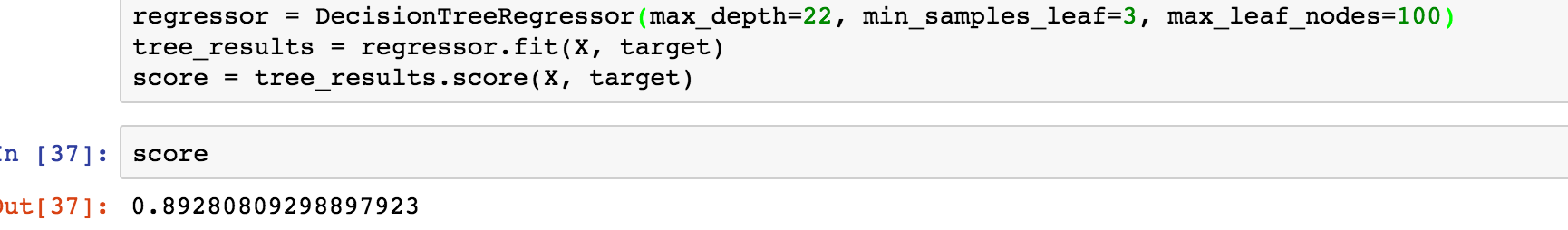
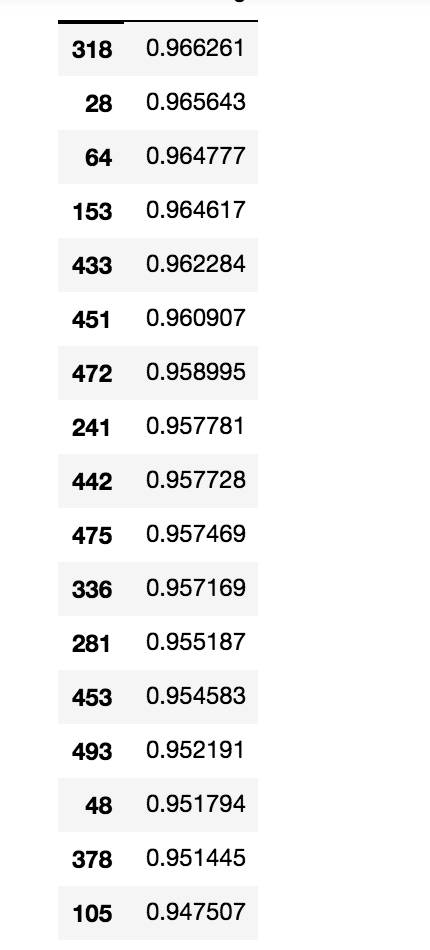
**Dataset Conclusion**

The first madelon dataset I worked with had only 500 features. These features required little data cleaning. I imported a train and test data set. This meant that I did not have to a train test split since I already had the data separated. I then imported the label data set. This acted as my target in order to try to find the features that correlated best with the target. I did a benchmark models on the training dataset to see which model performing the best. The DecisionTree Regressor performed the best.

I then processed to use that model in a function that return the Residual squared scores. This gave the below top 20 features. After the top 20 features, the scores went negative which meant that model did fit the data well. The closer the Residual score is to 1, the better the model. It means the model is fitting well. I also then took 3 samples of 10% percent of 500 features and ran the same function in order to make sure it returned the top 20 features that it did.





Theses 20 features had the best score. After these top 20, score went negative. I then took the top 20 features and generated dataset with only those features and the target to run the DecisionTreeRegressor model with just those features. I used since it had scored the highest in the benchmark test.

For the bigger data set, I had to pull data from online dataset. I followed similar steps. I also took a sample of 6,506 data points to test because based on the overall data, with only a one percent margin of error and 90 percent confidence interval, this would give us the highest scoring sample amount. I also created a benchmark model test of different models to see how they performed to find which one to get the top features for. I used SelectKBest in order narrow down the features instead of the R2 score. The SelectKBest uses a function, in the case, set it to f\_classif, to find the highest scoring features based on how many features you have selected to test. These results gave the following top features: 271, 683, 703, and 922.

