**Comparison between algorithms**

I had two target variables: IsDiscontinued (Boolean) and PrimaryCategory (factor, 13 levels). Different algorithms worked well for different target variables.

**Nearest Neighbour knn**

Cannot deal with missing data: had to remove some columns (or would have to remove some rows). Either way you are missing out on important data.

Can only deal with numerical data.

Knn algorithm takes more time to run than any other algorithm we have learnt about so far.

Primary Category: there were 13 different types and only 3 types that had most of the data. All the categories were sorted into those 3 types.

IsDiscontinued: most of the predicted results were false because there was a higher proportion of false’s (~88% false and ~12% true) in the data.

**Naïve Bayes**

Cannot deal with missing data: had to remove some columns (or would have to remove some rows). Either way you are missing out on important data.

Can only deal with categorical data.

Primary Category: the prediction were better compared to knn. The predictor subcategory helped bring about 99.3% correct predictions but without it, only about 75% of the results were predicted correctly.

isDiscontinued: following the same bias as seen in knn, most of the predicted results were false because there was a higher proportion of false’s (~88% false and ~12% true) in the data.

**Decision Trees**

Can deal with missing data ,

Can deal with both categorical and numerical data but couldn’t deal with categorical data that had too many levels.

isDiscontinued: worked amazingly here (the test error rate was only ~2% same as the training error rate)

Primary category: gave best results (~82% correct) compared to all other algorithms.

**Rule learning**

Can deal with missing data.

Can deal with both categorical and numerical data.

While JRip was a lot better than 1R for both target variables, both the rule learners were very susceptible to small changes in the data. I removed one level of variables from PrimaryCategory because there was too small a percentage present in the data and either the training or test dataset was entirely missing the category when the random split was done. The JRip and 1R results changes drastically.

Also, the training and test dataset error rates varied hugely (for e.g. where the training dataset had 86% errors, the test dataset had 2% errors).