*Section 1: Intro:*

The ability to analyze and understand pictures using machine learning techniques can create powerful technology for the end user. Imagine the ability to analyze thousands of photos in a split second to determine a conclusion.

During a machine learning conference in New York, a number of hedge funds were utilizing these techniques to analyze satellite imagery data in order to find if parking lots in retail centers were full or empty. The power of that data translated into a billion dollar short position against a retail company that may show a history of low foot traffic.

Outside of wall street, there are a number of other practices that utilize imagery data like the healthcare industry. A preliminary x-ray could indicate the makings of a tumor, a disease that may attack the immune system, or find discrepancies in an athlete’s running habits well before a doctor can see them. The importance of utilizing this technology can have significant benefits for society.

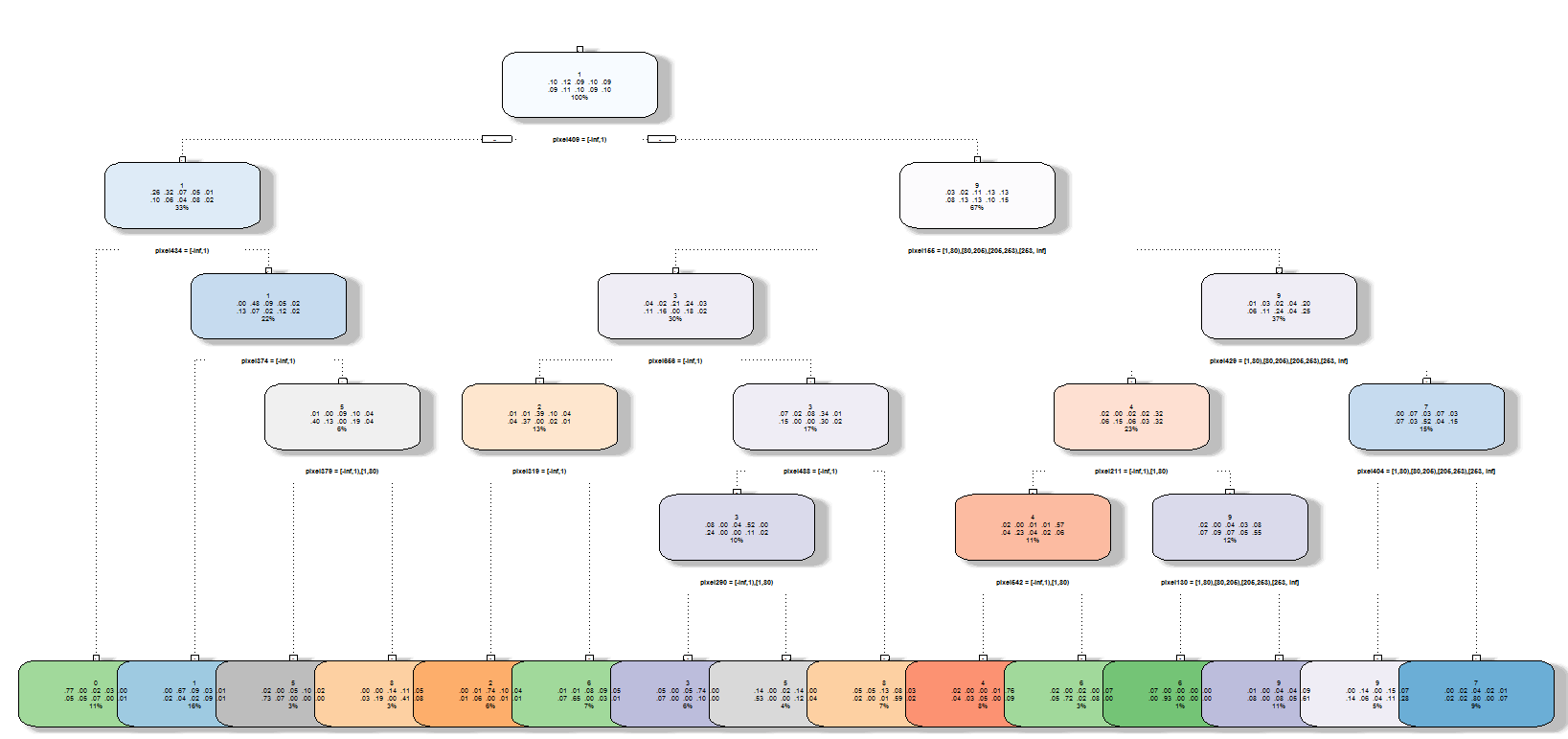
*Section 2: Decision Tree*

The data set is a data frame of values that correspond to the gradient of shading in each pixel. For example 255 (the highest value) indicates a very dark shading, or what the pixel would consider part of the number. The number 0 corresponds to non-shaded values or “white space”.

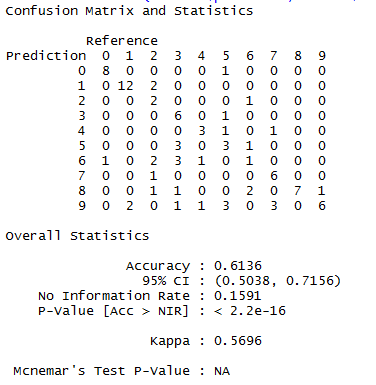
The decision tree followed a number of attempts to increase accuracy while also limiting complexity. First, different ways to discretize the data were utilized, and how to find important data points. Generally this was done by creating a fixed width:

* 0 values
* Numbers 1-25
* Numbers 26-50
* …
* Numbers 226-250
* Numbers 251-255

After running the decision tree analysis, Pixel 409 seems to be an important classifier. if it has a value (meaning that pixel has any kind of shading), the decision tree points to numbers 2-9.

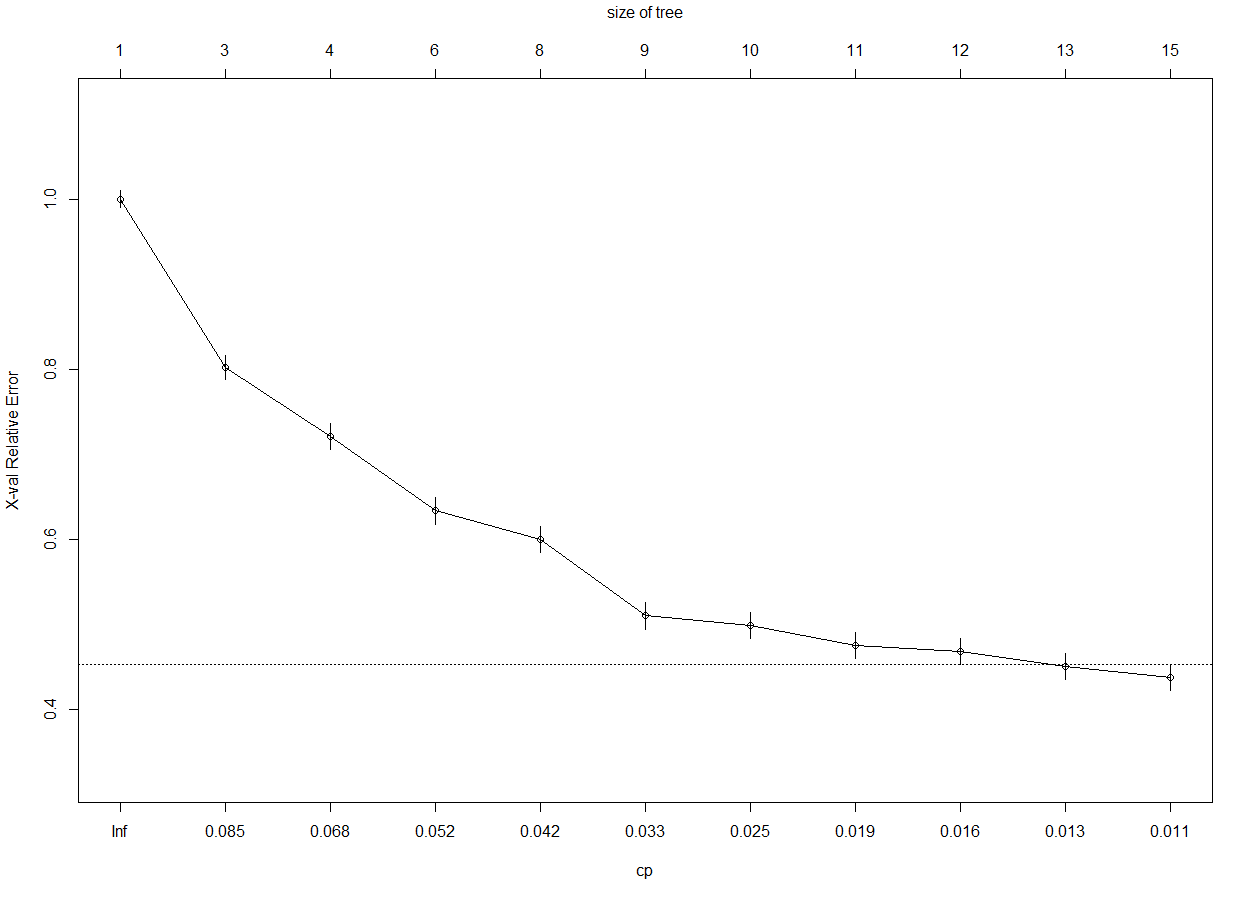
**Figure 1:**

**Figure 2:**

The confusion matrix shows that the decision tree performed at a 61.4% accuracy and showed a very low p-value. This seems to indicate that the decision tree did pretty well at finding numbers. But still 61.4% accuracy could be improved.

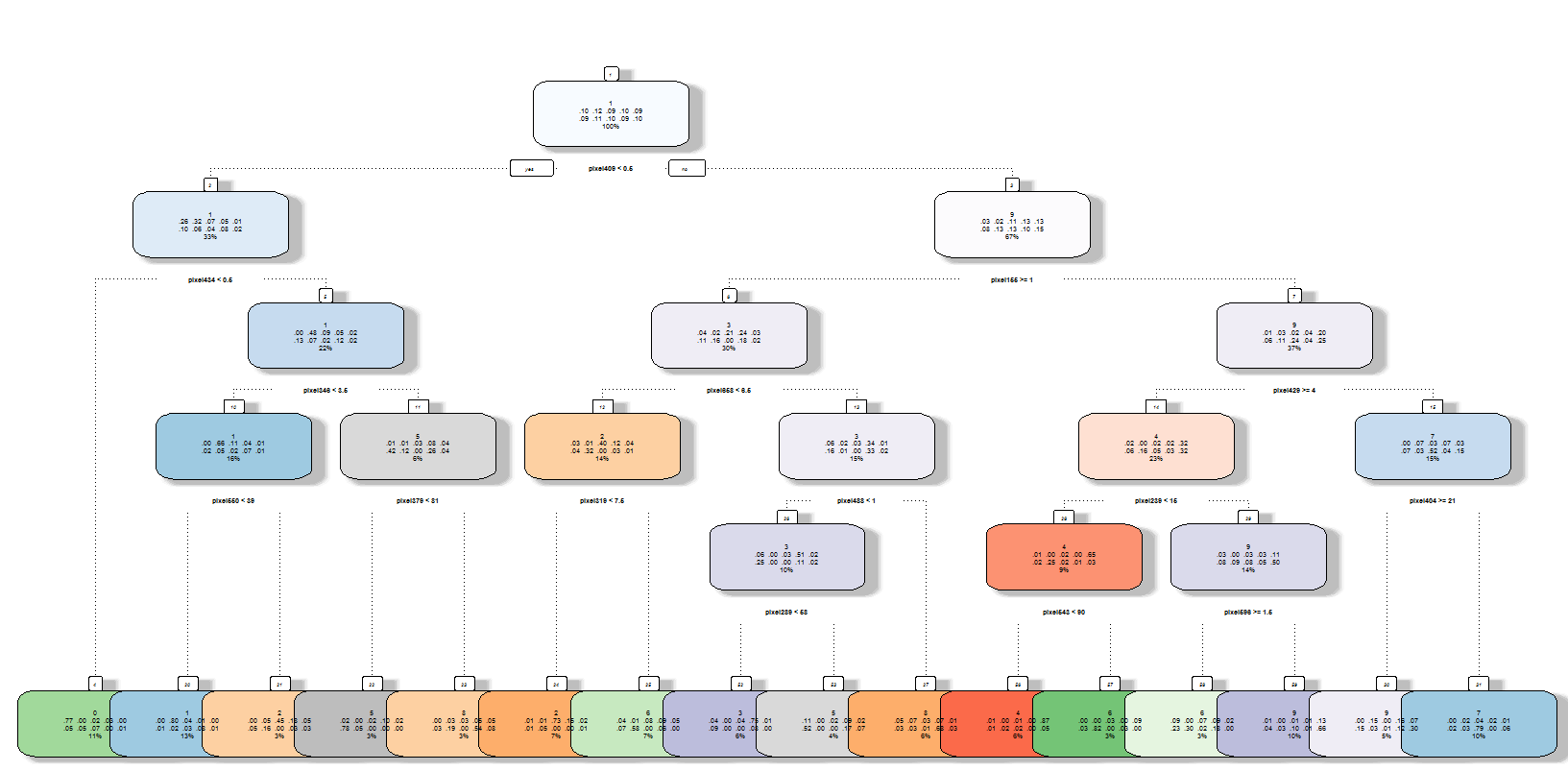
Next, pruning may help ascertain if the model can be improved and potentially simplified. First, calculate the complexity parameter (CP) to provide optimal prunings. The reason for this is to see if there is any overfitting error.

The CP error seems to indicate that optimal size of the tree is around 15. Which seems to be a complex tree, but the dataset is large. See figure 3 that shows the complexity error plotted against the size of tree. Figure 3 shows that further pruning is not necessary for the discretized data.

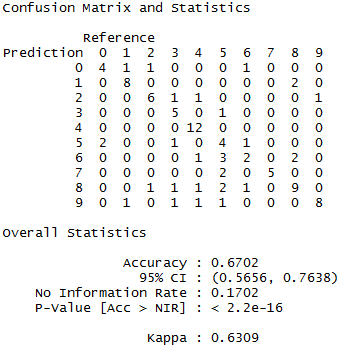
**Figure 3:**

Next the decision tree with the original data yields a slightly higher accuracy rate with more complexity. Again, pixel 409 seems to be an important node in the tree.

**Figure 4:**

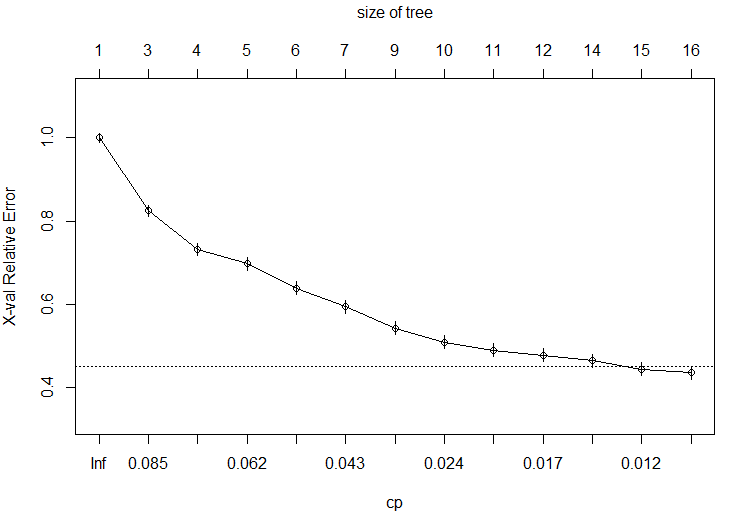


**Figure 5:**

The confusion matrix indicates a slightly higher accuracy, but as detailed above the tree is a little more complex, there may be an opportunity to prune the branches.

Again, figure 6 seems to indicate that there could be some pruning, to get the size of the tree to around 10 nodes. However, the accuracy of the model at 67% is fairly good. However, Naïve Bayes may be able to show better accuracies.

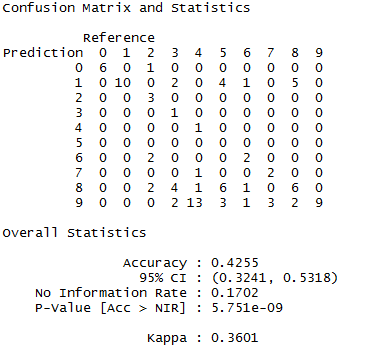
**Figure 6:**



*Section 3: Naïve Bayes*

Naïve Bayes attempts to estimate conditional probabilities form a training data set that assumes independence between the features. The first attempt with non-discretized dataset took some time as the number of probabilities to calculate was very large. The time to run the formulas to find a prediction was around one minute, which is generally too long for most settings and the accuracy was lower than the decision tree models at 42.6%.

**Figure 7:**

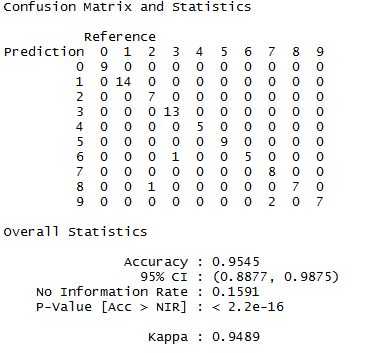
Because the model was very slow at it’s prediction and had issues in prediction accuracy a number of measures were implemented to yield more favorable results.

The data was discretized at different value ranges to lower the computing cost of the model. At first the ranges were cut into 5 equal number ranges based on number frequency. However the accuracy was very low (sub 5%).

The next attempt was a fixed length between values and then group the “ends” of the data together. Near the 255 had the highest frequency of non-zero values. So the discretized data created a categorical variable for:

* 0 values
* Numbers 1-25
* Numbers 26-50
* …
* Numbers 226-250
* Numbers 251-255

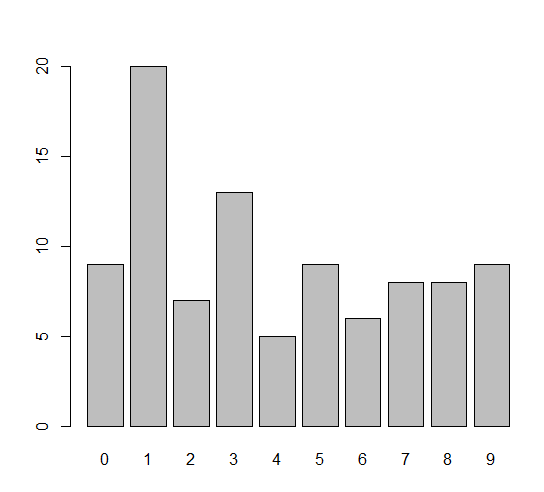
After running through the discretized process, the accuracy was very high at 95.5%. And the main area where the classifier ran into issues was differentiating between 7 and 9.

**Figure 7:**

The accuracy may be improved by increasing the training set to include more 7’s and 9’s. Also an area of concern in the prediction model was the numbers 7 and 9 represent smaller numbers in the test set in comparison to 1 and 3 which the model rarely erred on.

However, the accuracy measure for this measure of discretized should indicate a strong model for prediction.

**Figure 8:**



Section 4: Results

Overall the decision tree performed well if the data isn’t discretized, but Naïve Bayes tended to be more accurate once the data was manipulated. This solved two problems for the Naïve Bayes classifier, it lowered the computation cost to a minimal level and it yielded far more accurate predictions. A large reason the Naïve Bayes estimation was more accurate was due to independence of the vectorized data. Pixel 78 was likely independent for each picture and it is highly unlikely that one picture affected any other in the data set. Naïve Bayes performs well when the “Naïve” portion is true.

The decision tree on the other hand, isolates one pixel at a time and tries to create a binomial probability outcome for each node. Powerful, but in this case the model didn’t perform as well as Naïve Bayes.

*Section 5: Conclusion*

The results of the training set were very accurate and showed at 95% accuracy measure. This indicates Naïve Bayes can be a powerful predictor when using pixel data. Another strength of Naive Bayes, is the ability to train the algorithm in areas that it doesn’t perform particularly well in.

This is a powerful decision making tool can help wall street firms use pictures to predict sales for different companies, or allow doctors to diagnose maladies earlier, giving the patient added time to fight a disease.

Relatively simple techniques can yield unimaginable results in a matter of seconds with a simple understanding of the data set being analyzed, a tool every data scientist and casual user of data need in their toolbox.