Statistical Data Mining Swapnil Awasthi

## Homework 4

*Create a new dataset using the same price changes from the past 25 years of S&P Adjusted Closing Prices from Finance.Yahoo.Com.*

*Add to this the change in interest rates, similarly from the previous 25 years.*

*Choose a third category (oil, foreign exchange rates, CPI) and include those changes.*

*This creates 15 columns of data to predict the price change.*

*Modify this price change to a categorical value for:*

*Awful (Change <-1 standard deviation)*

*Bad (-1 stdev <= Change < -.3 stdev)*

*Unchanged( -.3 stdev <= Change < .3 stdev)*

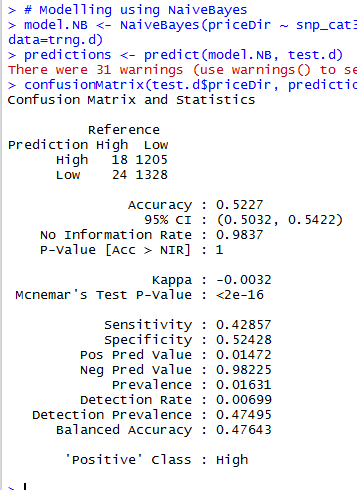
*Good (.3 stdev <= Change < 1 stdev)*

*Great (Change >= 1 stdev)*

*Model the price change using the three models and determine if any of them perform well. Determine a reasonable experiment (cross validation, testing/training) and give an executive summary of your findings.*

***Link to R code****: https://github.com/swapnilawasthi/sdmhw5/blob/master/hw4\_soln.R*

**Naïve Bayes**

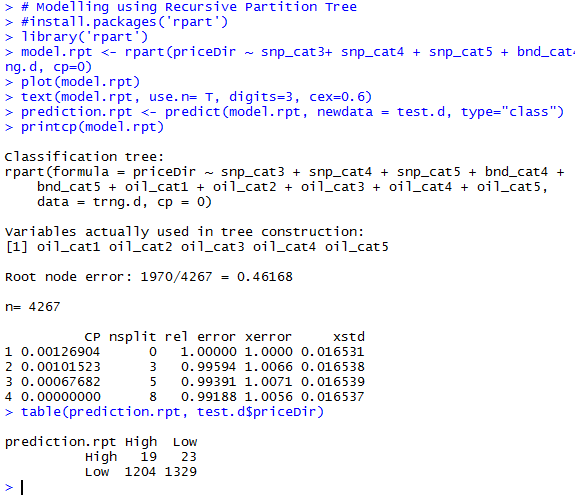


*Naïve Bayes confusion matrix*

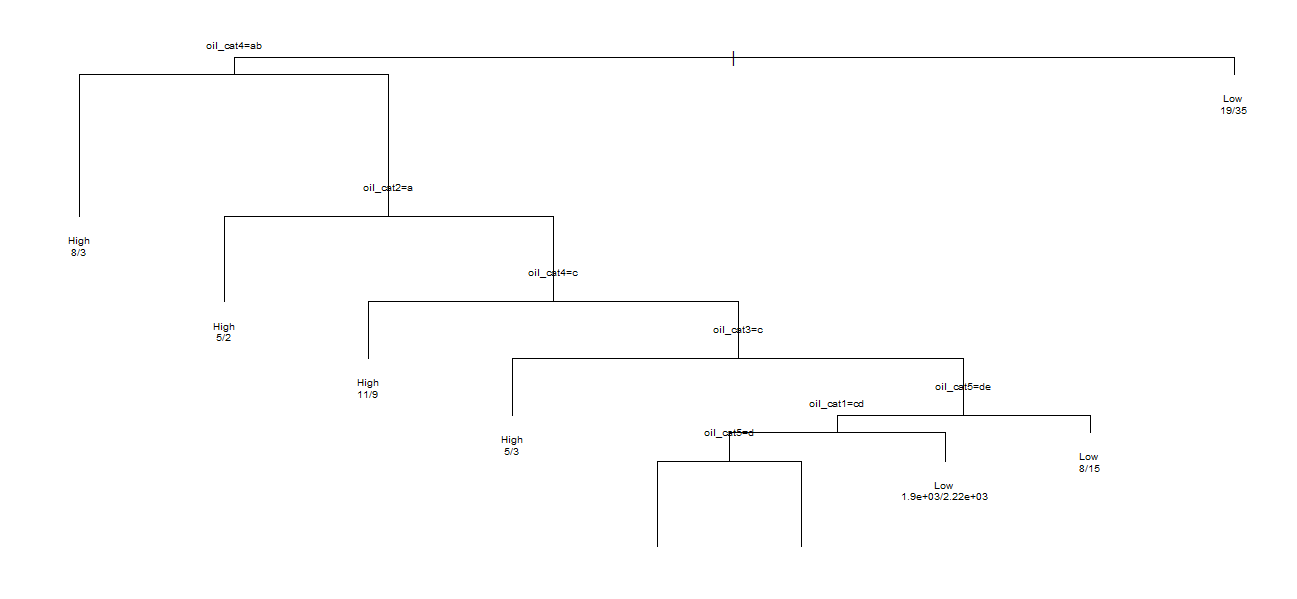
Summary: Our Naïve Bayes model is giving an average accuracy of 52.3% with an 95% confidence that our values will be between .5032 and .5422.

Our true positive rate is .428 and true negative rate is .524, our model is better at predicting proportion of negatives that are correctly identified.

**Recursive partition tree**



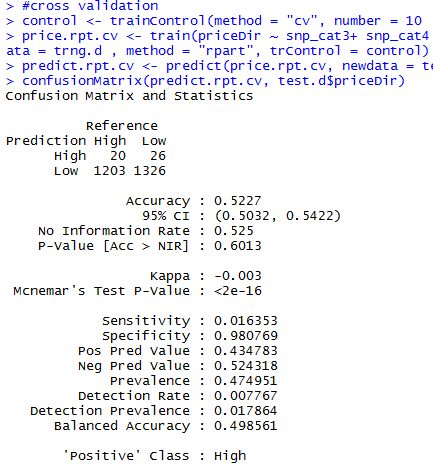
*Recursive partition*



*Recursive tree*

Our r part tree model is also performing pretty average.

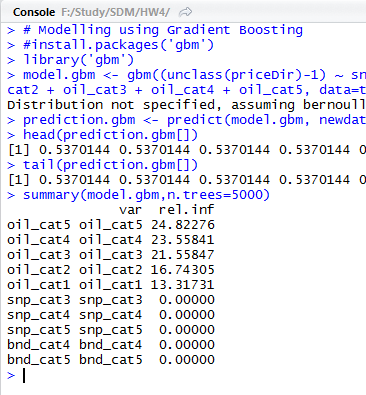
**Cross validation**



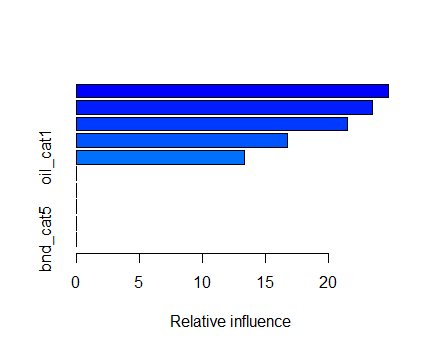
*Cross validation*

Cross validating our model 10 folds increases the specificity to 0.98.

**Gradient boosting model**



*Gradient boosting model*



*summary of gradient boosting model*