Unit 03 Homework – Wine

KAGGLE Name: *NikhilAgarwal*

Nikhil Agarwal
Northwestern University
PREDICT 411, Section 55

I am requesting a total **90** bingo bonus points for the unit 3 homework. Please see my justification below.

Points Requested	Category	Justification
20	R Code	Much of the homework has been rewritten in R. Please see the attached homework (NikhilAgarwal_HW3_RCode.r)
10	Macros	Extensive use of macros
20	Decision Tree	Decision trees were investigated and not used. Explanation is provided on page 10-11.
20	Logistic Hurdle Model	I created & discussed a Logistic Hurdle model (see page 19).
20	Decision Tree used to predict Target	I used JMP to create a decision tree and predict TARGET. See page 22.

Table of Contents

Introduction

The intent of this assignment is to develop a model that can predict the number of sample cases of wine purchased by distributors. Over 12,000 data points were used to construct five types of models: linear regression, Poisson, negative binomial, zero inflation Poisson, and zero inflation negative binomial. Many of the data points provided offer an insight into the chemical properties of each type of wine. Various model diagnostic parameters (e.g., AIC, SBC, RMSE) were used to determine the best model.

Results

Data Exploration

The original dataset contains over 12,000 data points consisting of 14 potential predictor variables with one response variable (see Table 1).

Table 1: Variable Overview

Variable	Description
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average
Alcohol	Alcohol Content
Chlorides	Chloride content of wine
CitricAcid	Citric Acid Content
Density	Density of Wine
FixedAcidity	Fixed Acidity of Wine
FreeSulfurDioxide	Sulfur Dioxide content of wine
LabelAppeal	Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customers don't like the design.
ResidualSugar	Residual Sugar of wine
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor
Sulphates	Sulfate content of wine
TotalSulfurDioxide	Total Sulfur Dioxide of Wine
VolatileAcidity	Volatile Acid content of wine
pH	pH of wine

Two of the variables, LabelAppeal and STARS, are interesting in the sense that as both values increase, the generally accepted thought is that net sales are also higher. As for the chemical properties, the analysis posited assumes that no prior knowledge is available. The response variable TARGET (not listed in Table 1) is essentially the number of cases of wine samples sold.

Table 2 illustrates the descriptive statistics (PROC MEANS) of the variables contained within the original dataset.

Table 2: Descriptive Statistics on Original Variables

Variable	N	N Miss	Minimum	Maximum	Median	Mean	1st Pctl	99th Pctl
TARGET	12795	0	0	8	3	3.0290739	0	7
FixedAcidity	12795	0	-18.1	34.4	6.9	7.0757171	-10.9	24.4
VolatileAcidity	12795	0	-2.79	3.68	0.28	0.3241039	-1.865	2.59
CitricAcid	12795	0	-3.24	3.86	0.31	0.3084127	-2.18	2.66
ResidualSugar	12179	616	-127.8	141.15	3.9	5.4187331	-91	99.2
Chlorides	12157	638	-1.171	1.351	0.046	0.0548225	-0.859	0.957
FreeSulfurDioxide	12148	647	-555	623	30	30.8455713	-388	469
TotalSulfurDioxide	12113	682	-823	1057	123	120.7142326	-531	767
Density	12795	0	0.88809	1.09924	0.99449	0.9942027	0.9168	1.06981
pH	12400	395	0.48	6.13	3.2	3.2076282	1.32	5.125
Sulphates	11585	1210	-3.13	4.24	0.5	0.5271118	-2.13	3.16
Alcohol	12142	653	-4.7	26.5	10.4	10.4892363	0.1	20.3
LabelAppeal	12795	0	-2	2	0	-0.009066	-2	2
AcidIndex	12795	0	4	17	8	7.7727237	6	13
STARS	9436	3359	1	4	2	2.041755	1	4

This dataset contains exactly 12,795 observations, but from Table 2, it is clear to see that eight (ResidualSugar, Chlorides, FreeSulferDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, and STARS) of the 14 variables have missing values and will require imputation. Another peculiar observation has to do with the minimum values for several of the chemical properties. General intuition dictates that many of the chemical properties cannot have negative values as this would against basic laws of matter, chemistry, and physics. For instance, it is not possible to have negative values for alcohol content. The same is true for all of the other chemical properties listed. Therefore, all of the chemical properties, at a minimum, will be imputed to the absolute value of the provided negative value. If the value is missing, it will be imputed using either the mean or the median (explained in greater detail in the section "Data Imputation"

The variable LabelAppeal can have a negative value as well as a positive value. A negative value would simply imply that the appearance of the label has a detrimental effect on the net sales of the particular wine. It is not clearly understood of the impact of the variable AcidIndex. The variable STARS implies that the higher number of stars that a wine may have, the higher the likelihood of the wine selling in larger quantities.

Table 3 is the correlation matrix of the variable TARGET versus the 18 different predictor variables. The intent here is to understand which set of variables have the strongest correlation to the response variable TARGET. Furthermore, this table can also be used to understand the variable's impact on the response variable. For instance, if it has a negative sign, then it has a detrimental impact.

Table 3: Correlation Matrix of Target vs. Predictor Variables

	TARGET
TARGET	1
FixedAcidity	-0.04901
VolatileAcidity	-0.08879
CitricAcid	0.00868
ResidualSugar	0.01649
Chlorides	-0.03826
FreeSulfurDioxide	0.04382
TotalSulfurDioxide	0.05148
Density	-0.03552
pH	-0.00944
Sulphates	-0.03885
Alcohol	0.06206
LabelAppeal	0.3565
AcidIndex	-0.24605
STARS	0.55879

No variable has an impressively strong correlation (positive or negative) to the response variable TARGET. Of particular note is the variable STARS. It is clear to see that it has a moderately strong correlation of about 0.56. This is also unsurprising since it is generally accepted that the more stars a wine has, the more likely it will sell in larger quantities. Similarly, the variable LabelAppeal also has a moderate positive impact on TARGET.

Note how AcidIndex has a moderate negative impact on TARGET. This suggests that the higher a wine's AcidIndex is, the more likely it is that it will sell in fewer quantities. Surprisingly, the variables CitricAcid and pH do not seem to have any correlation (very little) to the TARGET. Similarly, it is also surprising to note that the variable Alcohol also has a weak correlation to the response variable.

Several histograms and boxplots were constructed to understand the distribution of each of the predictor variables. An example of the histogram and boxplot are represented in Figure 1. Note how the distribution is not necessarily normal and has some skewness. The boxplot clearly shows that there are many outliers. It is important to note that at this exploration point, no values have been corrected for negativity (as explained earlier) nor have any imputations taken place.

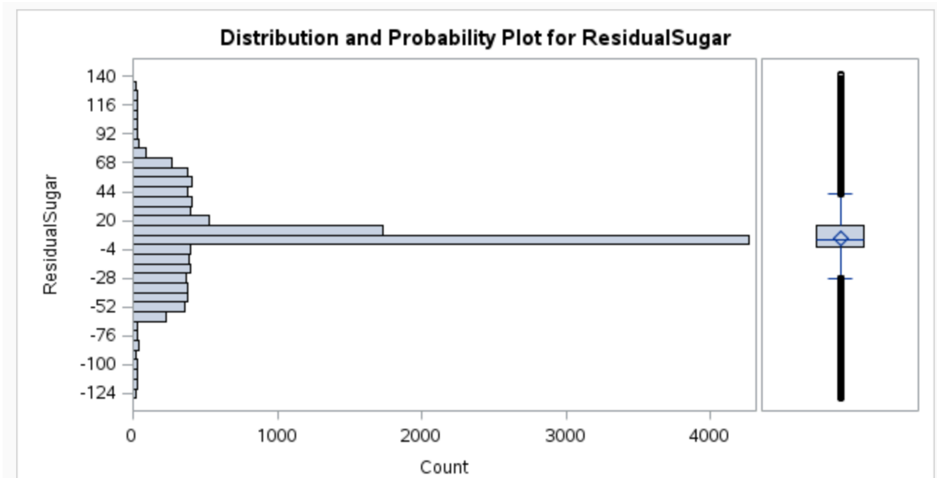

Figure 1: Histogram and Boxplot for Variable ResidualSugar

For the sake of brevity in terms of graphical outputs, Table 4 highlights the skewness[1] and kurtosis[2] for each of the predictor variables. Recall that if a variable has a skewness of 1 and a kurtosis value of 3, then it can be construed as normally distributed.

Table 4: Skewness, Kurtosis, and Variance for Predictor Variables

Variable	Skewness	Kurtosis	Variance
TARGET	-0.3263776	-0.8767876	3.7108945
FixedAcidity	-0.0225913	1.6768536	39.9126188
VolatileAcidity	0.0203847	1.8341516	0.6146783
CitricAcid	-0.0503188	1.8398842	0.7431816
ResidualSugar	-0.053136	1.886761	1139.02
Chlorides	0.0304347	1.7906222	0.1014214
FreeSulfurDioxide	0.0063946	1.8385435	22116.02
TotalSulfurDioxide	-0.0071811	1.6766257	53783.74
Density	-0.0186981	1.9019373	0.000704247
pH	0.0442987	1.6481659	0.4619745
Sulphates	0.0059134	1.7546612	0.868865
Alcohol	-0.0307234	1.5413715	13.8966348
LabelAppeal	0.0084314	-0.2614968	0.79404
AcidIndex	1.6488825	5.1938712	1.752781
STARS	0.4473775	-0.6917759	0.8145785

It’s clear to see that all of the variables are not perfectly normally distributed. Not the large variance for the variable TotalSulfurDioxide.

[1] Skewness measures symmetry of a distribution with 0 meaning equal distribution on both sides.
[2] Kurtosis describes the ‘shape’ of a distribution in terms of heavy tailed or light tailed relative to a normal distribution.

Data Preparation

Conversion to Non-Negative Values

The first step in the data preparation process was to convert all variable (except for LabelAppeal) values to the absolute values. This would ensure that no non-negative values would exist. For this process, two new types of variables were created. A variable starting with "n_" was created to act as a negative flag indicator with a default value of 0. Essentially, if the variable value was negative and then converted to an absolute value, the flag value was changed from 0 to 1. The negative value could be predictive to the overall model. Note that a negative flag indicator was not created for the variable LabelAppeal as this variable was deemed to predictive by itself. The next variable created started with the syntax "imp_". This new variable held the absolute value of the original value. This enables the analysis to maintain data integrity for the original values. Finally, the variable names following the aforementioned prefixes were also shortened for brevity and simplicity during the coding process (see Table 5).

Table 5: Summary of Variable Names

Original Variable	New Variable Name	Missing Flag Indicator Variable Name	Negative Flag Indicator Variable Name
FixedAcidity	imp_fa	m_fa	n_fa
VolatileAcidity	imp_va	m_va	n_va
CitricAcid	imp_ca	m_ca	n_ca
ResidualSugar	imp_rs	m_rs	n_rs
Chlorides	imp_chlo	m_chlo	n_chlo
FreeSulfurDioxide	imp_fsd	m_fsd	n_fsd
TotalSulfurDioxide	imp_tsd	m_tsd	n_tsd
Density	imp_density	m_density	n_density
pH	imp_ph	m_ph	n_ph
Sulphates	imp_sulf	m_sulf	n_sulf
Alcohol	imp_alcohol	m_alcohol	n_alcohol
LabelAppeal	imp_la	m_la	NOT CREATED
AcidIndex	imp_ai	m_ai	n_ai
STARS	imp_stars	m_stars	n_stars

Table 6 highlights the new descriptive statistics using the new non-negative values. Note that at this point, no missing data have been imputed. Of particular interest is the "Delta Mean" column. This column describes the difference between the Mean in Table 6 and Table 2 (descriptive statistics of the original values).

Table 6: Descriptive Statistics for New Predictor Variables

Variable	Minimum	Maximum	Median	Mean	1st Pctl	99th Pctl	Delta Mean
imp_fa	0	34.4	7	8.0632513	0.2	24.4	0.9875342
imp_va	0	3.68	0.41	0.6410856	0.03	2.65	0.3169817
imp_ca	0	3.86	0.44	0.686315	0.01	2.86	0.3779023
imp_rs	0	141.15	12.9	23.3678093	0.9	112.7	17.9490762
imp_chlo	0	1.351	0.098	0.2225586	0.011	1.065	0.1677361
imp_fsd	0	623	56	106.6790418	3	503	75.8334705
imp_tsd	0	1057	154	204.31912	9	771	83.6048874
imp_density	0.88809	1.09924	0.99449	0.9942027	0.9168	1.06981	0
imp_ph	0.48	6.13	3.2	3.2076282	1.32	5.125	0
imp_sulf	0	4.24	0.59	0.8466681	0.03	3.16	0.3195563
imp_alcohol	0	26.5	10.4	10.5237775	1.6	20.3	0.0345412
imp_la	-2	2	0	-0.009066	-2	2	0
imp_ai	4	17	8	7.7727237	6	13	0
imp_stars	1	4	2	2.041755	1	4	0

Note the substantial difference in means for the variables imp_fsd (related to FreeSulfurDioxide) and imp_tsd (related to TotalSulfurDioxide). This implies that the change in using the absolute value has had a major impact on the overall understanding of the variables. Table 7 highlights the skewness and kurtosis of the new variables using the absolute values rather than negative values.

Table 7: Skewness, Kurtosis, and Variance for New Variables

Variable	Skewness	Kurtosis	Variance	Delta skewness	Delta kurtosis	Delta variance
imp_fa	1.174556	1.968638	24.9612014	1.1971473	0.2917844	-14.9514174
imp_va	1.6533658	3.0859297	0.3087071	1.6329811	1.2517781	-0.3059712
imp_ca	1.6431954	2.9499946	0.3672423	1.6935142	1.1101104	-0.3759393
imp_rs	1.4691618	2.2353592	622.2863083	1.5222978	0.3485982	-516.7336917
imp_chlo	1.4811473	2.1772043	0.0548908	1.4507126	0.3865821	-0.0465306
imp_fsd	1.5301341	2.445134	11686.19	1.5237395	0.6065905	-10429.83
imp_tsd	1.6112749	3.0384944	26607.12	1.618456	1.3618687	-27176.62
imp_density	-0.0186981	1.9019373	0.000704247	0	0	0
imp_ph	0.0442987	1.6481659	0.4619745	0	0	0
imp_sulf	1.6918105	3.2105619	0.429827	1.6858971	1.4559007	-0.439038
imp_alcohol	0.1825913	1.0607644	13.170759	0.2133147	-0.4806071	-0.7258758
imp_la	0.0084314	-0.2614968	0.79404	0	0	0
imp_ai	1.6488825	5.1938712	1.752781	0	0	0
imp_stars	0.4473775	-0.6917759	0.8145785	0	0	0

The last three columns in Table 7 describe the change in skewness, kurtosis, and variance from the original variable values (compared to the values in Table 4). Through the use of the absolute values, skewness and kurtosis for almost all of the variables has changed. As an example, the kurtosis for alcohol has been reduced. In terms of variance, the change in values has been somewhat impressive. The variance for imp_rs (ResidualSugar) has been greatly reduced as well as for both imp_fsd (FreeSulfurDioxide) and imp_tsd (TotalSulfurDioxide). Nevertheless, the variance for both imp_fsd and imp_tsd is quite high compared to the other variables.

Imputation

Recall from Table 2 that there are eight predictor variables that have missing values. For the imputation process, the means of each the variables (except for imp_stars) after converting to non-negative values was used. A decision tree was investigated, but it was discovered that the decision trees for each of the eight different variables did not produce satisfying results.

JMP was used to construct decision trees for seven of the eight predictor variables. Table 8 highlights the R-square value obtained after five splits were constructed for each of the seven predictor variables.

Table 8: R-Square Value of Decision Tree Splits for Variables Needing Imputation

Variable	R^2 Value
imp_alcohol	0.021
imp_chlo	0.006
imp_fsd	0.005
imp_tsd	0.011
imp_ph	0.008
imp_sulf	0.006
imp_rs	0.004

The obtained R^2 values for these seven variables is unacceptably low. Furthermore, the obtained trees are needlessly complex and suspected of doing providing little value to the overall imputation process. Figure 2 highlights the decision tree for the variable imp_alcohol.

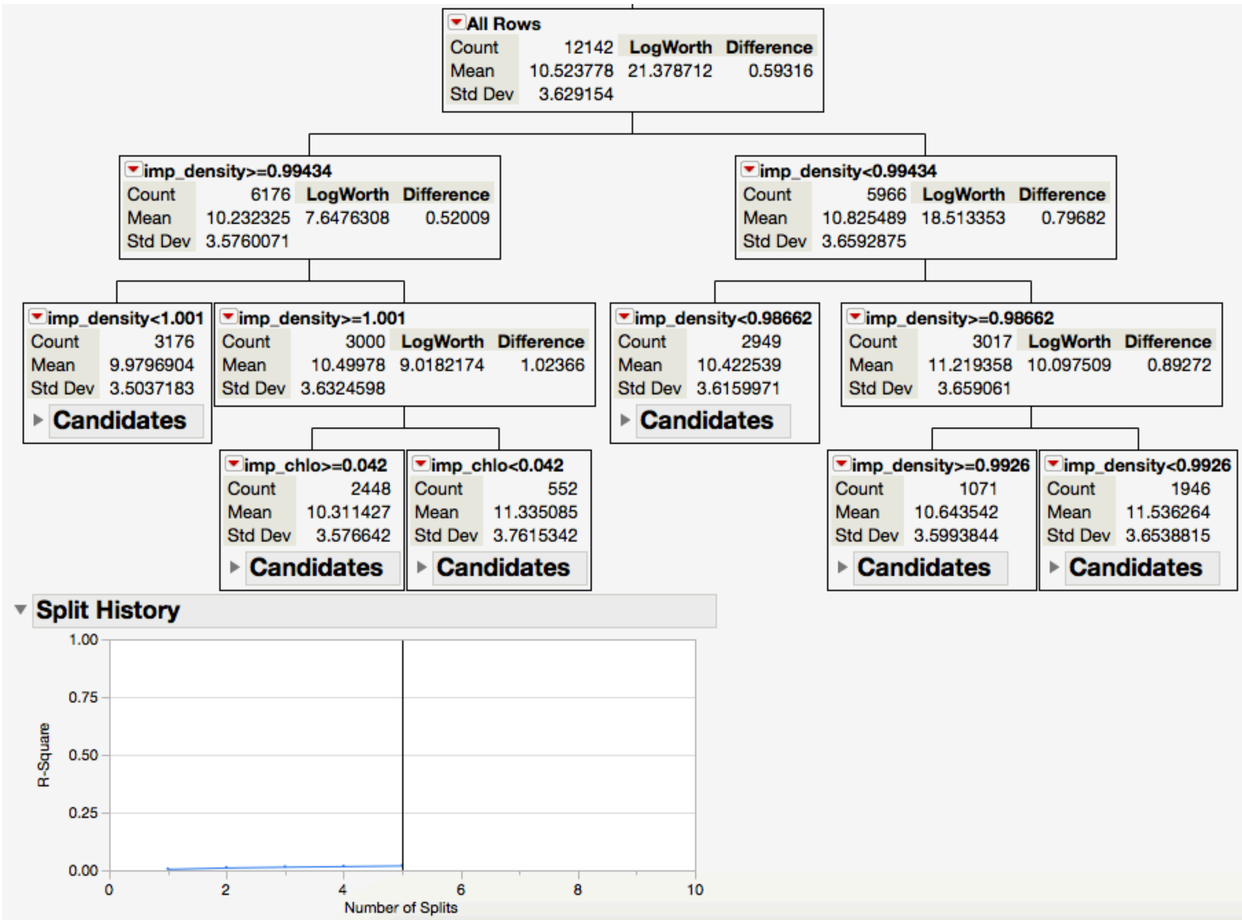

Figure 2: Example Decision Tree Split for imp_alcohol

In this tree, the imputed values for imp_alcohol tend to be between 10 and 11 (approximately). Recall from Table 6, that the mean for imp_alcohol is 10.5. Therefore, it is much simpler and cost effective to utilize the mean of the variables rather than a decision tree. Table 9 provides an overview of the imputation method and the imputed value for all of the eight predictor variables with missing values.

Table 9: Imputation Method & Value Summary

Variable	Method of Imputation	Imputed Value
imp_alcohol	mean	10.524
imp_chlo	mean	0.222
imp_fsd	mean	106.679
imp_tsd	mean	204.319
imp_ph	mean	3.207
imp_sulf	mean	0.847
imp_rs	mean	23.368
imp_stars	defaulting to 0	0

For the variable imp_stars, a value of 0 was used to impute missing values. The thought behind this step is that the lack of stars may be predictive and therefore indicative of the net sales. In an effort to reduce errors when using new data points after a model has been constructed, a safety check was built into the coding. This safety check imputes the mean for all variables (except imp_stars). As an example, it was found that the variable imp_density did not have any missing values. However, the mean of imp_density (according to Table 6) is approximately 0.994. This value has been used as the imputed value if and only if a value is missing for imp_density. In terms of the training dataset, this variable did not have any missing values. However, future data may have missing values for this variable.

Finally, a missing flag variable was created (annotated with the prefix "m_") with a default value of 0. If there was a missing value, then the flag would change to 1 indicating true. The thought here is that the missing value in itself may be predictive.

Model Development

Six types of models were constructed: Regression, Poisson, Negative Binomial, Zero Inflation Poisson, Zero Inflation Negative Binomial, and Logistic Hurdle. For each type, multiple models were constructed utilizing a variety of parameters. Ultimately, a routine was written to calculate both the Average Error and the Root Mean Square Error (RMSE). The chosen model was based primarily on the derived RMSE in combination with parameter value intuition. For the sake of brevity, only one model for each type will be discussed in this analysis.

Model 1 – Regression

For this model, a multiple linear regression model was constructed. This model utilized the stepwise variable selection method. 41 various predictor variables were fed into the model. These included 14 missing flag indicators (one for each predictor variable, with the prefix "m_"), 13 negative flag indicators (one for each predictor variable excluding LabelAppeal, with the prefix "n_"), and 14 imputed variables (imputations of the original variables; containing the prefix "imp_"). The regression equation for this model can be constructed as:

$$TARGET = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

β_0 is the intercept, β_n is the parameter estimate of the variable number *n*, X_n is the observed value in the dataset for the variable number *n*, and ε is the standard error term.

When executing the regression modeling with stepwise variable selection, only 18 of the 41 predictor variables were selected. Table 10 provides the appropriate parameter values for this model.

Table 10: Parameter Estimate for Model 1 (Regression)

Parameter	Notation	Value
Intercept	β_0	4.40897
m_fsd	β_1	0.09741
m_ph	β_2	-0.09652
m_stars	β_3	-0.68713
n_va	β_4	0.07151
n_fsd	β_5	-0.05997
n_alcohol	β_6	0.07468
imp_va	β_7	-0.11876
imp_ca	β_8	0.03167
imp_chlo	β_9	-0.08509
imp_fsd	β_{10}	0.00024562
imp_tsd	β_{11}	0.00025049
imp_density	β_{12}	-0.8292
imp_ph	β_{13}	-0.03095
imp_sulf	β_{14}	-0.03383
imp_alcohol	β_{15}	0.01398
imp_la	β_{16}	0.46529
imp_ai	β_{17}	-0.20255
imp_stars	β_{18}	0.78051

Several key observations can be made from this output. For instance, notice how the coefficient for imp_fsd (FreeSulfurDioxide) and imp_tsd (TotalSulfurDioxide) is almost 0. It also seems that as the density of the wine increases, the more of a negative impact on overall cases sold. This is also the case with the variable imp_ai (AcidIndex). This suggests that as the wine's AcidIndex increases, the relationship to cases sold is inverse. Unsurprisingly, the more stars (imp_stars) a wine has, the more of a positive impact to the cases sold. This implies that the higher a wine is rated, the more likely it will be sold in larger quantities. Perhaps a surprising discovery is the positive value associated with the negative flag variable n_alcohol. Could it be that a negative alcohol reported value has an impact on the sale of wine cases? The variable imp_alcohol suggests that there is a very slight positive impact to the overall wine sample case sales. Finally, the pH (imp_ph) coefficient of -0.031 suggests that as a wine becomes more basic, the lower the overall sales. Recall that the pH scale goes from 0 to 14 with a pH value of 7 being neutral, greater than 7 being basic, and a value of less than 7 being acidic. This seems to be in contrast to the AcidIndex (imp_ai) variable. Further clarification needs to be sought for the metric AcidIndex.

The variables chosen in this model (outlined in Table 10) will serve as the basis for all the subsequent models in this analysis. The stepwise variable selection method identified statistically significant variables and this is an acceptable starting point for the other types of models.

Model 2 – Poisson

The next model constructed was a Poisson regression model. Recall from prior discussion that the goal of this analysis is to predict how many sample cases of wine a customer may purchase. Figure 3 illustrates the histogram (generated using R) of the response variable, TARGET.

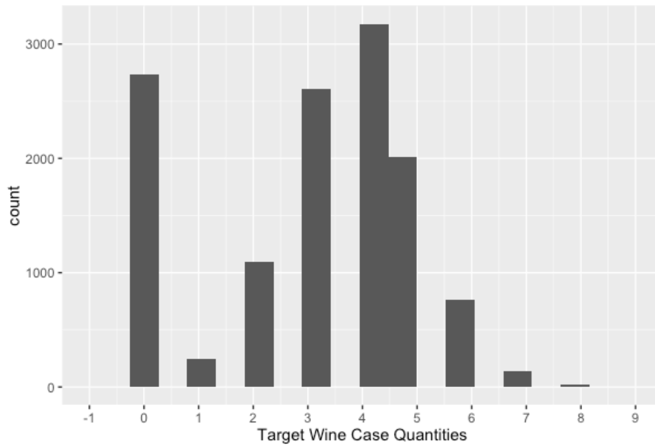

Figure 3: Histogram of TARGET

Note how there's a major spike at a value of 0. This clearly suggests that a zero-inflated model may be more appropriate. However, for illustrative purposes, a Poisson model is continued with. Furthermore, note how the distribution is approximately normally distributed (if the spike at 0 is excluded).

One of the key characteristics of a Poisson model is the need to predict a target variable that is a positive integer. The response variable, in this analysis, has a minimum value of 0 and can be any positive integer value. Furthermore, in a Poisson distribution, the mean and variance are also the same (not necessarily the same, but approximately close). The mean for the TARGET variable is approximately 3.03 and the variance is approximately 3.71. The general equation for a Poisson regression is like the following:

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \cdots \beta_n X_n + \varepsilon$$
$$Y = e^{\ln(Y)}$$

The first step essentially a natural log of the count and the second step will convert the natural log to a meaningful value. β_0 is the intercept, β_n is the parameter estimate of the variable number *n*, X_n is the observed value in the dataset for the variable number *n*, and ε is the standard error term.

Recall from the Regression Model that only 18 predictor variables were selected. Since there was not an easy way to do automatic variable selection with Poisson regression, only these 18 predictor variables were used to model the response variable TARGET. Table 11 provides the appropriate parameter values for this model.

Table 11: Parameter Estimates for Model 2 (Poisson)

Parameter	Notation	Value
Intercept	β_0	1.7904
m_fsd	β_1	0.0313
m_ph	β_2	-0.038
m_stars	β_3	-0.6478
n_va	β_4	0.0232
n_fsd	β_5	-0.0216
n_alcohol	β_6	0.0215
imp_va	β_7	-0.0412
imp_ca	β_8	0.0094
imp_chlo	β_9	-0.0301
imp_fsd	β_{10}	0.0001
imp_tsd	β_{11}	0.0001
imp_density	β_{12}	-0.2889
imp_ph	β_{13}	-0.013
imp_sulf	β_{14}	-0.0129
imp_alcohol	β_{15}	0.004
imp_la	β_{16}	0.1587
imp_ai	β_{17}	-0.0813
imp_stars	β_{18}	0.1883

Similar conclusions to the ones made for the Regression model can be drawn upon again. Of interest is the continued predictability drawn from the variable m_stars. Recalls that this variable has a binary (has a value of 0 meaning false or 1 meaning true) value indicating if the original data point is missing or not. Missing stars for a wine continue to have a negative impact on the overall unit sales. Note how the density of a wine (imp_density) continues to have a negative impact. A data point that would assist this analysis is the type of wine. From the original dataset, it is not clear if a wine is red, white, port, Riesling, etc. This would enable a clearer understanding of what kind of wine has what density.

Chlorides (imp_chlo) continue to have a negative impact on the overall sales, but this is where a subject matter expert could assist the findings to ensure that the negative impact makes sense. Surprisingly, the sulfur dioxide variables (imp_fsd and imp_tsd) seem to have negligible impact on the overall sales. Another surprising note is the minimal impact of the alcohol (imp_alcohol). This is another aspect that could be explored further with a subject matter expert.

Model 3 – Negative Binomial

Recall that the mean and variance for the response variable, TARGET, are approximately 3.03 and 3.71 respectively. Since the variance is larger, a negative binomial model may be more appropriate[3]. Conceptually, the Poisson regression is a special scenario of the negative binomial, thus, the underlying regression equation structure remains the same as the Poisson regression.

[3] Recall from Figure 3 that there is a large spike at a value of 0, suggesting that a zero-inflation model may be more appropriate.

Using the 18 predictor variables selected from the Regression Model, a negative binomial model was constructed. However, the results obtained were exactly the same as found in Table 11. In order to create some contrast, only the following variables were kept in the revised negative binomial model: m_stars, imp_va, imp_alcohol, imp_la, imp_ai, and imp_stars. This list was obtained by removing any variable from the Poisson regression whose 95% Wald Confidence Intervals contained the value of 0. The reasoning is that if the parameter estimate could be 0, then that variable could be potentially be meaningless in the overall model. Table 12 provides the appropriate parameter values for this model.

Table 12: Parameter Estimates for Model 3 (Negative Binomial)

Parameter	Notation	Value
Intercept	β_0	1.4831
m_stars	β_1	-0.6511
imp_va	β_2	-0.0406
imp_alcohol	β_3	0.0037
imp_la	β_4	0.1587
imp_ai	β_5	-0.0815
imp_stars	β_6	0.1884

This significantly reduced model does not contain significant surprises. It continues to highlight the fact that missing stars have a negative impact whereas stars in general improve the likelihood of wine sales. As seen in earlier models, the variable imp_alcohol continues to have negligible impact.

Model 4 – Zero Inflated Poisson (ZIP)

The ZIP (Zero Inflated Poisson) model is quite similar in its assumptions as the Poisson model. The major difference is that this model assumes a spike at a value of 0. Note how Figure 3 is a clear example of where a ZIP model makes more sense than a Poisson or Negative Binomial model. When developing ZIP model, two different scenarios are created: a model with non-zero cases and a model with only zero cases (i.e., a logistic model). These two models are then used to put the overall model into a production-ready model.

First, a similar approach to what was done in the Poisson Model (Model 3) is used:

$$\ln(Y1) = \beta_0 + \beta_1 X_1 + \cdots \beta_n X_n + \varepsilon$$
$$P\_Target\_All = e^{\ln(Y1)}$$

Second, a logistic model approach is executed for the likelihood of obtaining a 0:

$$Log\ Odds = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

$$P\_Target\_0 = \frac{e^{LogOdds}}{1 + e^{LogOdds}}$$

With the two probabilities obtained, the expected value is calculated using the following equation:

$$P\_Score\_ZIP = P\_Target\_All * (1 - P\_Target\_0)$$

This approach provides a comprehensive derivation of the ZIP model. Utilizing this methodology, Table 13 provides the appropriate parameter values for this model with both components.

Table 13: Parameter Estimates for Model 4 (Zero Inflation Poisson)

Non-Zero			Zero		
Parameter	**Notation**	**Value**	**Parameter**	**Notation**	**Value**
Intercept	β_0	1.4436	Intercept	β_0	-4.1789
m_fsd	β_1	0.0132	imp_va	β_1	0.2205
m_ph	β_2	-0.0097	imp_ca	β_2	-0.1047
m_stars	β_3	0.0369	imp_chlo	β_3	0.0614
n_va	β_4	0.0119	imp_fsd	β_4	-0.0007
n_fsd	β_5	-0.0091	imp_tsd	β_5	-0.0012
n_alcohol	β_6	0.0096	imp_density	β_6	0.275
imp_va	β_7	-0.0147	imp_ph	β_7	0.2087
imp_ca	β_8	-0.0013	imp_sulf	β_8	0.196
imp_chlo	β_9	-0.0172	imp_alcohol	β_9	0.0266
imp_fsd	β_{10}	0	imp_la	β_{10}	0.7226
imp_tsd	β_{11}	0	imp_ai	β_{11}	0.4404
imp_density	β_{12}	-0.2903	imp_stars	β_{12}	-2.3945
imp_ph	β_{13}	0.0052			
imp_sulf	β_{14}	0.0058			
imp_alcohol	β_{15}	0.0072			
imp_la	β_{16}	0.2323			
imp_ai	β_{17}	-0.0189			
imp_stars	β_{18}	0.107			

This model is quite different from the outputs of the other models. Immediately note how the variables imp_fsd and imp_tsd have a coefficient value of 0 in the Non-Zero component. Similarly, in the Zero component, these two variables have a negligible impact. However, there is a telling story when exploring the variable imp_density. It appears that as the density increases, the likelihood of no wine sales increases (see the Zero component). This contrasts with the imp_density parameter value in the Non-Zero component (-0.2903). In this one, as the density increases, the lower the likelihood of selling wine cases. Similarly, the variable imp_stars implies that if a wine as no stars, there is a larger likelihood that no sales will occur – a stark contrast to if a wine has stars. Another interesting note is how the variable imp_ph has a negligible impact on the Non-Zero component, but a large positive impact on the Zero component (0.0052 vs. 0.2087 respectively). From this, it could be concluded that the pH of a wine plays a stronger role in having 0 cases sold versus any non-zero cases sold. Finally, it is interesting to note that the variable m_stars has a positive value. This suggests that if a star is missing, it may contribute in a positive way to improving sales. Intuitively, this could be indicative that certain customers are willing to try out new wines that have not yet been rated to assess their viability for their customers.

Perhaps the most interesting aspect is how the Zero component did not have any parameters for the different flags (negative and missing flags). This may be explored further, but is outside the scope of this analysis.

Model 5 – Zero Inflated Negative Binomial (ZINB)

The final model constructed was the Zero Inflated Negative Binomial (ZINB) model. This model is very similar to the ZIP model (see section Model 4). As seen in Model 4, the ZINB model will also output two different components: a Zero component and a Non-Zero component. All of the derivations are similar to Model 4. Table 14 provides the appropriate parameter values for this model with both components.

Table 14: Parameter Estimates for Model 5 (Zero Inflation Negative Binomial)

Non-Zero			**Zero**		
Parameter	**Notation**	**Value**	**Parameter**	**Notation**	**Value**
Intercept	β_0	1.4492	Intercept	β_0	-4.01
m_fsd	β_1	0.0126	imp_va	β_1	0.2105
m_ph	β_2	-0.0095	imp_ca	β_2	-0.1033
m_stars	β_3	0.0313	imp_chlo	β_3	0.0639
n_va	β_4	0.0118	imp_fsd	β_4	-0.0006
n_fsd	β_5	-0.0088	imp_tsd	β_5	-0.0011
n_alcohol	β_6	0.0093	imp_density	β_6	0.2318
imp_va	β_7	-0.0145	imp_ph	β_7	0.2012
imp_ca	β_8	-0.0014	imp_sulf	β_8	0.1883
imp_chlo	β_9	-0.0173	imp_alcohol	β_9	0.0258
imp_fsd	β_{10}	0	imp_la	β_{10}	0.6926
imp_tsd	β_{11}	0	imp_ai	β_{11}	0.4248
imp_density	β_{12}	-0.2931	imp_stars	β_{12}	-2.2366
imp_ph	β_{13}	0.0053			
imp_sulf	β_{14}	0.0057			
imp_alcohol	β_{15}	0.0072			
imp_la	β_{16}	0.2319			
imp_ai	β_{17}	-0.0186			
imp_stars	β_{18}	0.1053			

As previously mentioned, the ZINB model is quite similar to the ZIP model. Note how the ZINB model also has a parameter estimate of 0 for the variables imp_fsd and imp_tsd. There are times where a ZINB model may converge to the same value as a ZIP model, but such is not the case. The parameter estimates are quite close to each other (when comparing ZIP and ZINB), but they are not exact. Many of the observations made for Model 4 can be applied to Model 5 as well.

Model 6 – Logistic Hurdle

An optional model that was explored is the Logistic Hurdle model. This model is somewhat similar to the ZIP and ZINB models. The Logistic Hurdle model also has two components: a component to predict if the count is zero (or non-zero) and a component to predict the count (assuming the count is non-zero). The end result is obtained by multiplying the probability of non-zero and the count to obtain the expected count.

In order to develop this model, two variables were created. The first variable TARGET_FLAG was created to indicate a 1 if the response variable TARGET was greater than 0. Otherwise, TARGET_FLAG would be set to 0. The second variable TARGET_AMT was shifted down by one unit. The intent here was to eliminate the need for modeling sales of 0 units. TARGET_AMT was adjusted to null if the variable TARGET_FLAG was zero. This would ensure that the zero sales counts would be negated. In the final scoring process, the shift of one unit was reversed (1 was added back into the overall model) to ensure that the dataset was not shifted needlessly.

Table 15 highlights the parameter estimates for the probability of obtaining a non-zero count (Non-Zero) and the predicted count (Count).

Table 15: Parameter Estimates for Logistic Hurdle Model

Non-Zero				Count			
Parameter	**Sub-Parameter**	**Notation**	**Value**	**Parameter**	**Sub-Parameter**	**Notation**	**Value**
Intercept		β_0	20.1126	Intercept		β_0	1.7889
n_va		β_1	0.1578	m_stars		β_1	-0.4372
imp_va		β_2	-0.2021	imp_alcohol		β_2	0.0096
imp_tsd		β_3	0.000949	imp_la	-2	β_3	-1.4595
imp_ph		β_4	-0.1836	imp_la	-1	β_4	-0.8074
imp_sulf		β_5	-0.1568	imp_la	0	β_5	-0.4331
imp_alcohol		β_6	-0.021	imp_la	1	β_6	-0.192
imp_la	-2	β_7	1.8192	imp_la	2	β_7	0
imp_la	-1	β_8	1.3239	imp_ai		β_8	-0.0207
imp_la	0	β_9	0.9137	imp_stars	0	β_9	0
imp_la	1	β_{10}	0.3668	imp_stars	1	β_{10}	-0.3753
imp_ai		β_{11}	-0.3955	imp_stars	2	β_{11}	-0.2342
imp_stars	0	β_{12}	-17.5075	imp_stars	3	β_{12}	-0.1244
imp_stars	1	β_{13}	-15.6792	imp_stars	4	β_{13}	0
imp_stars	2	β_{14}	-13.2538				
imp_stars	3	β_{15}	-0.1485				

For the Predicted Count component, statistically insignificant variables were removed prior to the modeling process. This component made use of the PROC GENMOD function that does not have the ability for automatic variable selection. In contrast, the Non-Zero component used the PROC LOGISTIC function with the stepwise automatic variable selection method. In both instances, the variables imp_la (LabelAppeal) and imp_stars (STARS) were categorized as class variables (i.e., categorical variables).

Note how, in the Non-Zero component, the significant impact that 0, 1, or 2 stars have on the overall predicted probability of non-zero sales. Surprisingly, the label appeal (imp_la) coefficients have a positive value. This is somewhat counter intuitive as one would think that the increased label appeal would have more non-zero sales and negative label appeal would have a detrimental impact on the sales.

Model Selection

For each model, a model validation algorithm was developed. This algorithm provided the average error and root mean square error. Furthermore, for each model, the AIC and BIC values (generated through the appropriate SAS functions) were also recorded. Table 16 summarizes these findings.

Table 16: Model Validation Summary

	AIC	BIC	Avg. Error	RMSE
Model 1 - Regression	6951.07	6953.14	1.56201	1.92732
Model 2 - Poisson	45757.4834	45899.1628	1.03433	1.31595
Model 3 - NB	45762.9136	45815.1113	1.03285	1.31651
Model 4 - ZIP	40899.1174	41137.7353	0.96187	1.27407
Model 5 - ZINB	40965.0954	41211.1701	0.96597	1.27308
Model 6 – Logistic Hurdle	n/a	n/a	0.96339	1.26456

From an AIC or BIC perspective, the best performing model is the regression model (Model 1). However, this model is not the most intuitive as it does a poor job of modeling for the spike in zeros that the response variable has (see Figure 3). For this analysis, the intent of the regression model was to have the model identify which variables are statistically significant using the stepwise automatic variable selection method. In this event, 18 of the 41 predictor variables were identified. These 18 variables were then used for the subsequent models.

The metric of choice here is to use the RMSE metric. This metric will sufficiently enable an understanding of the validity of each model. Note that the lower the RMSE value, the 'better' a model may be. Model 3 (negative binomial) was not be selected since it was heavily modified from Model 2. Recall that the original version of Model 3 produced the exact same values as Model 2. Nevertheless, the RMSE value for the modified version of Model 3 is surprisingly similar to Model 2 while using a substantially smaller set of predictor variables.

The ZINB (Zero Inflation Negative Binomial) model has the second lowest RMSE value (1.27308) followed by the ZIP (Zero Inflation Poisson) model (1.27407). The model with the lowest RMSE is the Logistic Hurdle model (Model 6). Although this model performs the best in terms of RMSE, it has not been chosen due to the lack of simplicity. Therefore, the chosen model is the ZINB (Model 5) model. It's parameter estimates provide an intuitive understanding of the variables' relationship to the response variable TARGET.

Conclusion

Several models of each type (Regression, Negative Binomial, etc.) were constructed to predict the amount of sample cases a wine distributor may purchase. The original dataset contained 14 predictor variables and over 12,000 observations. Most of the variables captured the chemical composition of a wine (e.g., pH, free sulfur dioxide, alcohol, etc.). A few of the variables offered insight into general characterizations of the wine such as the star rating, acid index, or the label appeal. One of the key challenges in this data was the inability to classify the type of wine investigated. There was no variable that clearly identified each wine as being red, white, etc. Surprisingly, many of the variables are numeric, but could considered categorical such as stars. For this analysis, all variables were explored as numeric variables and not categorical variables. Many of the data points were imputed if missing or adjusted if negative. A subject matter expert in wine composition should be consulted to understand the applicability of the signs in each parameter estimate as well as support the assumption that no chemical value could be negative. This will require further investigation which is outside the scope of this analysis. Finally, it would be prudent to monitor this chosen model's performance for the next three months as new data are fed into it. Next steps may include (but not limited to) refining the model and focusing on improved data imputation.

Bingo Bonus – Using a Decision Tree to Predict Target Cases

For this bingo bonus, I used JMP to create a decision tree to predict the number of wine cases sold. This process required creating the necessary flags (both missing and negative) as well as data imputation. Note that the steps discussed in this analysis were applied directly to JMP. Approximately four splits were created for an R^2 value of 0.514. Figure 4 illustrates the decision tree that was created and the appropriate steps.

Figure 4: JMP Decision Tree for TARGET

Loosely put, the logic can be described as follows:

```
If the imp_stars is less than 2 and m_stars is greater than or equal to 1
then the predicted target value is 1.18. Otherwise it is 2.58. If the
imp_stars is greater than or equal to 2 and if imp_la is less than 1 and if
imp_la < 0 then the predicted target value is 3.045. If the imp_la value is
greater than or equal to 0, then the predicted target value is 4.034. If the
imp_la value is greater than or equal to 1 then the predicted target value is
5.06.
```

Table 17 summarizes the model validation values (Avg. Error and RMSE) for the previous six models in addition to this new model using a decision tree.

Table 17: Model Validation Summary with Decision Tree Model

	AIC	BIC	Avg. Error	RMSE
Model 1 - Regression	6951.07	6953.14	1.56201	1.92732
Model 2 - Poisson	45757.4834	45899.1628	1.03433	1.31595
Model 3 - NB	45762.9136	45815.1113	1.03285	1.31651
Model 4 - ZIP	40899.1174	41137.7353	0.96187	1.27407
Model 5 - ZINB	40965.0954	41211.1701	0.96597	1.27308
Model 6 – Logistic Hurdle	n/a	n/a	0.96339	1.26456
Model 7 – Decision Tree	n/a	n/a	1.03202	1.34251

The RMSE for this decision tree model (Model 7) is worse than the RMSE for the Poisson model (Model 2). Note how the prior models were using several variables to predict target cases. However, the decision tree relied on only three distinct variables: imp_stars, m_stars, imp_la. It is possible that more splits could have been made to make use of more predictor variables. However, as can be seen in Figure 4, over 10 splits were made and only incremental changes were seen in the R^2 value. In an effort to find a parsimonious model, fewer splits were kept. Needless to say, this model would not have been chosen as compared to the prior models.