M6: Data Assignment

Jason Merten

Data Exploration/Preparation:

The wine data sets used in this analysis contains 16 columns of data across a training and a test data sets. The training data set consists of 12,795 rows of data while the test data set consists of 3,335 rows of data. Of the 14 predictor variable columns, there are 8 columns that contain missing values: **ResidualSugar**, **Chlorides**, **FreeSulfurDioxide**, **TotalSulfurDioxide**, **pH**, **Sulphates**, **Alcohol**, and **STARS**. I also looked at the distribution of the target variable and there's a significant amount of 0 entries which will change the way we create our models. Looking at the variables, I noticed that most are centered around their means which makes imputing missing values very easy for this exercise. Looking at the variables after imputation, the distributions don't seem to be significantly altered which means we can continue with creating a few feature variables that may be useful with the models. The new feature variables, all ending with **REDFLAG** as well as the variable **TallyUp**, are to help differentiate between red and white wines. I also created another variable, **TARGET_AMT**, to account for 0 sale scenarios.

Building Models:

I created six models for this assignment:

Model 1 (lm_fit): This is a linear model that was fitted using backwards selection. Nothing special about this model, it's essentially used as a base model to judge the accuracy of the other models.

Model 2 (lm_fit_stepwise): Another linear model trained with a more advanced technique. This model also uses 10-fold cross-validation to achieve a higher r-squared and RMSE than the first model. This model is actually the third best model out of all six models with a r-squared value of .5432 and a RMSE value of 1.302.

Model 3 (poisson_model): This model is a Poisson regression that uses repeated 10-fold cross-validation to achieve good r-squared and RMSE values. This model has a r-squared value of .5363 and a RMSE value of 1.312.

Model 4 (NBR_Model): This model is a Negative Binomial Regression model, trained using 10-fold cross-validation. As with the previous two models, I only included the imputed variables, flag variables, and target variable to ensure no missing values were evaluated. This model yielded a r-squared value of .5364 and a RMSE value of 1.312, the same as the poisson_model.

Model 5 (ZIP_Model): This model is a Zero-inflated Poisson model, trained on the full training data set (no cross-validation, I couldn't figure out how to get it to work correctly), and using the predictors that Model 2 selected as the best terms. This model yielded a r-squared value of .5718 and a RMSE value of 1.2604.

Model 6 (ZINB_Model): The last model is a Zero-inflated Negative Binomial model, trained on the full training set (no cross-validation), again using the predictors that Model 2 selected. The r-squared value for this model was .5718, and the RMSE was 1.2604.

Selecting a Model:

I tried various methods to analyze all six models and decided to primarily use r-squared and RMSE (and AIC where available) to help determine the best model as these values were all available or could be calculated for each model. I chose the Zero-inflated Poisson model as my champion model due to its high r-squared value, low RMSE, and low AIC compared to the other five models.