Wine Sales Predictive Model Report

Kaggle Name: Rahul Sangole

Nov 20 2016

PREDICT-411 Section 55

Total Points Claimed:

- Write up	150
- Standalone scoring program	50
- Scored data file	50
- Hurdle model	20
- Decision trees using Angoss & SAS JMP	20
- Usage of SAS macros	10
	Total 300

Table of Contents

Abstract

This document reports the analysis performed to predict the number of cases of commercially available wine purchased by wine distribution companies after sampling the wine. Various data preparation steps are studied prior to model development. A total of six models run across three competing data preparation strategies are described, before a final model is proposed based off of common performance metrics.

Wine Sales Predictive Model Report

The data set analyzed consists of the number of cases of commercially available wine purchased by wine distribution companies after sampling the wine. Since these sample cases are used to demo the wine in restaurants and stores, a higher the cases of wine samples purchased is indicative of a positive taste profile of the wine, which is indicative of higher expected sales of the wine.

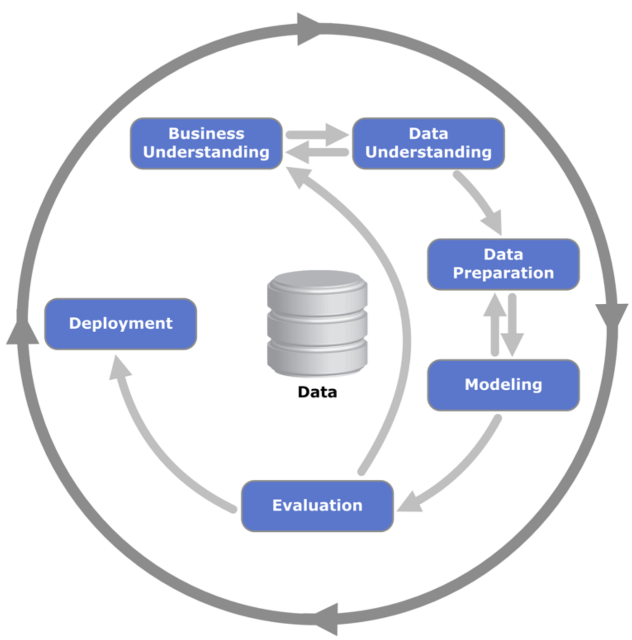
The data set consists of information of 12,795 wines. The response variable of interest - which contains an integer between 0 and 8 – is labeled Target, which is the number of wine cases purchased by the distribution company. There are some missing values in this variable.

The predictor variables available for modeling range from those describing the acidity of the wine, sugar and alcohol contents, other chemical properties and lastly, a marketing score indicating the appeal of the label design for consumers. Some of these variables have missing data, which has been addressed prior to modeling.

Overview of Methodology Used

The CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is followed in this project. The figure below (Chapman, 2000) provides an overview of the life cycle of a data mining and modeling project such as this one in a CRISP-DM framework.

There are six key phases in the life cycle of a data mining project. These phases are not linearly arranged. The two way arrows and outer circle symbolize the cyclic nature of the task of data mining.



The input given to the analyst by the business has been an output of the first step – Business Understanding. This report focuses on the tasks of Data Understanding, Data Preparation, Modeling, and Evaluation. Each of these steps is summarized below.

Data Understanding

This section focuses on developing an intimate understanding of the data provided by the database administrator. Exploratory data analysis in terms of univariate and bivariate statistics, as well as graphical methods are employed. Comments are made on the distributions of predictor and response variables as well as their data quality.

Data Preparation

This section describes the work done for clean-up of the data. This involves addressing potential outliers and influential points, negative values, imputation of missing data. Certain engineered and derived variables are also explored here. Test and training datasets are also created here.

Modeling

This section describes the various models explored in the project – Linear, Poisson, Negative Binomial, Zero Inflated Models and the Hurdle model. The variable selection methods are explored, model diagnostics evaluated and overall model fits assessed.

Evaluation

The models created in previous step are evaluated in this step. Various model evaluation criteria such as AIC, SBC, Mean Error etc are explored. These criteria are balanced against model complexity, business reasoning to select the final model.

Data Understanding

The data set consists of 1 target variable and 16 predictor variables. A total of 12,795 observations in the data set. The details of the data set are shown in the table below.

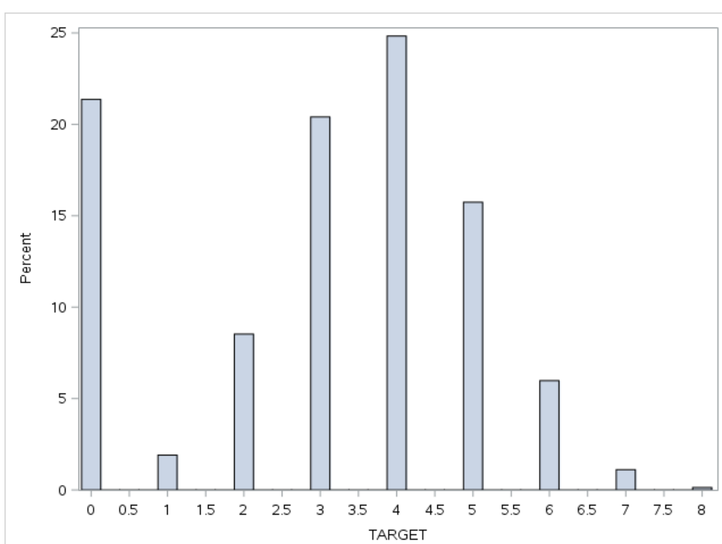
VARIABLE NAME	DEFINITION	PREDICTED EFFECT
Target	Number of Cases Purchased	-
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average	Too high acidity negatively affects the taste
Alcohol	Alcohol Content	-
Chlorides	Chloride content of wine	-
CitricAcid	Citric Acid Content	Too high acidity negatively affects the taste
Density	Density of Wine	-

FixedAcidity	Fixed Acidity of Wine	Too high acidity negatively affects the taste
FreeSulfurDioxide	Sulfur Dioxide content of wine	-
LabelAppeal	Marketing Score indicating the appeal of label design for consumers	Higher visual appeal may suggest higher sales
ResidualSugar	Residual Sugar of wine	-
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor	Higher star rating may suggest higher sales
Sulphates	Sulfate content of wine	-
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	-
VolatileAcidity	Volatile Acid content of wine	Too high acidity negatively affects the taste
pH	pH of wine	Too high acidity negatively affects the taste

The response variables can be divided into Acid-related, Chemical-composition-related, physical-property-related and and rating-related variables. For the most part, the predicted effect of the response variables on the target is unknown. Each variable is explored below.

Target Variable

This variable is a count of how many cases of wine purchased. The histogram shows the discrete values of the variable, which indicate that a Poisson or Negative Binomial model could be used to predict this variable. The histogram also shows 21% of the observations are zero, indicating that no cases were purchased. This indicates that we should also investigate usage of a Zero Inflated Poisson (ZIP) or Zero Inflated Negative Binomial (ZINB) regression model.



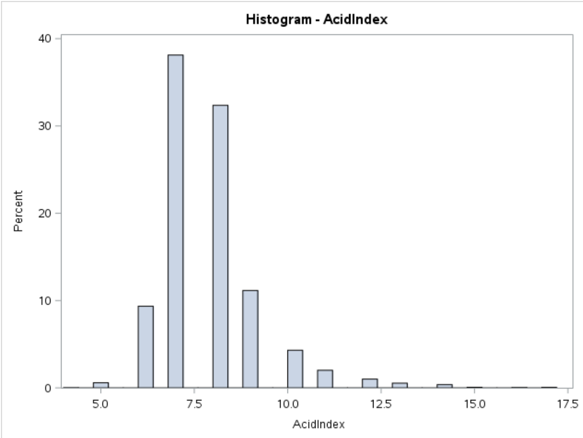
Response Variables

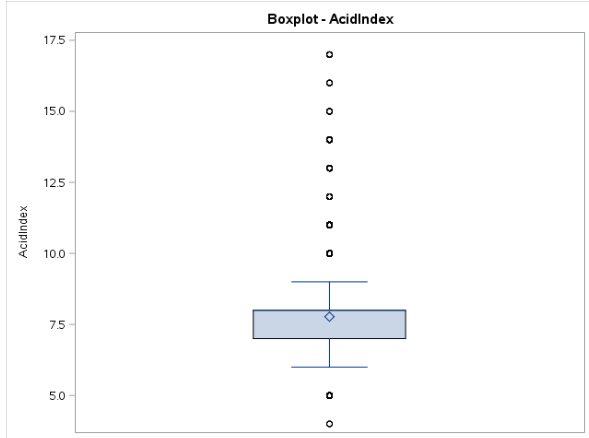
Univariate analysis of the response variables results in two important observations:

1. 8 of the 14 variables have missing data. This has been addressed in the data imputation section.
2. Sulphates, TotalSulfurDioxide, Alcohol, FreeSulfurDioxide, Alcohol, FreeSulfurDioxide, Chlorides, ResidualSugar, CitricAcid, FixedAcidity and VolatileAcidity have negative values. In a physical sense, none of these variables could have negative values. This points to either a data recording error, or some possible data transformation performed after the data was collected. Univariate analysis points to the latter, though we cannot be certain at this point. The strategy adopted in this project is to keep the negative values without transformation if the resulting model is highly predictive.

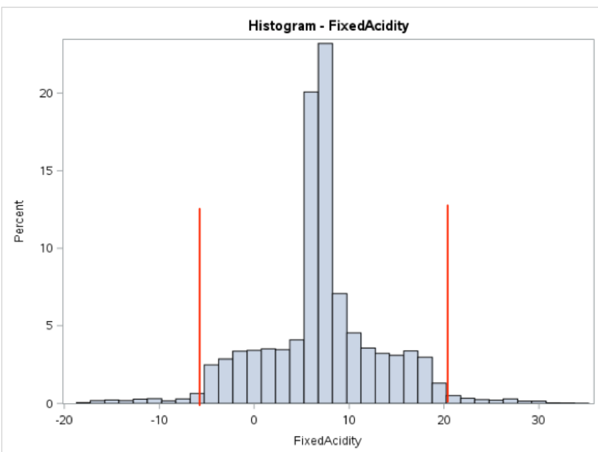
Variable	N Miss	% Missing	Min	Max	Mean	Median
STARS	3359	36%	1	4	2.04	2.00
Sulphates	1210	10%	-3.13	4.24	0.53	0.50
TotalSulfurDioxide	682	6%	-823	1057	120.71	123.00
Alcohol	653	5%	-4.7	26.5	10.49	10.40
FreeSulfurDioxide	647	5%	-555	623	30.85	30.00
Chlorides	638	5%	-1.171	1.351	0.05	0.05
ResidualSugar	616	5%	-127.8	141.15	5.42	3.90
pH	395	3%	0.48	6.13	3.21	3.20
AcidIndex	0	0%	4	17	7.77	8.00
CitricAcid	0	0%	-3.24	3.86	0.31	0.31
Density	0	0%	0.888	1.099	0.99	0.99
FixedAcidity	0	0%	-18.1	34.4	7.08	6.90
LabelAppeal	0	0%	-2	2	-0.01	0.00
VolatileAcidity	0	0%	-2.79	3.68	0.32	0.28

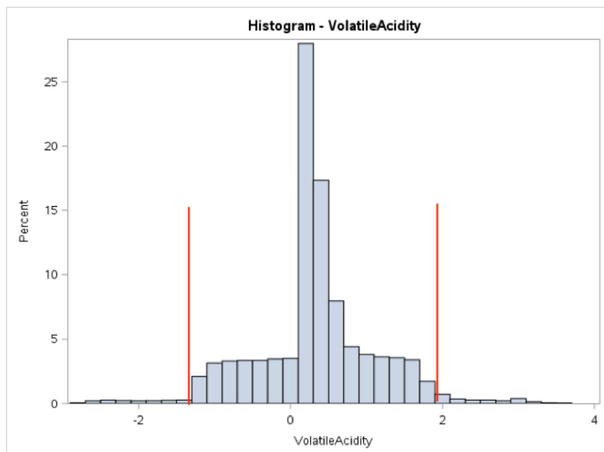
Acid Related Variables. The AcidIndex variable is a right skewed variable with outliers present at either end. These outliers (<5%: 6, >95%: 10) can be curbed using thresholds in the data preparation stage.

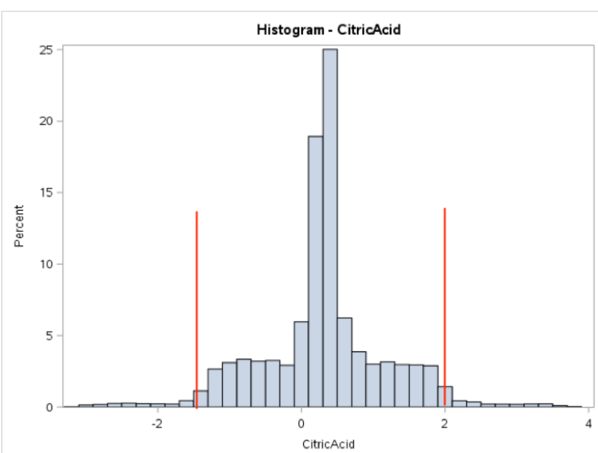


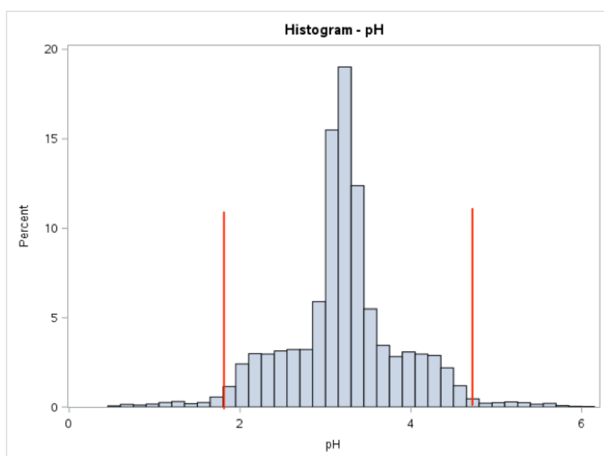


CitricAcid, FixedAcidity, VolatileAcidity, and pH all show a peculiar symmetric distribution of a large central peak, fairly flat mid-section and long tails. The first three variables have negative values, which does not intuitively make sense. CitricAcid and VolatileAcidity have a mean of zero and a standard deviation close to 1 (0.862 and 0.784 respectively). This points towards some type of data standardization performed beforehand.

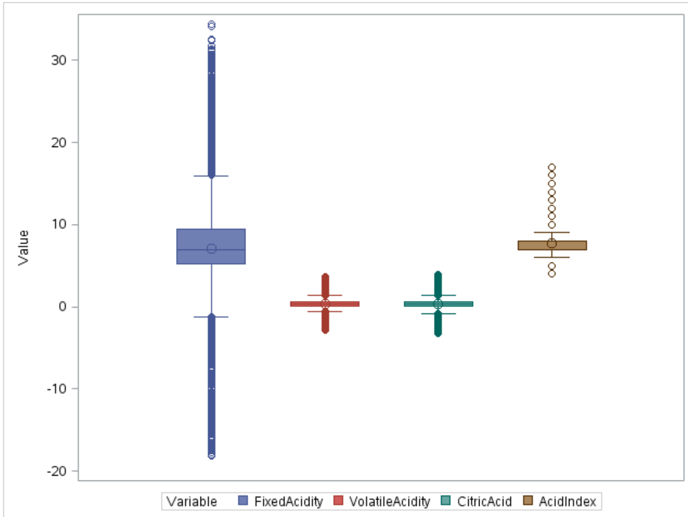




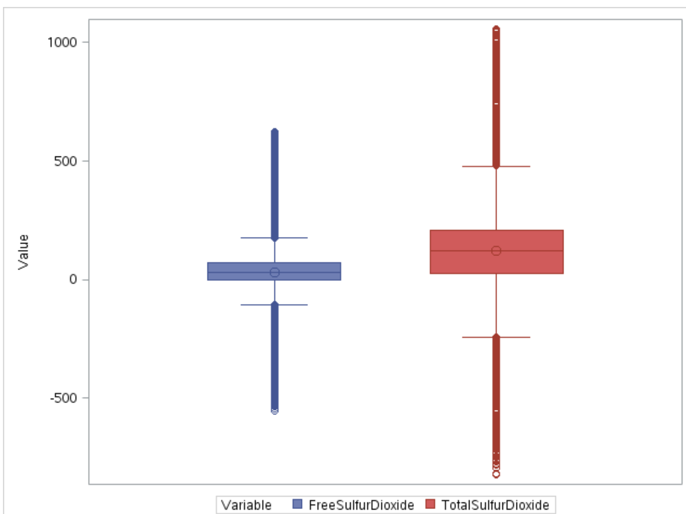




While all four variables show outliers, FixedAcidity has a large number of them.

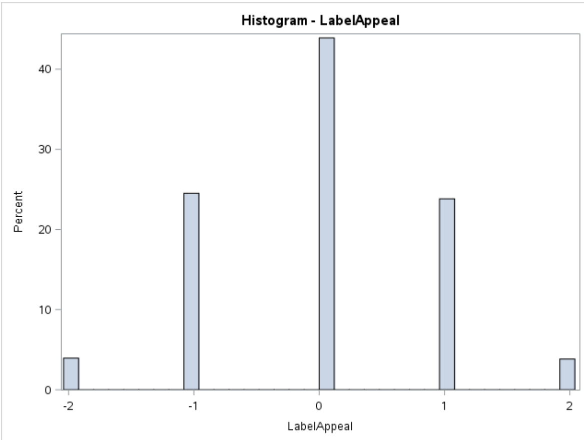


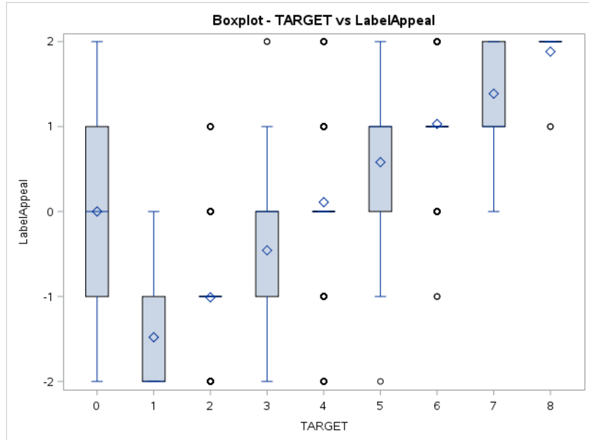
Chemical Composition Related Variables. ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSurfurDioxide and Sulphates have similar distributions as the Acid related variables. TotalSurfurDioxide has higher variability compared to FreeSulfurDioxide.



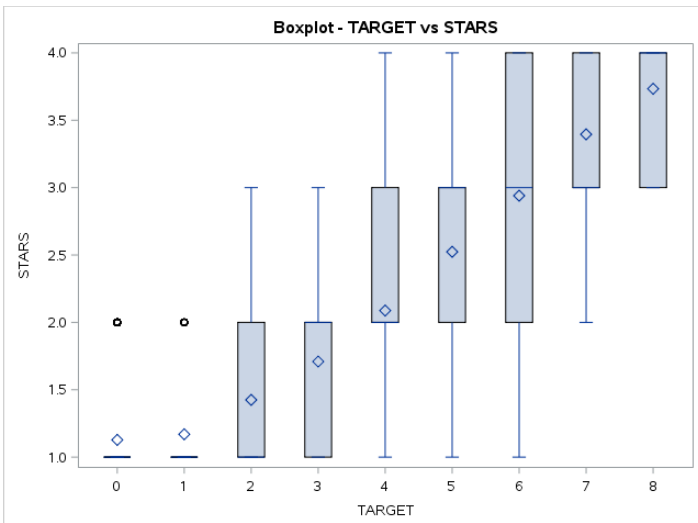
Physical Composition Related Variables. Density follows a similar distribution as before. Alcohol shows a smoother distribution; potential cutoffs seem to be 0 and 20.

Rating Related Variables. LabelAppeal is a categorical variable between -2 and 2, centered around 0. The boxplot of LabelAppeal categorized by Target shows that there is a strong correlation between the label's appeal to the customer and the number of cases ordered. But, for zero cases, the label appeal has no correlation and a high variance.





Number of stars given to the wine varies between 1 and 4, and also shows a strong correlation to the number of cases purchased. This makes intuitive sense since the higher the rating after a taste test, the higher the chances an order will be placed for the wine.



Data Preparation

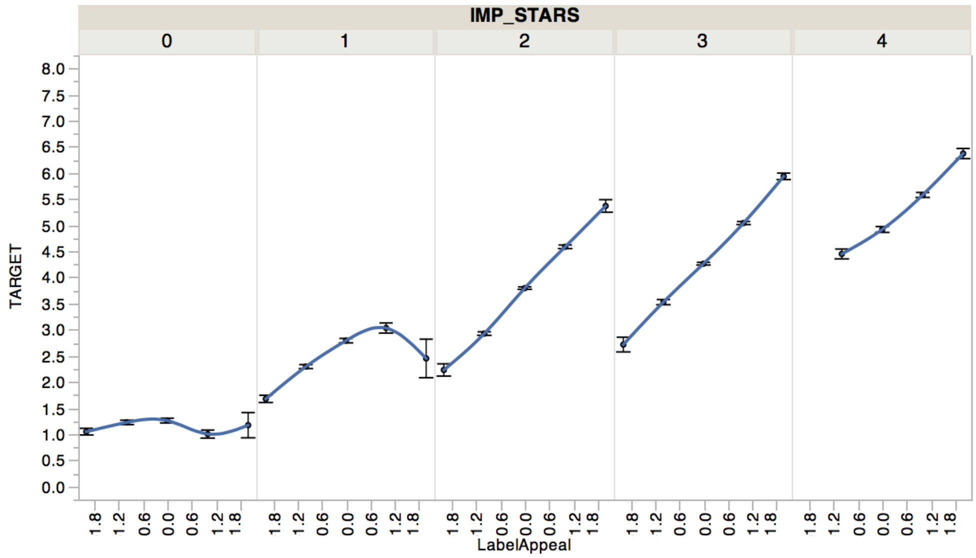
These are the data preparation steps taken prior to modeling:

Missing Value Imputation Variables

STARS. This variable is highly predictive as shown above. If the mean value of the cases sold is observed against the number of stars, we can see when the data is missing, the mean value of cases sold is significantly low.

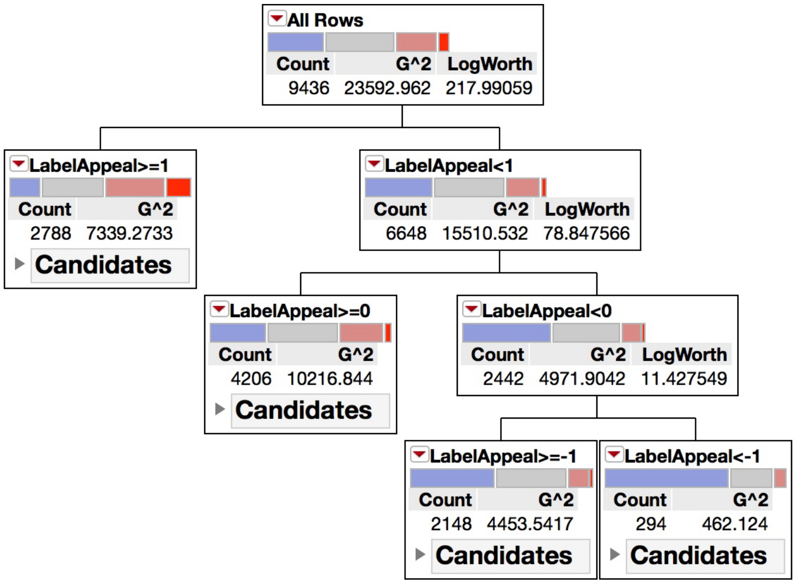
Stars	Missing	1	2	3	4
Mean (Target)	1.184	2.580	3.798	4.544	5.421

As a result, for any observations with missing stars, an imputed value of 0 is used. Post imputation, we can observe the strong predictive relationship of Stars, LabelAppeal and Target. This is best observed through a visual relationship shown below. Each point is the mean value of the response variable, with 1 standard deviation errors bars plotted. For each value of Stars, we can see a clear increase in cases sold as the label appeal increases. Also, as the stars themselves increase, the number of cases sold increases.



Each error bar is constructed using 1 standard error from the mean.

Note. An imputation strategy of using decision trees for the Stars variable was explored, but the trees made negligible difference (0% - 1% difference in performance) in MAE, RMSE, SBC performance of the models. The performance of a tree can be seen by the division of colors in the tree. Except, LabelAppeal < - 1, the splits in stars are uniform amongst 1 and 2 stars. The confusion matrix also shows the training model heavily under predicts stars =3 and stars = 4.



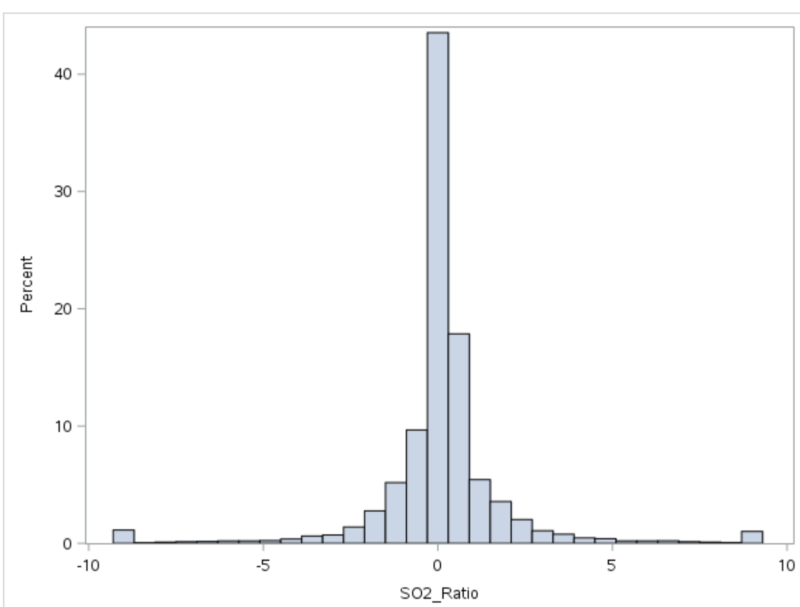
Training

Actual STARS	Predicted Count 1	2	3	4
1	1211	1831	0	0
2	919	2651	0	0
3	283	1929	0	0
4	29	583	0	0

All Other Variables. For all other variables with missing values, two strategies are employed: (i) usage of decision trees, and (ii) imputation with the median value. Decision trees are created using Angoss KnowledgeStudio and SAS JMP Pro 11.0. Both strategies are explored in the ‘Data Sets Created’ section described below.

Derived Variables

One engineered variable is created called SO2_Ratio, which is the ratio of IMP_FreeSulfurDioxide and IMP_TotalSulfurDioxide. It compares how much of the total sulfur dioxide in a free sulfur dioxide form. The variable is winsorized at the 1% and 99% cutoff points of -9 and +9, to curb outliers where the denominator IMP_TotalSulfurDioxide is very large or very small. A histogram of SO2_Ratio is as follows.



Winsorizing

An approach of winsorizing the variable is investigated in data set D3. This involves replacing outliers (<1% and >99%) with the corresponding 1% and 99% cutoff values.

Variable Name	1% Cutoff	99% Cutoff
IMP_Chlorides	-0.848	0.952
Density	0.917	1.070
LabelAppeal	-2	2
VolatileAcidity	-1.865	2.590
CitricAcid	-2.18	2.66
IMP_pH	1.33	5.12
AcidIndex	6	13
IMP_Alcohol	0.2	20.2
FixedAcidity	-10.9	24.4
IMP_ResidualSugar	-89.6	97.1
IMP_FreeSulfurDioxide	-382	464
IMP_TotalSulfurDioxide	-516	746

Data Sets Created, and Training & Test Data Sets

Each model is run on three datasets, defined in the following table. Each data set is prepared with a different approach towards data preparation.

Data Set Name	Decision Tree for Missing	Median Value for Missing	Winsorized Variables	Missing Stars Imputed Zero
X1	X			X
X2		X		X
X3	X		X	X

Each data set (X1, X2 and X3) is split into a Training dataset (which constitutes 80% of total records, or 10,137) and Test dataset (which constitutes 20% records, or 2,658) using a uniform random distribution. The models described in the section below are created using the training dataset, while the test data set is used to evaluate model performance, over-fitting, and final model selection.

Modeling

In the modeling phase, a total of 6 types of models are investigated:

1. Multiple Linear Regression (MLR)
2. Poisson Regression (POI)
3. Negative Binomial Regression (NB)
4. Zero Inflated Poisson Regression (ZIP)
5. Zero Inflated Negative Binomial Regression (ZINB)
6. Hurdle Model

An MLR model with stepwise variable selection is first investigated – even though the response variable is a count variable, which results in violations of some of the basic assumptions of linear regression – because MLR models are simple to generate, easy to interpret and fairly robust to deviations from the basic assumptions.

Model types 2 through 6 are more appropriate when the response variable consists of counts. As discussed before, although ZIP and ZINB are more appropriate given the large number of zeros in the target variable, POI and NB models are still evaluated given their simplicity.

ZIP models are used when the mean and variance of the response variable are equal, or almost equal. In the wine data, the target variable does have a variance greater than the mean. This constitutes an ‘over dispersed model’, for which NB or ZINB is the more appropriate model selection.

TARGET

Mean	Variance
3.029	3.711

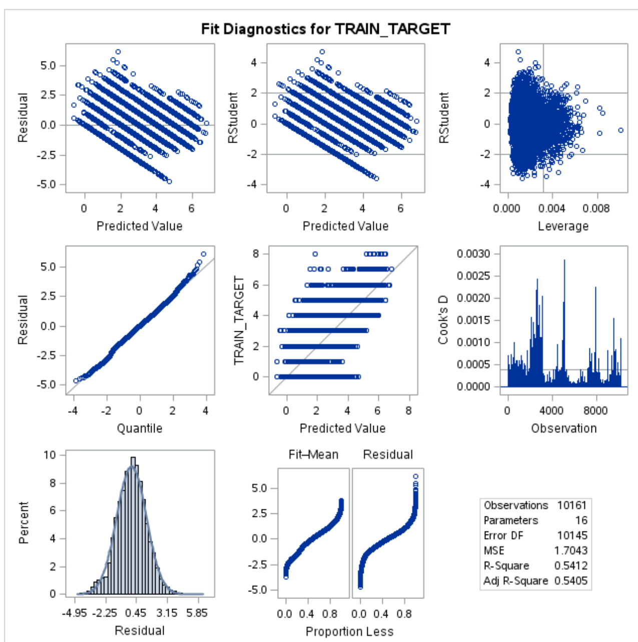
Multi Linear Regression Model

The first model run is a MLR with stepwise automated variable selection to identify the right variables. A total of 15 variables are retained in the model out of a potential 23 variables. The resulting model is statistically significant with a p-value < 0.001 for the overall F-test. An adjusted R^2 of 54% tell us that the model explains a little more than half the variation in the data. All the variables in the model are left are significant at the 0.05 level. Low VIF values indicate no issues of multicollinearity.

Variable	Parameter Estimate	Pr > |t|	Variance Inflation

Intercept	4.51591	<.0001	0
VolatileAcidity	-0.10006	<.0001	1.00702
CitricAcid	0.03113	0.0476	1.00714
Density	-0.79586	0.1166	1.00379
LabelAppeal	0.46443	<.0001	1.10738
AcidIndex	-0.21967	<.0001	1.05646
M_STARS	-0.67569	<.0001	2.41682
M_pH	-0.14950	0.0472	1.00056
IMP_STARS	0.78140	<.0001	2.59043
IMP_Sulphates	-0.03501	0.0210	1.00377
IMP_TotalSulfurDioxide	0.00020735	0.0006	1.00626
IMP_Alcohol	0.01344	0.0003	1.00686
IMP_FreeSulfurDioxide	0.00028484	0.0051	1.20735
IMP_Chlorides	-0.11176	0.0096	1.00373
IMP_pH	-0.03105	0.1211	1.00488
SO2_Ratio	-0.01226	0.0945	1.20673

The model makes sense intuitively for three key variables: Each increase in label appeal increases the number of wine cases sold by 0.46, all other variables kept constant. Each increase in the star count increases the number of wine cases sold by 0.78 while if the stars are missing from the dataset, the number of wine cases decreases by 0.67. Model diagnostics do not show any obvious issues with the model.



Poisson & Negative Binomial Model

POI and NB models are appropriate for count variables. Four models are performed on the data: a POI and NB model with all the input variables, and two more models with any variables with p-values <0.05. The latter two models, discussed here, are very similar results when measured by AIC (POI: 36316, NB: 36410), SBC (POI: 36312, NB: 36420) and number of variables (POI:9, NB: 10). The point estimates for the parameters are shown for the POI model.

Parameter		Estimate	Pr > ChiSq
Intercept		1.1216	<.0001
VolatileAcidity		-0.0324	<.0001
LabelAppeal	-2	-0.6942	<.0001
LabelAppeal	-1	-0.4383	<.0001
LabelAppeal	0	-0.2502	<.0001
LabelAppeal	1	-0.1174	<.0001
AcidIndex		-0.0855	<.0001
M_STARS	0	0.6394	<.0001
IMP_STARS		0.1884	<.0001
IMP_Sulphates		-0.0131	0.0505
IMP_TotalSulfurDioxide		0.0001	0.0072
IMP_Alcohol		0.0041	0.0130
IMP_FreeSulfurDioxide		0.0001	0.0566
IMP_Chlorides		-0.0356	0.0604

The shortlisted variables are the ones expected – LabelAppeal, missing Stars, Imputed Stars etc. These are common to the final selected model, and thus, their interpretation is explored in the final section of the report, titled 'Interpretation of Hurdle Model'.

Zero Inflated Poisson Model

The zero inflated Poisson model is appropriate for the wine data given the large number of zeros. As before the variables with P values < 0.05 are removed from the model. The ZIP model not only calculates the chances of a sale happening, but also a count of the number of wine cases sold if a sale happens. The ZIP and ZINB model assume that the same process governs both the processes.

The Poisson model for calculating the number of variables is given below. It consists of only 4 variables and is thus a very frugal model.

Parameter		Estimate	Pr > ChiSq
Intercept		1.5726	<.0001
VolatileAcidity		-0.0131	0.0895
LabelAppeal	-2	-1.0597	<.0001
LabelAppeal	-1	-0.6314	<.0001
LabelAppeal	0	-0.3429	<.0001
LabelAppeal	1	-0.1521	<.0001
AcidIndex		-0.0216	<.0001
IMP_STARS		0.1012	<.0001
IMP_Alcohol		0.0076	<.0001

The Logistic model for calculating if a sale occurs is given below. It consists of 11 variables and more expensive than the Poisson model above.

Parameter		Estimate	Pr > ChiSq
Intercept		-3.1492	<.0001
VolatileAcidity		0.1787	0.0004
LabelAppeal	-2	-3.3525	<.0001
LabelAppeal	-1	-1.9103	<.0001
LabelAppeal	0	-1.1021	<.0001
LabelAppeal	1	-0.4043	0.0942
AcidIndex		0.4648	<.0001
M_STARS	0	1.5756	<.0001
IMP_Sulphates		0.1426	0.0014
IMP_STARS		-3.6545	<.0001
IMP_TotalSulfurDioxi		-0.0009	<.0001
IMP_Alcohol		0.0274	0.0131
IMP_FreeSulfurDioxid		-0.0011	0.0005
IMP_pH		0.2283	<.0001
SO2_Ratio		0.0537	0.0091

Zero Inflated Negative Binomial Model

The ZINB, like the ZIP model, is appropriate for this data – perhaps even more given the overdispersed nature of the data. As before the variables with P values < 0.05 are removed from the model.

The Poisson model for calculating the number of variables is given below. It consists of only 4 variables; just as frugal as the ZIP model explored above.

Parameter		Estimate	Pr > ChiSq
Intercept		1.5682	<.0001
LabelAppeal	-2	-1.0601	<.0001
LabelAppeal	-1	-0.6328	<.0001
LabelAppeal	0	-0.3437	<.0001
LabelAppeal	1	-0.1523	<.0001
AcidIndex		-0.0210	0.0001
IMP_STARS		0.1011	<.0001
IMP_Alcohol		0.0076	<.0001

The Logistic model for calculating if a sale occurs is given below. It consists of 11 variables.

Parameter		Estimate	Pr > ChiSq
Intercept		-2.5457	<.0001
VolatileAcidity		0.1900	0.0001
LabelAppeal	-2	-3.1990	<.0001
LabelAppeal	-1	-1.8291	<.0001
LabelAppeal	0	-1.0334	<.0001
LabelAppeal	1	-0.3874	0.0805
AcidIndex		0.4496	<.0001
M_pH	0	-0.4858	0.0243
IMP_STARS		-2.2252	<.0001
IMP_Sulphates		0.1374	0.0016
IMP_TotalSulfurDioxi		-0.0009	<.0001
IMP_Alcohol		0.0261	0.0156
IMP_FreeSulfurDioxid		-0.0010	0.0006
IMP_pH		0.2198	0.0001
SO2_Ratio		0.0511	0.0106

As before, the variables are common to the final selected model, and thus, their interpretation is explored in the final section of the report, titled 'Interpretation of Hurdle Model'.

Hurdle Model

A hurdle model is similar to a ZIP or ZINB model in that there are two models run separately: one to determine if there was a sale, and another to predict how many cases were sold. The difference between the ZIP/ZINB model and the hurdle model is that the hurdle model can assume different processes for each part of the model.

To create a hurdle model, first, two additional variables are created:

1. TARGET_FLAG: This variable is 1 if the Target response variable is positive, else it is 0.
2. TARGET_AMT: This variable is the counts of wine cases, if there are any sales at all, i.e. if TARGET_FLAG equals 1.

The two parts of the model are:

1. Logistic regression with stepwise regression to predict TARGET_FLAG
2. Poisson model to predict TARGET_AMT

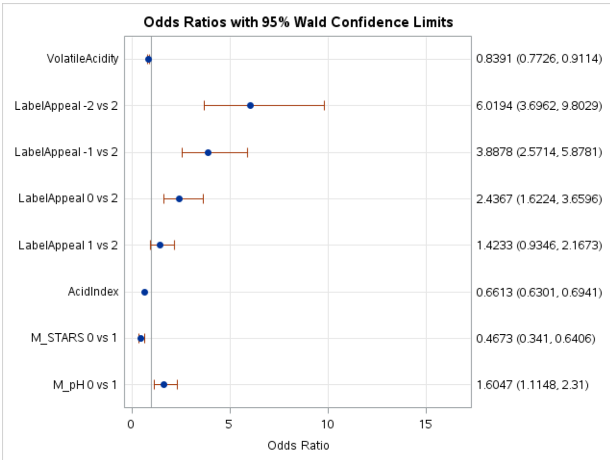
The final score for the hurdle model equals the product of the two probabilities:

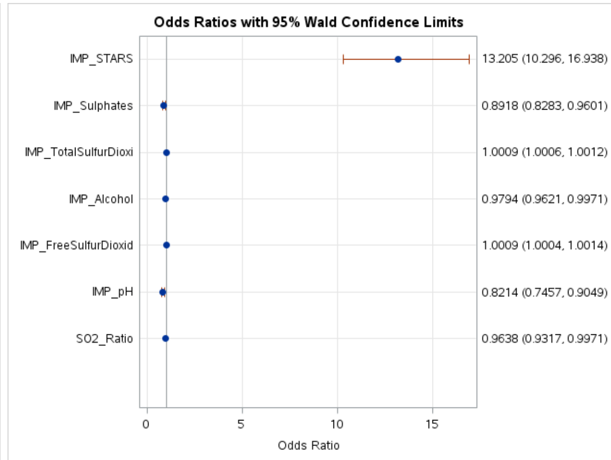
PREDICTED_TARGET_FLAG x PREDICTED_TARGET_AMT

Logistic Model. This model consists of 12 variables, with the point estimates of the coefficients given in the table below.

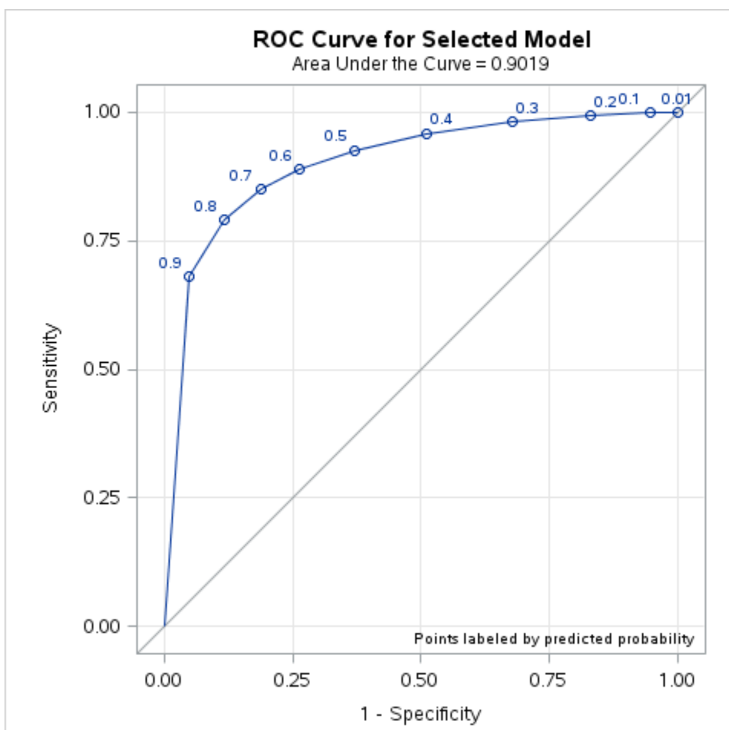
Parameter		Estimate	Pr > ChiSq	Exp(Est)
Intercept		2.3000	<.0001	9.974
VolatileAcidity		-0.1754	<.0001	0.839
LabelAppeal	-2	1.7950	<.0001	6.019
LabelAppeal	-1	1.3578	<.0001	3.888
LabelAppeal	0	0.8906	<.0001	2.437
LabelAppeal	1	0.3530	0.1000	1.423
AcidIndex		-0.4135	<.0001	0.661
M_STARS	0	-0.7607	<.0001	0.467
M_pH	0	0.4729	0.0109	1.605
IMP_STARS		2.5806	<.0001	13.205
IMP_Sulphates		-0.1146	0.0023	0.892
IMP_TotalSulfurDioxi		0.000864	<.0001	1.001
IMP_Alcohol		-0.0208	0.0230	0.979
IMP_FreeSulfurDioxid		0.000887	0.0005	1.001
IMP_pH		-0.1967	<.0001	0.821
SO2_Ratio		-0.0368	0.0334	0.964

The 95% Wald confidence limits for the point estimates are shown graphically below. Visually, we can see that LabelAppeal, IMP_Stars have the largest impact of the response variable. With each additional star, the odds of purchasing at least one case of wine versus not purchasing any increases by 13.





The performance of this model is excellent, with an Area Under the Curve for the ROC curve of 90.19%. The percent concordant, which is the number of responses with a lower ordered response value (TARGET_FLAG = 0) has a lower predicted mean score than the observation with the higher ordered response value (TARGET_FLAG=1) – the higher this value, the better. The % concordant is 87%.



Poisson Model. After eliminating the variables not statistically significant at the 0.05 level, the Poisson model has 6 variables.

Parameter		Estimate	Pr > ChiSq
Intercept		1.3709	<.0001
VolatileAcidity		-0.0144	0.1007
LabelAppeal	-2	-1.4329	<.0001
LabelAppeal	-1	-0.8009	<.0001
LabelAppeal	0	-0.4248	<.0001
LabelAppeal	1	-0.1840	<.0001
AcidIndex		-0.0239	0.0001
M_STARS	0	-0.0572	0.0477
IMP_STARS		0.1250	<.0001
IMP_Alcohol		0.0099	<.0001

Evaluation

A total of 31 models are compared – comprising of the 6 model types across the 3 data sets. The models are evaluated against 6 criteria:

1. AIC: Akaike Information Criterion. AIC is used for the comparison of nonnested models on the same sample. The model with the smallest AIC is considered the best, although the AIC value itself is not meaningful.
2. SBC: Schwarz Criterion. Like AIC, SC penalizes for the number of predictors in the model and the smallest SC is most desirable and the value itself is not meaningful.
3. MAE: Mean Average Error. This is calculated as the average value of the absolute error. Error is defined as the difference between the number of cases sold (response) and the predicted number of cases sold.
4. RMSE: Root Mean Square Error. This is calculated as the average value of the root mean square of the error value.
5. Model complexity: This is defined by the number of predictors included in the equation, as well as the intuitive understanding of the model.
6. Sum(T): This is the sum of the response variable. Each model should be able to predict the total number of cases sold in the original dataset correctly.

The table on page 23 compares all the models using these criteria for two datasets X1 and X3. in the interest of brevity, data set X2 is not shown. Data set X2 (usage of median values for imputation) was clearly the lowest ranking amongst the three data sets.

Key Observations

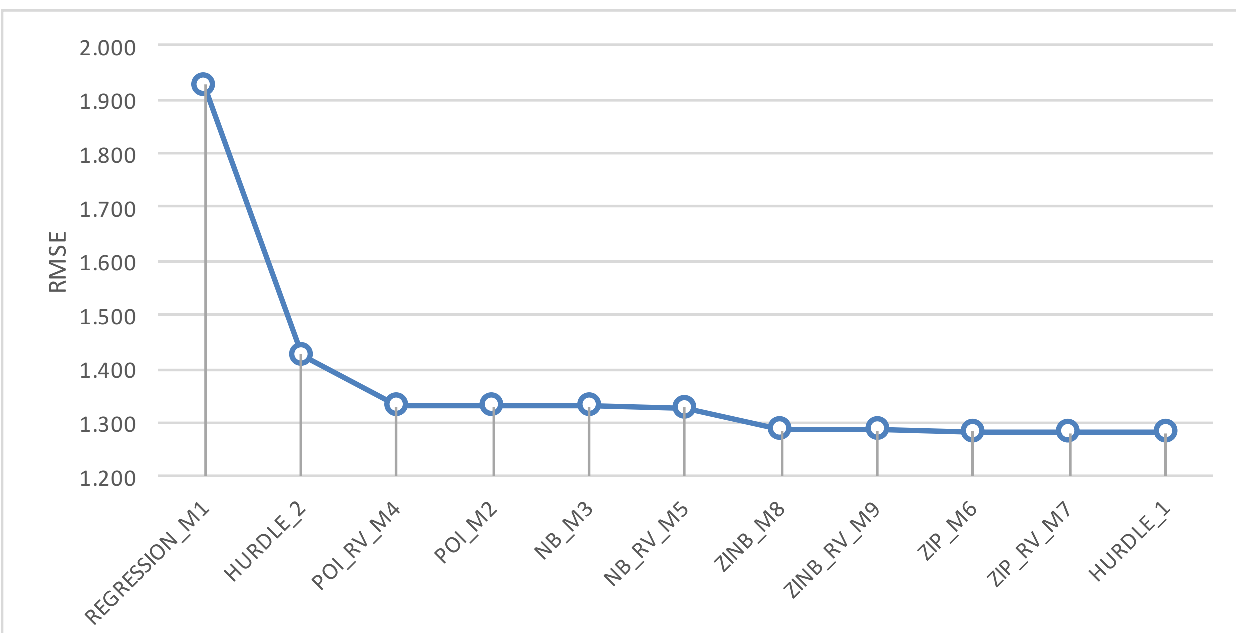
1. Data set X3 clearly out performs data set X1. Thus, winsorizing the data is the right approach for the final model.
2. Based off of the AIC and SBC criteria, the ZIP models with reduced number of variables is the best model, followed closely by the ZINB model. The hurdle model cannot be compared using the AIC SBC criteria given it's a combination of two models.
3. The top performing model based off of the RMSE performance indicator is the hurdle model (ID28). Although the hurdle model ID29 is better in MAE, it performs considerably worse in RMSE and Sum(T). So ID29 is not a good model.
4. None of the models (except ID29) show signs of overfitting. This is observed by the strong performance agreement between the Training and Test data sets, compared on page 24. The errors in MAE and RMSE between the training and test models is very low. Furthermore, the Sum(T) error to the original target variable is also very low (within +/- 3%).

ID	Model Name	Type	Data Set	AIC	SBC	k	Zero Model k	Training MAE	Training RMSE	Training Sum(T)[1]	Training MAE	Training RMSE	Training Sum(T)[2]	Kaggle Score
1	REGRESSION_M1	Linear Regr	X1	-	-	14	-	1.000	1.926[3]	30900	1.009	1.928	7969	-
2	POI_M2	Poisson	X1	36323	36518	14	-	1.000	1.310	30769	1.013	1.330	7911	-
3	NB_M3	Neg Binomial	X1	36325	36528	14	-	1.000	1.310	30769	1.013	1.330	7911	-
4	POI_RV_M4	Poisson	X1	36316	36410	9	-	1.002	1.311	30611	1.008	1.331	7905	-
5	NB_RV_M5	Neg Binomial	X1	36312	36420	10	-	0.999	1.311	30800	1.004	1.328	7926	-
6	ZIP_M6	Zero Inflated Poi	X1	32433	32823	14	14	0.915	1.257	30997	0.931	1.283	7923	-
7	ZIP_RV_M7	Zero Inflated Poi	X1	32389[4]	32570	5	12	0.917	1.258	30901	0.938	1.282	7921	1.378
8	ZINB_M8	Zero Inflated NB	X1	32559	32957	14	14	0.922	1.260	30777	0.942	1.286	7900	-
9	ZINB_RV_M9	Zero Inflated NB	X1	32477	32651	5	10	0.920	1.261	30796	0.940	1.285	7899	1.388
30	HURDLE_1	Logistic w/ Step-wise + Poisson	X1	-	-	6	12	0.929	1.257	30865	0.950	1.280	7916	-
31	HURDLE_2[5]	Logistic w/ Step-wise + Poisson	X1	-	-	6	12	0.878	1.405	28003	0.898	1.427	7158	-
19	REGRESSION_M1	Linear Regr	X3	-	-		-	1.001	1.926	30907	1.013	1.928	7959	-
20	POI_M2	Poisson	X3	36321	36516	14	-	0.999	1.311	30771	1.013	1.330	7928	-
21	NB_M3	Neg Binomial	X3	36323	36525	14	-	0.999	1.311	30771	1.013	1.330	7928	-
22	POI_RV_M4	Poisson	X3	36307	36416	11	-	0.998	1.312	30792	1.009	1.328	7936	-
23	NB_RV_M5	Neg Binomial	X3	36311	36419	10	-	0.998	1.312	30804	1.005	1.328	7939	-
24	ZIP_M6	Zero Inflated Poi	X3	32424[6]	32814	14	-	0.916	1.257	30996	0.931	1.282	7917	-
25	ZIP_RV_M7	Zero Inflated Poi	X3	32384	32558	5	11	0.917	1.258	30926	0.939	1.281	7923	1.378
26	ZINB_M8	Zero Inflated NB	X3	32550	32947	14	14	0.921	1.259	30764	0.934	1.285	7897	-
27	ZINB_RV_M9	Zero Inflated NB	X3	32468	32641	4	11	0.917	1.260	30819	0.934	1.284	7891	1.417
28	HURDLE_1	Logistic w/ Step-wise + Poisson	X3	-	-	6	12	0.925	1.257	30856	0.951	1.279	7915	1.374
29	HURDLE_2[1]	Logistic w/ Step-wise + Poisson	X3	-	-	6	12	0.877	1.379	28478	0.890	1.392	7277	1.598

[1] Sum of the response variable in the training data. Sum(T) for input data = 30883
[2] Sum of the response variable in the test data. Sum(T) for input data = 7874
[3] All values marked in red represent the highest (worst) 5 models in each column
[4] All values marked in yellow represent the lowest (best) 15% models in each column
[5] Cutoff selected for logistic portion of the model selected = 0.6, based off of the ROC curve
[6] All values marked in green represent the lowest (best) 10% models in each column

		Training	Test	Training to Test Error	
ID	Model Name	Sum(T) Error	Sum(T) Error	MAE	RMSE
19	REGRESSION_M1	100.1%	101.1%	1.1%	0.1%
20	POI_M2	99.6%	100.7%	1.4%	1.4%
21	NB_M3	99.6%	100.7%	1.4%	1.4%
22	POI_RV_M4	99.7%	100.8%	1.1%	1.3%
23	NB_RV_M5	99.7%	100.8%	0.7%	1.2%
24	ZIP_M6	100.4%	100.5%	1.6%	2.0%
25	ZIP_RV_M7	100.1%	100.6%	2.4%	1.8%
26	ZINB_M8	99.6%	100.3%	1.4%	2.1%
27	ZINB_RV_M9	99.8%	100.2%	1.8%	1.9%
28	HURDLE_1	99.9%	100.5%	2.8%	1.8%
29	HURDLE_2	92.2%	92.4%	1.6%	0.9%

A comparison of the RMSE performance of the 20% test dataset reveals that the hurdle model, ZIP reduced variables, ZIP full model, and ZINB reduced variables are the best performers. Between these four models though, the differences in performance are marginal. Thus, it's best to pick the model that's easiest to explain and/or the most frugal.



This rules out ZIP_M6 since it uses a large number of variables. HURDLE_1 and ZIP_RV_M7 have similar number of variables and similar variables as well. I choose HURDLE_1 as the model of choice given it's also the lowest on the Kaggle competition, which itself is an exposure to 10% of a new test data set untouched by the modeling process.

Interpretation of Selected Hurdle Model

Logistic Sub-Model. The intercept indicates when LabelAppeal is zero (i.e. Label Appeal = +2), and missing stars is true (i.e. M_STARS=1), missing pH value is true (i.e. M_pH = 1), the log-odds of a wine case purchase is 2.3; in other words, it is 9.9 times more likely for a wine purchase under these conditions.

Label appeal is monotonically inversely related to the response variable. With each decrease in label appeal from +2, the odds of purchasing a wine case reduce by 1.4 times for LA=1, 2.4 times for LA=0, 3.8 times for LA=-1 and 6 times for LA=-2.

The missing value of stars is highly predictive too. This makes sense since it's highly unlikely that a wine appreciated for its taste would go unranked in the dataset. Barring errors in data entry, it indicates that if the stars are not entered, the chances of purchasing a case drop by 50%.

Volatile acidity and acid index both point towards how acidic a wine is. Increase in VA causes wines to smell more pungent and taste bad. Similarly, increase in the acid index causes wine to go towards tasting like vinegar. This is represented in the estimates: a unit increase in either reduces the changes of wine purchase by a factor of 0.8 and 0.6 respectively.

The low values of the point estimates for Total Sulfur Dioxide and Free Sulfur Dioxide indicate that these variables may not be as important to the final model, but an expert should comment on this.

Parameter		Estimate	Pr > ChiSq	Exp(Est)
Intercept		2.3000	<.0001	9.974
VolatileAcidity		-0.1754	<.0001	0.839
LabelAppeal	-2	1.7950	<.0001	6.019
LabelAppeal	-1	1.3578	<.0001	3.888
LabelAppeal	0	0.8906	<.0001	2.437
LabelAppeal	1	0.3530	0.1000	1.423
AcidIndex		-0.4135	<.0001	0.661
M_STARS	0	-0.7607	<.0001	0.467
M_pH	0	0.4729	0.0109	1.605

IMP_STARS	2.5806	<.0001	13.205
IMP_Sulphates	-0.1146	0.0023	0.892
IMP_TotalSulfurDioxide	0.000864	<.0001	1.001
IMP_Alcohol	-0.0208	0.0230	0.979
IMP_FreeSulfurDioxide	0.000887	0.0005	1.001
IMP_pH	-0.1967	<.0001	0.821
SO2_Ratio	-0.0368	0.0334	0.964

Poisson Sub-Model. On similar lines as the model above, the counts of wine cases, if cases were purchased, are a function of only 6 key variables. The general interpretation of this model is that wines with lower acidity, higher label appeal, no missing stars, and a higher number of stars results in a higher purchase count of wine cases.

All other variables kept constant, if the number of stars are missing, it results in 100*(e^-0.0572-1) = 5.56% reduction in the number of cases purchased.

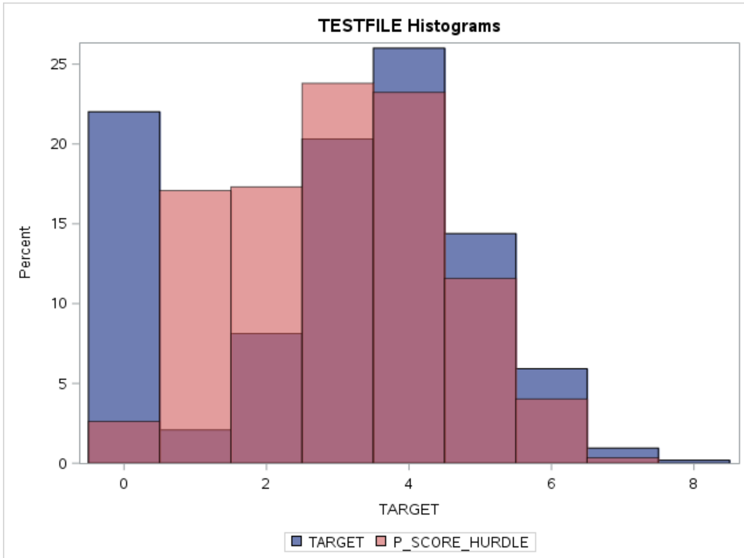
Increase in one star increases the number of cases purchased by 1.13 (on average) controlling for other variables.

Increase in acidity reduces the number of cases purchased by 0.96 on average for 1-unit increase in VA and AI combined.

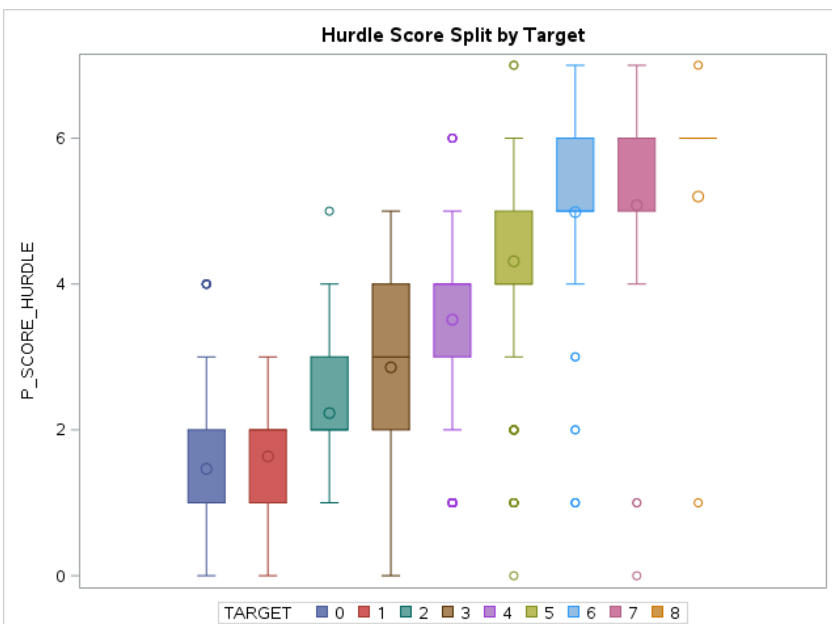
Label appeal has a strong impact on the number of cases ordered. A drop from +2 to +1 in appeal reduces the number of cases by 16.8%, while moving from +2 to -2 reduces the number of cases by 76%.

Parameter		Estimate	Pr > ChiSq
Intercept		1.3709	<.0001
VolatileAcidity		-0.0144	0.1007
LabelAppeal	-2	-1.4329	<.0001
LabelAppeal	-1	-0.8009	<.0001
LabelAppeal	0	-0.4248	<.0001
LabelAppeal	1	-0.1840	<.0001
AcidIndex		-0.0239	0.0001
M_STARS	0	-0.0572	0.0477
IMP_STARS		0.1250	<.0001
IMP_Alcohol		0.0099	<.0001

The resulting predictions when compared against the original response variable shows the following frequency distribution. What is observed is that the hurdle model (shown in red) under predicts the number of zero cases, but over predicts the number of 1 and 2 cases purchased. Having said this, the overall metric of Sum(T) is 99.9% accurate for training, 100.5% accurate for the test data set. (Training set: 30883 original to predicted 30856, testing set: 7874 original to 7915 predicted).



This accuracy in the predictions can be observed in a boxplot split by the original target variable. Overall, the mean hurdle scores monotonically increase with the response variable. For the lower scores of target = 0 or 1, the predicted score as a mean higher than expected of ~1.2, with some outliers at 4, which supports the conclusion above that the model under predicts the number of zero cases.



Conclusion

A variety of models were investigated for the wine sales problem data set. Two competing models ZIP and the Hurdle model were shortlisted. Ultimately, the Hurdle model was selected due to its high performance in AIC, SBC and RMSE metrics in the 20% test dataset, as well as the higher Kaggle ranking.

References

Chapman, P., & Clinton, J. (2000, August). CRISP-DM 1 - The Modeling Agency. Retrieved November 8, 2016, from https://the-modeling-agency.com/crisp-dm.pdf

Hu, M.-C., Pavlicova, M., & Nunes, E. V. (2011). Zero-inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. The American Journal of Drug and Alcohol Abuse, 37(5), 367–375. http://doi.org/10.3109/00952990.2011.597280

Hoffmann, J. P. (2004). *Generalized linear models: An applied approach*. Boston: Pearson A and B.

Introduction to SAS. UCLA: Statistical Consulting Group. from http://www.ats.ucla.edu/stat/sas/notes2/ (accessed November 20, 2016).