Unit 3 Assignment: “Wine Sales Project”

Alan Kessler, MSDS 411, Section 56

Intro

The goal of the assignment is to predict the number of cases sold for a wine based on its attributes. The data contains information about wine rating and how appealing the label is as well as information about the chemical composition of the wine.

Several models are fit to the data with strong consideration to how well they fit a validation data set representing 30% of the training data. The fit involves added complexity due to a zero-inflated feature in the data.

Bonus

For bonus points, a decision tree model is considered and the combination of a logistic regression and a Poisson regression is ultimately selected.

Data Exploration

Observing the target variable can provide input as to which techniques will fit the data best. Figure 1 shows the distribution of the target in the training data. There is a large mass of observations at zero. One reason for this is that if someone decides to buy a particular wine, they are likely to buy more than one case of it.

This data structure indicates that zero inflated or hurdle models may improve the fit. For wines that are purchased, the number of cases in excess of one has a mean greater than the variance. This is an important piece of information if a Poisson hurdle model is considered.

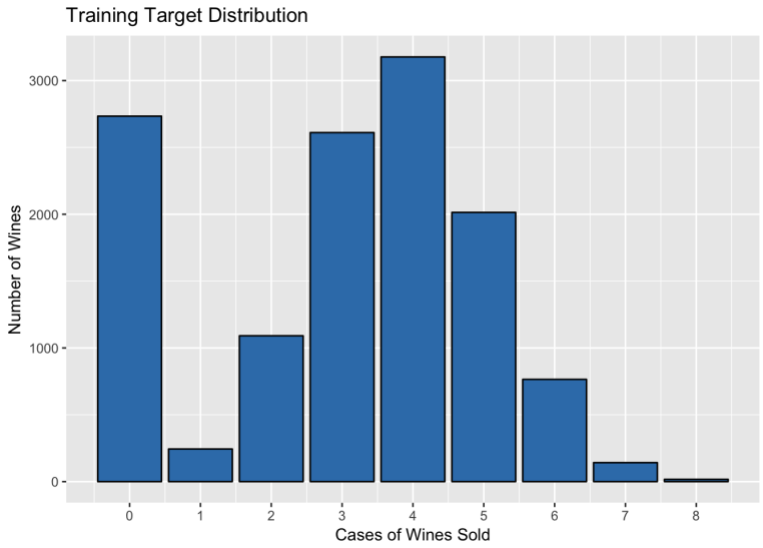

Figure 1 – Training Target Distribution

Next, missing values are observed. Several of the chemical measurements such as pH and alcohol levels are missing in the data. Figure 2 shows the portion of observations with missing values for those variables with missing data. The presence of missing values does not appear

consistent across all measurements as a wine with a missing measurement is likely to have other completed measurements. Approximately a third of the wines are missing a rating as well. This may indicate that not all wines are rated.

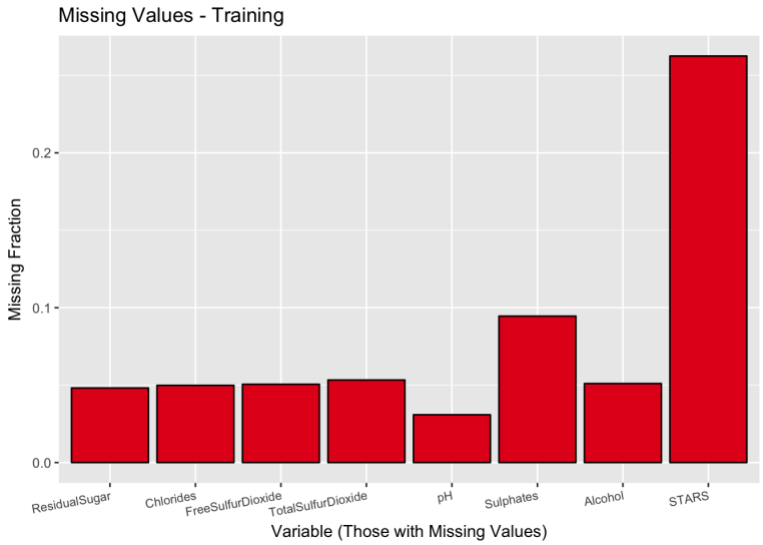

Figure 2 – Missing Values in Training Data

Figure 3 shows the correlation in the training data. Multicollinearity appears to not be very likely as measurements lack strong correlations with each other. However, they are also only marginally correlated with the number of cases sold.

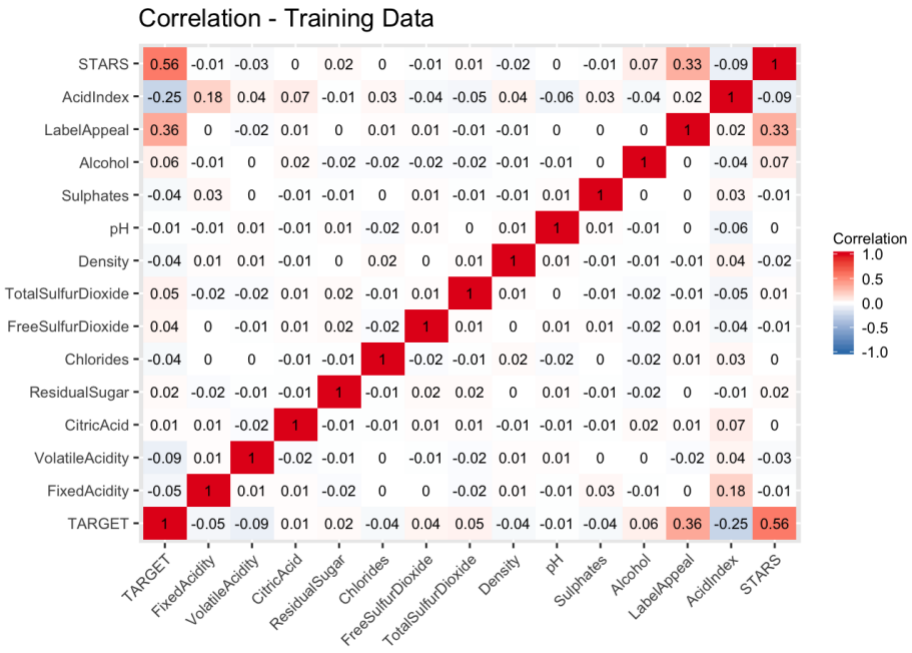

Figure 3 – Training Data Correlation

Figure 4 shows box plots of the predictors. The training and testing data appear to have similar distributions and many of the measurements have heavy tails.

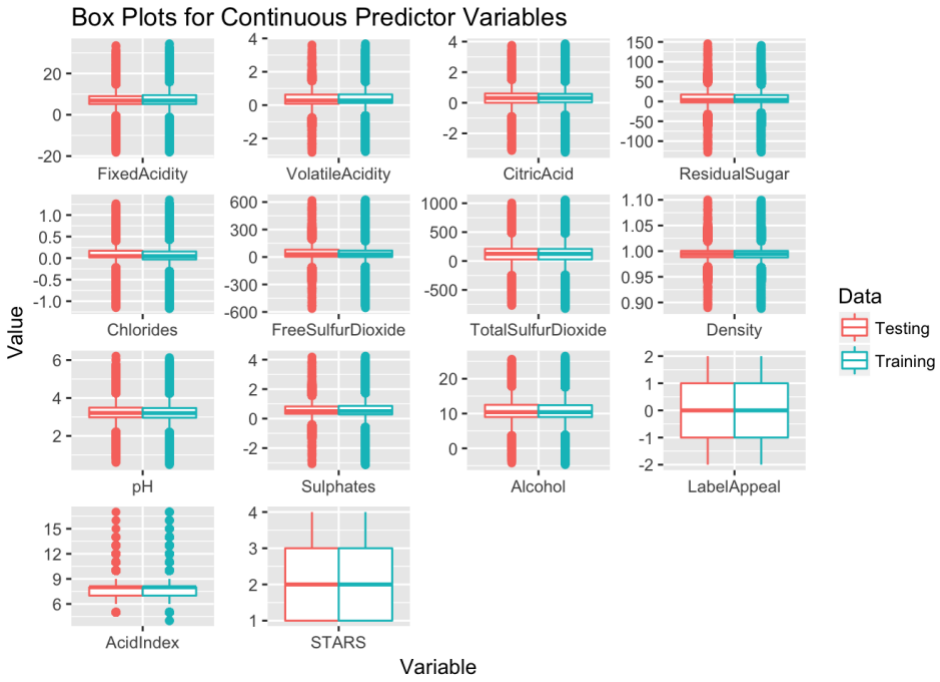

Figure 4 – Predictor Variable Box Plots

The outliers make it difficult to observe the univariate relationships between predictors and the target. An alternative way to do this comparison is to plot the target instead on the x-axis and show the average predictor value for each value of the target on the y-axis. This type of plot is shown in Figure 5. Due to the difference in the plot type, the only interpretation that can be made is that the predictor variables do differ on average by the number of cases sold.

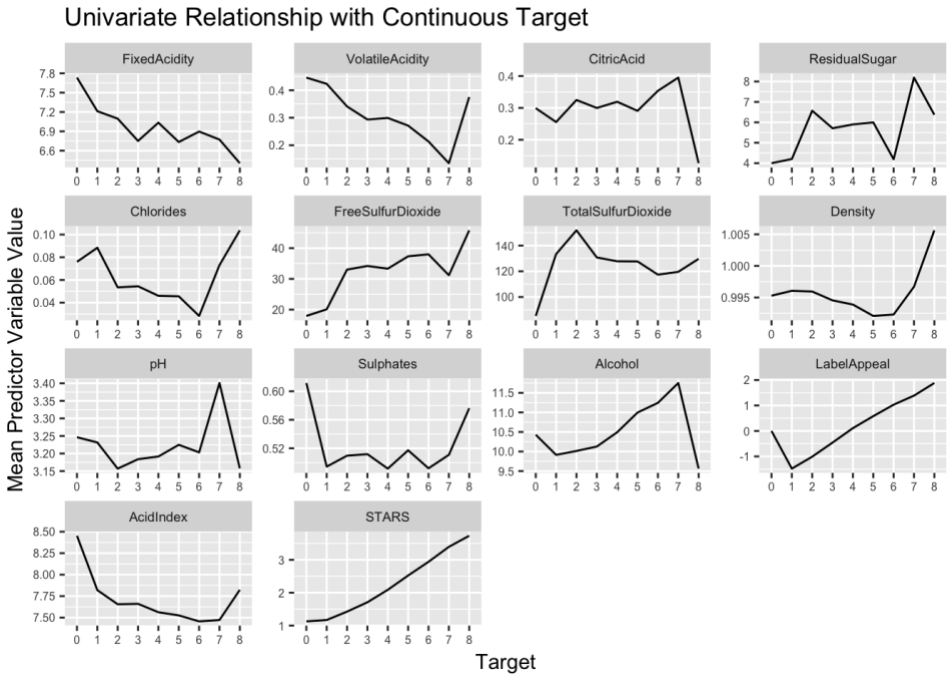

Figure 5 – Univariate Relationship (Axis Switched)

Data Preparation

First, the ratings are coded as categorical variables. This allows for missing values to be in a distinct category. Given that the rating was the variable most correlated with the target, it is important to provide sufficient flexibility to this variable.

Label appeal is another important variable from a univariate perspective. Given the low number of distinct values it can have, it is also coded as a categorical value.

To account for missing measurements, missing values are imputed with the mean of the training data. Indicator variables are created based on the presence of a missing value.

The outliers in the data result in the need to cap and floor the predictor variables. The top and bottom 2.5% of the training data define the capped and floored transformed variables. An outlier wine may intuitively be less likely to sell as much, so indicators are created for cases in which a flooring or capping occurs.

Build Models

The models built on the data are assessed based on their performance on a 30% hold out partition of the training data. The first model is a multiple linear regression baseline including only the variables most correlated with the target: rating, label appeal, and acid index. All of the variables in the regression are significant, match intuitive relationships, and variance inflation factors (VIF) are all lower than two. The mean squared error (MSE) on the validation set is close to that of the training at 1.735.

Model 2 incorporates the chemical measurements by applying backwards selection to a multiple linear regression. VIFs continue to stay small which is expected based on the correlation plot. The variables which have intuitive relationships like ratings and labels persist. The measurement values and related indicators make it into the model. The capped and floored indicators have negative coefficients which match the hypothesis that niche wines will sell fewer cases. While the training MSE shows some improvement, the validation MSE's improvement is minimal at 1.725.

Model 3 applies backwards selection to the same data using a Poisson regression instead. Again, the intuitive relationships hold and multicollinearity is not a concern. The validation MSE shows an improvement at 1.716.

Model 4 uses zero-inflated Poisson regression. Due to issues getting the model to converge, manual variable selection is used starting with the variables found in Model 3. While the structure of Model 4 matches the problem better than previous models the convergence issues are likely to blame for the relatively high MSE of 1.727.

Model 5 uses the same terms as Model 3 except instead of Poisson regression, negative binomial regression is used instead. The coefficients and fit statistics are identical in this case which is an interesting feature of these kinds of regression. Other selection methods are considered but ultimately not used due to convergence issues. While the coefficients are not identical, equal validation MSE is generated for Model 6 as well which is the zero-inflated negative binomial version of Model 3.

Models 7 and 8 use a hurdle approach where Model 7 is a logistic regression predicting whether or not an individual wine is purchased. Model 8 is a Poisson regression predicting the number of cases in excess of one purchased. Both use backwards stepwise selection.

An advantage of this analysis is that it allows for the interpretation of factors that create sales and factors that lead to larger sales separately. In this case the direction of the variables were similar in almost all cases, but the Poisson regression incorporated fewer variables in the backwards selection. It appears that the information available is better suited to predicted a sale than predicting how much is sold. Model 8 included the same variables as Model 1 with the addition of capped volatile acidity and the capped/floored indicator for chlorides as variables that negatively impact sales when they increase in value and alcohol level as a variable that increases sales when it increases. Model 7 included additional measurement variables but interestingly none of the capped/floored indicators were selected. This could mean that even niche wines may be likely to be purchased but when they are, fewer are purchased.

One of the limitations of this approach is that the data in excess of one is actually under-dispersed. This limits the appropriateness of the application of Poisson regression. However, the model may still produce good rankings of wine that will sell.

As expected, the VIFs are low for these models. The validation MSE for the combination is also an improvement over the other models at 1.636.

As another point of comparison, Model 9, a decision tree with default parameters from the “rpart” package is fit to the data. The resulting MSE is 1.766 which is impressive given the simplicity of the algorithm. One element to consider in the future is applying ensemble tree approaches such as Random Forest models.

Select Models

The different model forms restricted the variety of model metric comparisons to some extent. The focus for model selection was the MSE on the 30% of the data used for validation. The combination of Models 7 and 8 results in a significant improvement in MSE over the other models considered. Additionally, this approach did not have any convergence issues and the hurdle feature adds to the interpretability of the model at large. As a result, this combination is the selected approach.

Conclusion

Predicting the number of cases sold for a particular wine involved predicting a zero-inflated target. To account for this, nine different models were fit to the training data that both ignored and explicitly considered the zero-inflated feature. The final selected model was actually a combination of a logistic regression and a Poisson regression. The two models together fit the data well and are extremely interpretable.