K-Nearest Neighbor

→ Starting with the K nearest Neighbor Methodology we chose the variables to implement this with, based on if they were of quantitative values.

→ With the 11 quantitative variables used, the KNN method created training + test datasets (70/30 split) (with min-max normalization scaling) to develop the KNN model Object and predict the classification of the data.

→ The outputted Accuracy for K=1 K=3 and K=5 of the variables shows us that NumSeedInvestors, NumAngel or CapitalVentureInvestors, and Age may play an important role in the success of a start up company.

→ This is because of their higher accuracies than the others.

~~We did see some variance between tests of Age and this could be attributed to KNN's sensitivity to~~ Age however during the test varied occasionally making its precision/repeatability worse than the other seemingly important.

→ Of the lower Accuracy outputs from the less-important variables, we can see 3 potentially important variables, given their higher accuracy's than the rest and are marked with the red underline.

→ Because KNN is one of the simpler models, we can use its accuracy as a baseline for future models... our KNN model's accuracy averaged around 70% giving us the goal of 70-85% accuracy in future models.

Naive Bayes

→ Next we implemented Naive Bayes with qualilitative variables.

→ Begining with Business2Business Or Business to Consumer because of its moderate correlation, we compared the Naive Bayes predictions with actual data. The NaiveBayes model incorrectly classified 18/61 predictions outputting an error Rate of 0.3 or 30%.

→ Because NB works better with more features, we included Local Global, another moderate correlation variable to the method. (to be able to determine which class it would fall in)

→ Doing so improved our error Rate of assumption to 0.26 or 26% having misclassified only 16/61 occurances. this time.

→ Then repeating the process with a less correlated variable, HighestEdu (graphed 3 levels) is seen to improve our Naive Bayes assumptions to an error rate of 0.23 or 23%.

Lastly we compare the NB assumptions of all categories to the Actual Status data and see the error rate improve more, only misclassifying 8/61. We can see the improvement of our NB model with each feature included. The overall error rate is 13-22% based on our model which lets us know that it has an accuracy of 78-87% which is above our initial baseline goal. Making this an accurate model.