Assignment for Affinity Marketing Lab

1. The goal is to increase recall to cast a larger net and balancing helped that as opposed to weighting
2. Decisions
   1. Compare ROCs 3 in 07\_06
      1. I think that they considered using alternate model types, however decision trees performed the best
      2. Test
         1. Decision Tree
            1. Params

Switching to information gain and Gini were not advantageous nor was modifying leaf size or split size

* + - * 1. Modifying may gain some extra precision points
      1. Naïve Bayes
         1. This classifier was not considered so I want to see its performance
      2. Knn
         1. Params

K = 10 performed pretty well a k=100 was tested but looked closer to random so k was reduced

* + - * 1. This classifier was not considered so I want to see its performance
      1. Neural Net
         1. Params

Cycles = 200, learning rate =0.09

* + - * 1. This classifier was not considered so I want to see its performance
    1. Report
       1. It appears that the designers chose correctly using a decision tree with the parameters they chose. The results from the neural net were interesting however because it performed much worse than I thought it would as it was the worst performing model. One reason why regression models as opposed to Naïve Bayes would be due to the fact that Naïve Bayes considers factors independently, which does not make sense in banking data where a higher income might suggest a customer would pay for an account, and therefore have dependent attributes.
  1. Replace Missing Values 03 in 02 Preparing
     1. I think they chose zero and unknown in an effort to provide a conservative estimate as oppose to overestimate the number of new customers.
     2. I am changing family to single, and income to average in the replacing missing values operator
     3. I want to take a slightly less stringent estimate for the classifier to see if that helps classification or not.
     4. Recall 75.4 40.22 precision 93.79 12.05 74 72.75
     5. Test
        1. I changed family status to single and chose the average for the income
        2. I thought that this would be more ambitious in the classification since at one point every person was single and averaging the income level would bring some unknowns up and some down.
        3. After modification, the results were the same as the original data table. This makes me feel better about the original parameters as that means that taking a less conservative approach to these two parameters does not increase the information gained by the classifier.

1. My Change
   1. In the file 07\_02 Prepare data for mining, I was curious about looking into covariance as a method to determine related variables to see if any should be removed and what variable is influencing what
   2. I placed a multiply operator, then a covariance matrix between the Remove attributes not needed and the store table operation.
   3. The lowest correlated attributes are:
      1. Any attribute and Number of accounts, either 1,2,3 or all together
      2. Cash withdraws average and average account duration are the lowest correlated that is not something correlated with number of accounts.
   4. The highest correlated attributes are:
      1. Income average and insurance average
      2. Income average and credit card average
      3. Cash withdrawals average and income average
   5. I removed average account duration and cash withdrawals average to see if that would improve classifier performance
      1. Their removal decreased performance slightly, in the cross validation it changed 4 samples from true positives to false negatives which is decreasing the model’s accuracy.
      2. Removing the true attributes only removes information and based on the cross validation, information that is important to classification, thus they should be included. However, I am interested in using the covariance method on a larger dataset to see what variables are possibly related to each other in the future.