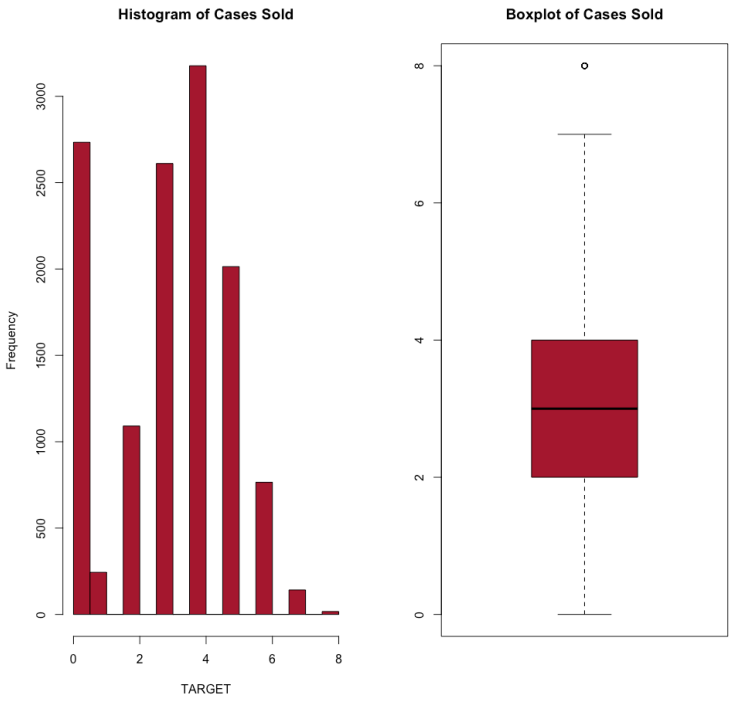
Bingo Bonus – Random Forest importance used in variable selection, Logistic regression + Poisson attempted along with Decision tree regression – see bonus section at the end.

Introduction

The data for the exercise are measurements for various attributes of wine. Each observation in the data set constitutes a profile for an individual wine. The target is the number of sample cases ordered for that wine. The goal of the modelling effort is to be able to predict the number of sample cases that will be sold, which in turn would permit the manufacturer to adjust inventory to maximize sales.

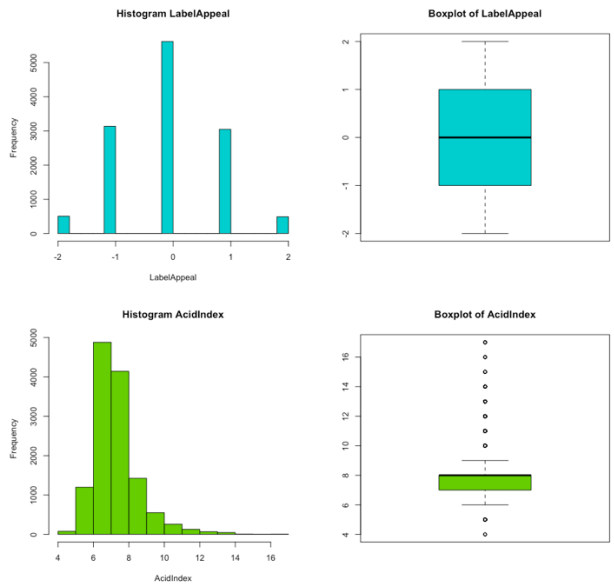
Data Exploration

An initial summary of the training data shows there are issues with missing values. The training data have a variance of 3.71 and a mean of 3.03; the variance is larger than the mean so this may be a case of the negative binomial solution being the better choice. The target variable distribution looks Poisson-like:

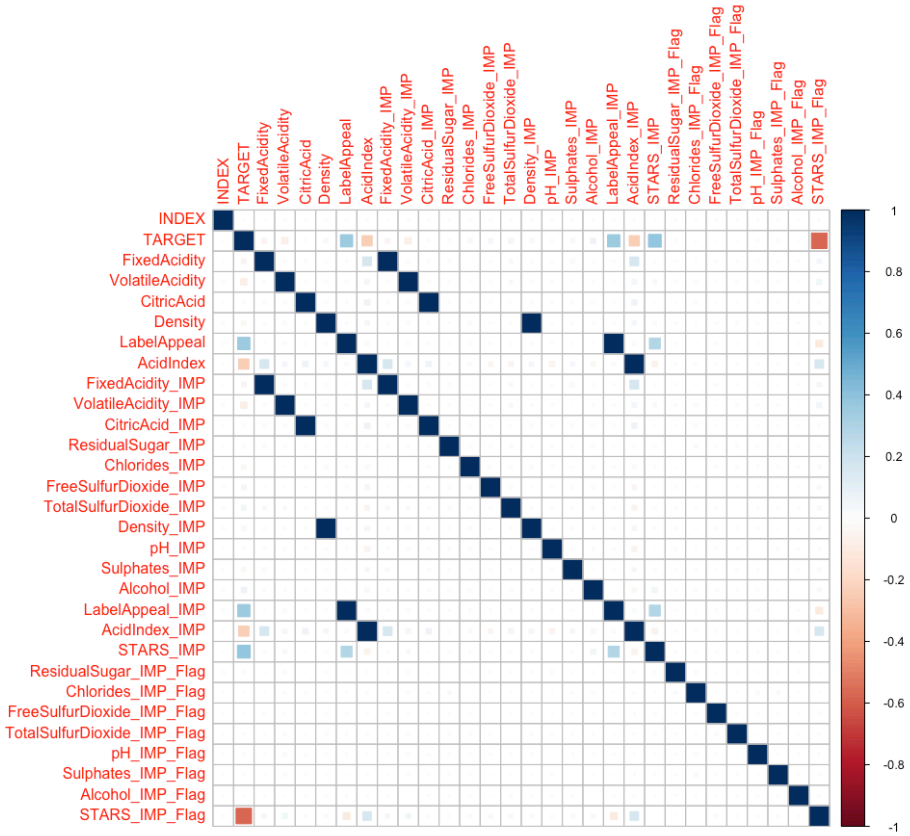


The outlier in the number of cases sold ("8 or more") is noted.

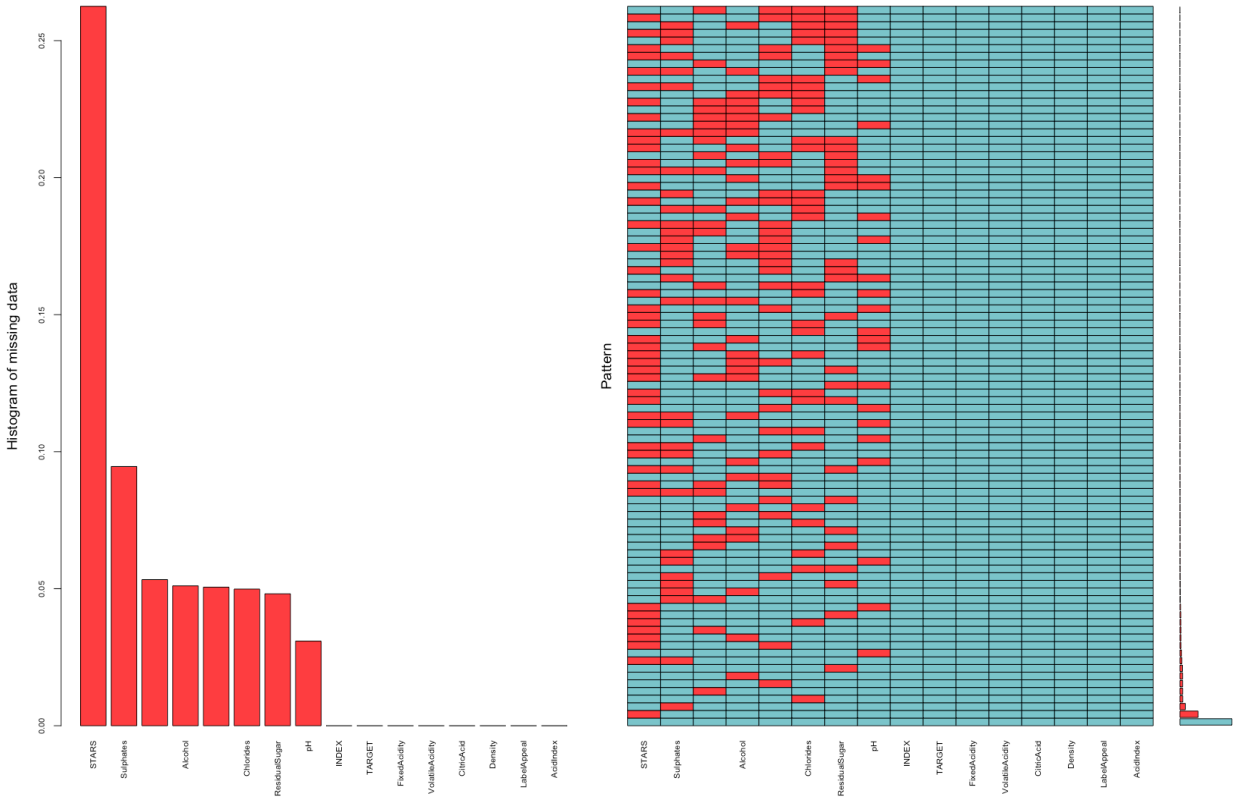
For the most part, the features have normal distributions, the exception being the Acid Index shown below. To address the skew in this feature, it is transformed as part of the data preparation.



Other than expected correlations between features and the imputed version of the features, there don't appear to be any multi-collinearity issues.



The STARS rating is the most frequently missing piece of data, which explains why knowing if it was imputed appears to have predictive power. The distribution of missing data is visualized as:

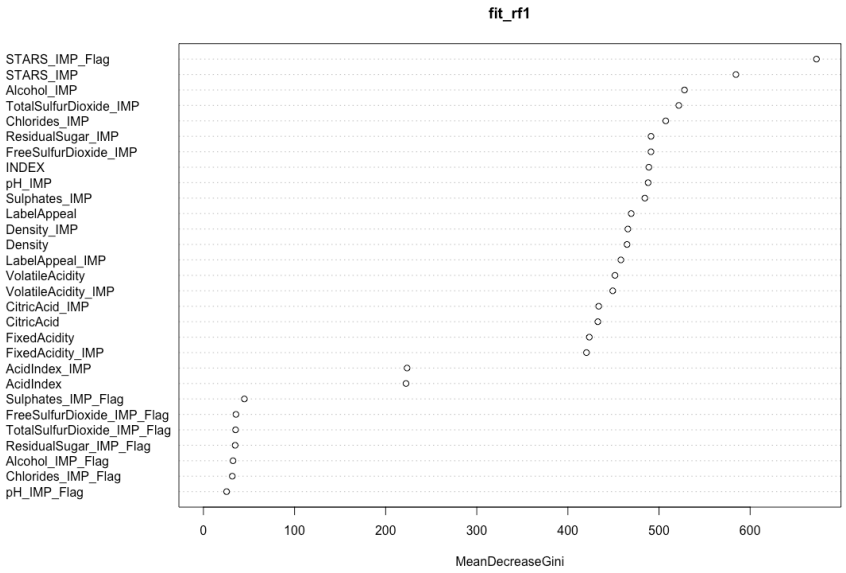


Data Preparation

Flag variables were created to track which features were imputed. The imputed, complete-cases data are stored in new columns as well, to separate them from the original data. Imputation is being done by using the mean of the non-NA column values. As mentioned above, the Acid Index feature required transformation to address skewness, it is transformed by taking the log. A data frame comprised of only complete cases was also created for doing correlations.

Model Construction

I began by running a random forest variable importance on the data, producing the following:



The relative importance was used to inform various model building attempts.

Two linear models were fit as reference; the first model uses fewer features than the second. Then Poisson, negative binomial, zero inflation, and zero inflation negative binomial models are fit using different combinations of features selected based on the feature importance.

Model Selection

Model selection was done based on the AIC and the -2*log(Likelihood). For my exploration, the best model was a zero-inflation negative binomial version (zero_inf_nb3). The selected model has an AIC of 41161.24 and contains the following features:

STARS_IMP_Flag, STARS_IMP, Alcohol_IMP, TotalSulfurDioxide_IMP, Chlorides_IMP, FreeSulfurDioxide_IMP, pH_IMP, ResidualSugar_IMP, Sulphates_IMP, LabelAppeal_IMP, Density_IMP, VolatileAcidity_IMP, CitricAcid_IMP, FixedAcidity_IMP

Conclusion

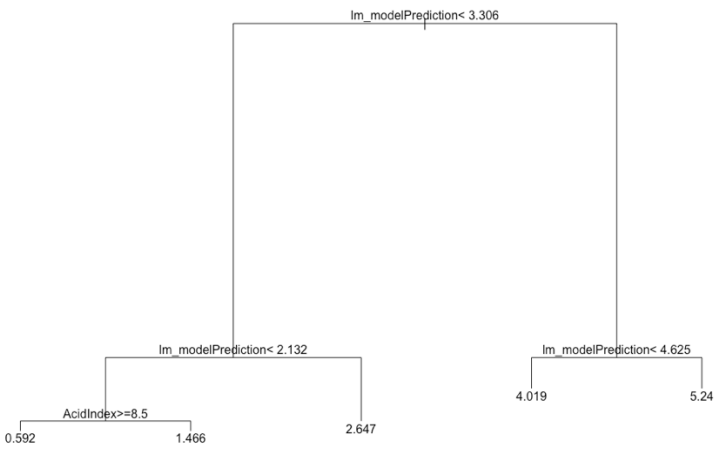
The comparison of the variance and mean indicated that the negative binomial regression may perform better than the Poisson regression for this data. After creating models using both regression methods, I selected my best performing model which was a zero-inflated negative binomial (ZINB) regression. The regular linear regression model did outperform the all of the Poisson models and the negative binomial models. The UCLA IDRE website suggests that ZINB is most appropriate when you are predicting count variables and the zeros can be modeled separately from the non-zeros. This seems to fit our situation; if a sample is ordered (the zero

case) could be modeled, and then the number of cases for those who do buy. So, the result of ZINB as the best model makes sense to me.

Bingo-Bonus Write up

1. I attempted a logistic model which then fed a Poisson model. According to the summary data, the mean is much lower than expected. The number of 0 targets is roughly the same percentage as in the training data. I don't know how to compare a 2-step model to a 1-step, the individual AICs are much help in comparing, so I didn't use this as my Champion model – I just don't know how to tell if it is "better".

2. I fit a decision tree to predict TARGET. The visualization for the tree is



Looking at the predictions, they aren't very good. The only predicted values are the leaf nodes shown in the graphic.