out
3) PCA
4) And if all else fails, exhaustive

There is also a chance that our data has such a wide range of success and failure because of our limited our training set is. In order to make up for this we are going to look into interpolating points between the ones we have in order to have a bigger training set.