Wine Sales Project

FY 2016

Author Eric Lewis

TABLE OF CONTENTS

Contents

Bonus

Scored File as SAS Data Set

Filename: wine_score_03. sas7bdat

SAS Code on page 32

SAS Macro Use

One of the advantages is using SAS macros is to create clear references towards file and data usage at the beginning of the SAS code, for example &INFILE., &TEMPFILE., and &SCRUBFILE. used in this model.

SAS Code on page 38

Develop Logistic / Poisson Model

The Kaggle score using the Logistic and Poisson model combined is 1.30570.

SAS Code on page 33

PROC Logistic

- AIC is 13277.788
- SC is 13285.245
- -2 Log L is 13275.788

PROC Poisson

- AIC is 41169.2589
- AICC is 41169.3529
- BIC is 41348.2224

Introduction

The objective of this data analysis is to build a model to predict the number of cases of wine that will be sold based upon properties of the wine within the given data set. Five models will be compared based upon their highest probability of predicting the target value, cases of wine sold. The characteristics of the wine are measures utilized in Linear, Logistic, and Poisson regressions to provide empirical for model comparison. The final determination will be determined by the lowest score using Kaggle, with the zero as the lowest score possible though not probably. A final model will be selected and a short and long term recommendation will be delivered from this analysis.

Exploratory Data Analysis

There are two main components in developing this predictive model:

- **Training data set** – utilized for exploratory data analysis, data preparation, building and selecting a predictive model. This data set contains 12,795 observations with the variables as shown in table 1 below.
- **Test data set** – utilized to score the model selected in the training phase of this analysis. The model results are being scored using Kaggle. This data set contains 3,335 observations less the variable Each line item in the data set contains the specific data on the insured.

This analysis will determine which data elements are the highest correlated towards determining the target, the number cases of wine that will be sold.

The following table provides the variable name, type, and definition as the initial step towards understanding the data.

VARIABLE NAME	TYPE	DEFINITION
INDEX		Identification Variable (do not use)
Target		Number of Cases Purchased
AcidIndex	Continuous	Proprietary method of testing total acidity of wine by using a weighted average
Alcohol	Continuous	Alcohol Content
Chlorides	Continuous	Chloride content of wine
CitricAcid	Continuous	Citric Acid Content
Density	Continuous	Density of Wine
FixedAcidity	Continuous	Fixed Acidity of Wine
FreeSulfurDioxide	Continuous	Sulfur Dioxide content of wine
LabelAppeal	Categorical	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	Continuous	Residual Sugar of wine
Stars	Categorical	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor
Sulphates	Continuous	Sulfate content of wine
TotalSulfurDioxide	Continuous	Total Sulfur Dioxide of Wine
VolatileAcidity	Continuous	Volatile Acid content of wine
pH		pH of wine

Table 1: Data Dictionary

The following table provides the variable name, and theoretical effect as an additional step towards understanding the data and how it relates towards building the predictive model.

VARIABLE NAME	THEORETICAL EFFECT
INDEX	None
Target	None
AcidIndex	
Alcohol	
Chlorides	
CitricAcid	
Density	
FixedAcidity	
FreeSulfurDioxide	
LabelAppeal	Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales.
ResidualSugar	
Stars	A high number of stars suggests high sales
Sulphates	
TotalSulfurDioxide	
VolatileAcidity	
pH	

Table 2: Data Dictionary {THEORETICAL EFFECT}

Data Exploration

Missing Data

The training data set contains the following statistical variables with missing data. Depending on the next step, which is correlating the statistical variables with the target, the number cases of wine that will be sold. The key observations from table 3 are the quantity of missing values per statistical variable along with the mean and standard deviation if we choose to impute the missing data elements.

Variable	N	N Missing	Mean	Std Dev
INDEX	12795	0	8069.98	4656.91
TARGET	12795	0	3.0290739	1.9263682
FixedAcidity	12795	0	7.0757171	6.3176435
VolatileAcidity	12795	0	0.3241039	0.7840142
CitricAcid	12795	0	0.3084127	0.8620798
ResidualSugar	12179	616	5.4187331	33.7493790
Chlorides	12157	638	0.0548225	0.3184673
FreeSulfurDioxide	12148	647	30.8455713	148.7145577
TotalSulfurDioxide	12113	682	120.7142326	231.9132105
Density	12795	0	0.9942027	0.0265376
pH	12400	395	3.2076282	0.6796871
Sulphates	11585	1210	0.5271118	0.9321293
Alcohol	12142	653	10.4892363	3.7278190
LabelAppeal	12795	0	-0.0090660	0.8910892
AcidIndex	12795	0	7.7727237	1.3239264
Stars	9436	3359	2.0417550	0.9025400

Table 3: Missing & Mean

We will decide whether to impute or exclude the statistical variable from the predictive model depending on the results from correlating each statistic with the correlation to the target, the number cases of wine that will be sold.

- ResidualSugar
- Chlorides
- FreeSulfurDioxide
- TotalSulfurDioxide
- pH
- Sulphates
- Alcohol
- Stars

Variable Correlation to Target Flag

Key observations from table 4 demonstrates that eleven of the fourteen variables within the wine sales dataset are correlated with the target, the number cases of wine that will be sold.

Variable	Target
TARGET	1.00000
Acidindex	-0.24605 <.0001
Imp_alcohol	0.06043 <.0001
Imp_chlorides	-0.03724 <.0001
Citricacid	0.00868 0.3260
Density	-0.03552 <.0001
Fixedacidity	-0.04901 <.0001
Imp_freesulfurdioxide	0.04269 <.0001
labelappeal	0.35650 <.0001
Imp_residualsugar	0.01607 0.0691
Imp_stars	0.40013 <.0001
Imp_sulphates	-0.03691 <.0001
Imp_totalsulfurdioxide	0.05163 <.0001
Volatileacidity	-0.08879 <.0001
Imp_ph	-0.00928 0.2939

Table 4: Variable Correlation with Target

Visual Representation of Variables

The purpose of the visual or graphical representation of the distribution within the variables is to provide observations toward the predictive model variable selection to complement the correlation with target as shown in figure 1 below.   The key observation of the Distribution of Target is that this distribution is that of a histogram of Poisson distribution. This provides an initial indication that when we are comparing the various regression models, the Poisson model may be likely the best method.

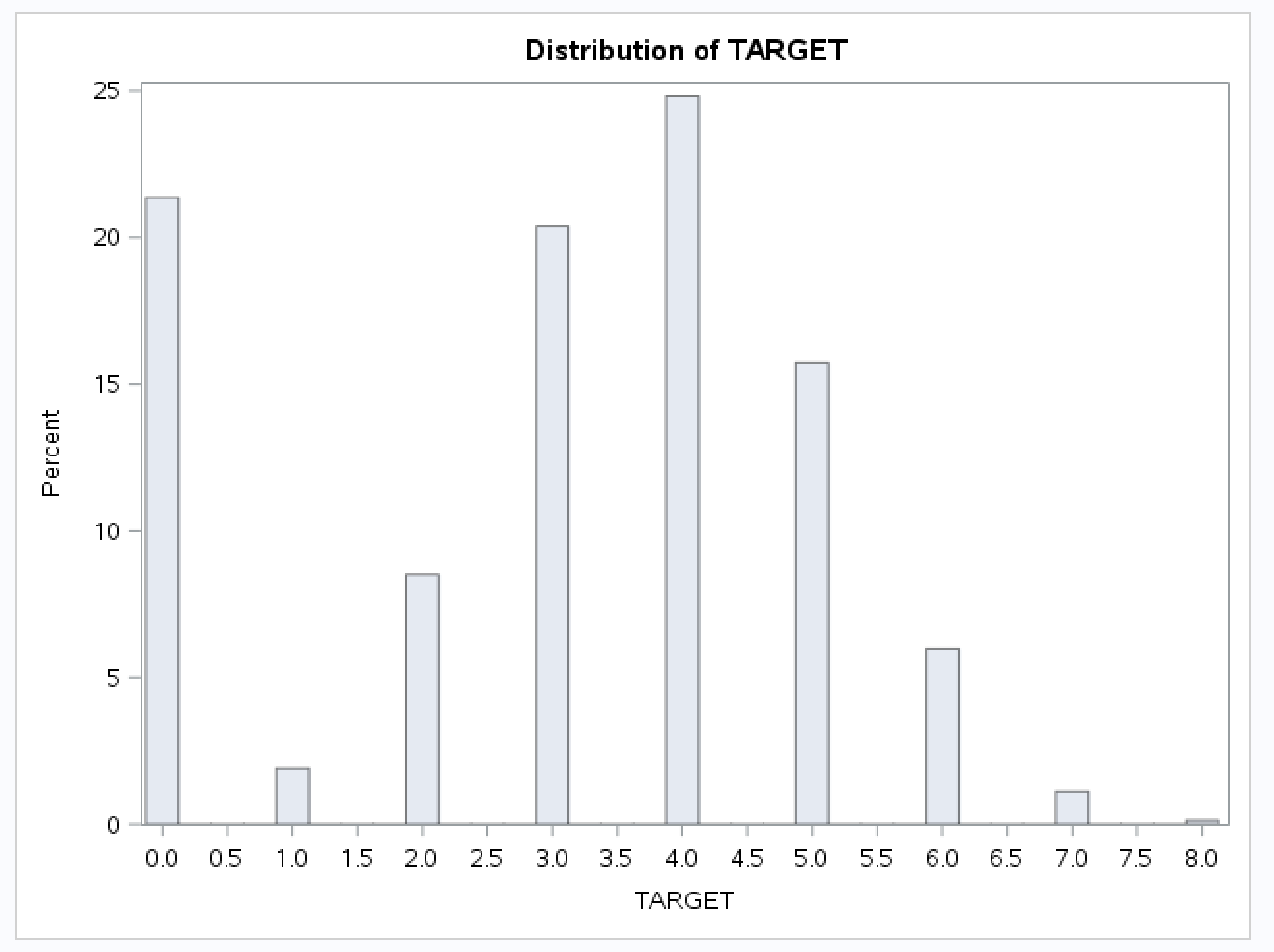


Figure 1: Distribution of Target

Data Preparation

The data preparation phase of this analysis encompasses preparing the data for modeling. Various techniques that will be reviewed are: imputing missing values, flagging missing variables, data transformation through combining variables and through the use of mathematical transformations.

Missing Values

The initial testing of the model will include imputing the following variable with missing values based upon their mean shown in table 2. The following variables are imputed using the mean in the case of missing data.

- YOJ – mean of 10.4992864
- ResidualSugar – mean of 5.4187331
- Chlorides – mean of 0.0548225
- FreeSulfurDioxide – mean of 30.8455713
- TotalSulfurDioxide – mean of 120.7142326
- pH – mean of 3.2076282
- Sulphates – mean of 0.5271118
- Alcohol – mean of 10.4892363
- Stars – mean of 2.0417550

Transforming Data into Buckets

The following variables are transformed into buckets based partly upon the theoretical effect of increasing wine sold and their analysis of maximum likelihood estimates.

- M_Stars using the value of 0, see frequency table 5.
- IMP_Stars using the values of 1, 2, 3, see frequency table 6.
- Label Appeal using the values of -2, -1, 0, 1, see frequency table 7.

Mathematical Data Transformations

Attempts were made to transform variables mathematically for example Logarithm and square root data transformations were attempted; however, the predictive value of the model demonstrated no improvement.

Combining Variables

Attempts were made to combine variables to perform ratio analysis; however, the predictive value of the model demonstrated no improvement.

Frequency Tables

The following three frequency tables provides key percentage to target flag observations on each variable array subscript. These tables are utilized in to determine which variable array subscript to use in the models as shown above.

Table of M_STARS by TARGET_FLAG			
M_Stars	TARGET_FLAG		
	0	1	Total
0	696 5.44 7.38 25.46	8740 68.31 92.62 86.87	9436 73.75
1	2038 15.93 60.67 74.54	1321 10.32 39.33 13.13	3359 26.25
Total	2734 21.37	10061 78.63	12795 100.00

Table 5: Frequency Table of M_Stars

Table of IMP_STARS by TARGET_FLAG			
IMP_Stars	TARGET_FLAG		
	0	1	Total
1	607 4.74 19.95 22.20	2435 19.03 80.05 24.20	3042 23.77
2	2127 16.62 30.70 77.80	4802 37.53 69.30 47.73	6929 54.15
3	0 0.00 0.00 0.00	2212 17.29 100.00 21.99	2212 17.29
4	0 0.00 0.00 0.00	612 4.78 100.00 6.08	612 4.78
Total	2734 21.37	10061 78.63	12795 100.00

Table 6: Frequency Table of IMP_Stars

Table of Label Appeal by TARGET_FLAG			
Label Appeal	TARGET_FLAG		
	0	1	Total
-2	102 0.80 20.24 3.73	402 3.14 79.76 4.00	504 3.94
-1	671 5.24 21.40 24.54	2465 19.27 78.60 24.50	3136 24.51
0	1193 9.32 21.24 43.64	4424 34.58 78.76 43.97	5617 43.90
1	660 5.16 21.65 24.14	2388 18.66 78.35 23.74	3048 23.82
2	108 0.84 22.04 3.95	382 2.99 77.96 3.80	490 3.83
Total	2734 21.37	10061 78.63	12795 100.00

Table 7: Frequency Table of Label Appeal

Building Models

Five models were utilized as comparison for this analysis using Linear, Logistic, and Poisson regression analysis.

First Model

This model is known as the base model. It is a model using linear regression with stepwise variable selection. The linear regression model scored using Kaggle 1.35994, which is not the best score of the five models. On the surface linear regression appears to fit based upon the means comparison in table 21; however, the means error procedure in table 22 demonstrates the linear regression model has a higher error mean than the selected model, the Zero Inflated Poisson.

The REG Procedure
Model: MODEL1
Dependent Variable: TARGET

Number of Observations Read	12795
Number of Observations Used	12795

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	25526	2836.20490	1651.88	<.0001
Error	12785	21951	1.71696		
Corrected Total	12794	47477			

Root MSE	1.31033	R-Square	0.5376
Dependent Mean	3.02907	Adj R-Sq	0.5373
Coeff Var	43.25838		

Table 8: Model Linear Regression

		Standard		
Parameter	Estimate	Error	t Value	Pr > |t|
Intercept	1	3.46236	0.08655	40.00
AcidIndex	1	-0.20007	0.00894	-22.38
IMP_Alcohol	1	0.01246	0.00320	3.89
IMP_Chlorides	1	-0.11742	0.03736	-3.14
IMP_FreeSulfurDioxide	1	0.00028171	0.00008008	3.52
LabelAppeal	1	0.46626	0.01367	34.10
IMP_STARS	1	0.78030	0.01568	49.77
M_STARS	1	-2.24712	0.02695	-83.39
IMP_TotalSulfurDioxide	1	0.00024441	0.00005634	4.34
VolatileAcidity	1	-0.09693	0.01482	-6.54

Table 9: Model Linear Regression

Second Model

This model is a model using SAS GENMOD with negative binomial distribution. This model demonstrated similar results as the linear regression and appears to fit based upon the means comparison in table 21; however, the means error procedure in table 22 demonstrates the linear regression model has a higher error mean than the selected model, the Zero Inflated Poisson.

Model Information

Data Set	WORK.FIXFILE
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	TARGET

Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	13E3	13777.2487	1.0776
Scaled Deviance	13E3	13777.2487	1.0776
Pearson Chi-Square	13E3	11306.8800	0.8844
Scaled Pearson X2	13E3	11306.8800	0.8844
Log Likelihood		8737.5361	
Full Log Likelihood		-22859.6352	
AIC (smaller is better)		45741.2704	
AICC (smaller is better)		45741.2910	
BIC (smaller is better)		45823.2953	
Criterion	DF	Value	Value/DF

Table 10: Model GENMOD NB

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.4480	0.0411	1.3674	1.5285	1240.97	<.0001
AcidIndex	1	-0.0804	0.0045	-0.0892	-0.0716	319.33	<.0001
IMP_Alcohol	1	0.0035	0.0014	0.0007	0.0062	6.14	0.0132
IMP_Chlorides	1	-0.0368	0.0165	-0.0690	-0.0045	4.99	0.0255
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.69	0.0056
LabelAppeal	1	0.1587	0.0061	0.1467	0.1707	671.33	<.0001
IMP_STARS	1	0.1882	0.0061	0.1762	0.2001	954.55	<.0001
M_STARS	1	-1.0246	0.0170	-1.0578	-0.9913	3642.43	<.0001
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	11.95	0.0005
VolatileAcidity	1	-0.0312	0.0065	-0.0440	-0.0184	22.95	<.0001
Dispersion	0	0.0000	0.0000	0.0000	0.0000		

Table 11: Model GENMOD Maximum Likelihood

Third Model

This model is a model using SAS Logistic distribution. This model demonstrated similar results as the linear regression and appears to fit based upon the means comparison in table 21; however, the means error procedure in table 22 demonstrates the linear regression model has a higher error mean than the selected model, the Zero Inflated Poisson.

Model Information	
Data Set	WORK.FIXFILE
Response Variable	TARGET_FLAG
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	13277.788	7675.111
SC	13285.245	7749.679
-2 Log L	13275.788	7655.111

Analysis of Maximum Likelihood Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.9599	0.2347	69.7334	<.0001
AcidIndex	1	-0.3836	0.0213	325.2904	<.0001
IMP_Alcohol	1	-0.0208	0.00788	6.9604	0.0083
IMP_Chlorides	1	-0.1497	0.0917	2.6662	0.1025
IMP_FreeSulfurDioxid	1	0.000592	0.000200	8.7965	0.0030
LabelAppeal	1	-0.4644	0.0332	195.3493	<.0001
IMP_STARS	1	2.5553	0.1118	522.7085	<.0001
M_STARS	1	-4.3686	0.1113	1541.4447	<.0001
IMP_TotalSulfurDioxi	1	0.000972	0.000139	48.5400	<.0001
VolatileAcidity	1	-0.1822	0.0364	25.0431	<.0001

Table(s) 12: Model Logistic Maximum Likelihood

Fourth Model

This model is a model using SAS GENMOD with negative binomial distribution, also known as Hurdle model 1. This model demonstrated less results than the previous model based upon the means comparison in table 21; however, the means error procedure in table 22 demonstrates this model has a higher error mean than all the other models with the exception of hurdle model 2.

Model Information	
Data Set	WORK.FIXFILE
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	TARGET

Criteria for Assessing Goodness of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13777.2487	1.0776
Scaled Deviance	13E3	13777.2487	1.0776
Pearson Chi-Square	13E3	11306.8800	0.8844
Scaled Pearson X2	13E3	11306.8800	0.8844
Log Likelihood		8737.5361	
Full Log Likelihood		-22859.6352	
AIC (smaller is better)		45741.2704	
AICC (smaller is better)		45741.2910	
BIC (smaller is better)		45823.2953	

Table(s) 13: Model GENMOD NB

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	1.4480	0.0411	1.3674	1.5285	1240.97	<.0001
AcidIndex	1	-0.0804	0.0045	-0.0892	-0.0716	319.33	<.0001
IMP_Alcohol	1	0.0035	0.0014	0.0007	0.0062	6.14	0.0132
IMP_Chlorides	1	-0.0368	0.0165	-0.0690	-0.0045	4.99	0.0255
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.69	0.0056
LabelAppeal	1	0.1587	0.0061	0.1467	0.1707	671.33	<.0001
IMP_STARS	1	0.1882	0.0061	0.1762	0.2001	954.55	<.0001
M_STARS	1	-1.0246	0.0170	-1.0578	-0.9913	3642.43	<.0001
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	11.95	0.0005
VolatileAcidity	1	-0.0312	0.0065	-0.0440	-0.0184	22.95	<.0001
Dispersion	0	0.0000	0.0000	0.0000	0.0000		

Table(s) 14: Model GENMOD NB Maximum Likelihood

Fifth Model

This model is a model using SAS GENMOD with negative binomial distribution and using a class of imp_stars, also known as hurdle model 2. This model demonstrated less results than the previous models based upon the means comparison in table 21; however, the means error procedure in table 22 demonstrates this model has the highest error mean than all the other models.

Model Information	
Data Set	WORK.FIXFILE
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	TARGET

Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	13E3	13662.0427	1.0688
Scaled Deviance	13E3	13662.0427	1.0688
Pearson Chi-Square	13E3	11303.1231	0.8842
Scaled Pearson X2	13E3	11303.1231	0.8842
Log Likelihood		8795.1391	
Full Log Likelihood		-22802.0322	
AIC (smaller is better)		45630.0644	
AICC (smaller is better)		45630.0929	
BIC (smaller is better)		45727.0029	

Table(s) 15: Model GENMOD NB, Class IMP_Stars

	Analysis of Maximum Likelihood Parameter Estimates							
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	1.5568	0.0401	1.4783	1.6354	1510.34	<.0001
AcidIndex		1	-0.0795	0.0045	-0.0883	-0.0706	311.50	<.0001
IMP_Alcohol		1	0.0038	0.0014	0.0011	0.0066	7.34	0.0067
IMP_Chlorides		1	-0.0386	0.0165	-0.0709	-0.0064	5.51	0.0189
IMP_FreeSulfurDioxid		1	0.0001	0.0000	0.0000	0.0002	6.81	0.0090
LabelAppeal		1	0.1591	0.0061	0.1471	0.1711	675.15	<.0001
IMP_STARS	2	1	0.3227	0.0143	0.2946	0.3508	506.66	<.0001
IMP_STARS	3	1	0.4417	0.0156	0.4111	0.4723	800.48	<.0001
IMP_STARS	4	1	0.5567	0.0217	0.5143	0.5992	660.30	<.0001
IMP_STARS	1	0	0.0000	0.0000	0.0000	0.0000	.	.
M_STARS		1	-1.0904	0.0182	-1.1261	-1.0547	3587.05	<.0001
IMP_TotalSulfurDioxi		1	0.0001	0.0000	0.0000	0.0001	11.63	0.0007
VolatileAcidity		1	-0.0307	0.0065	-0.0435	-0.0179	22.09	<.0001
Dispersion		1	0.0000	0.0001	0.0000	5.77E144		

Table(s) 16: Model GENMOD NB Maximum Likelihood, Class IMP_Stars

Selected Final Model

This model is a model using SAS GENMOD with zero inflated Poisson distribution and using a class of Label Appeal, IMP_Stars, M_Stars. This model demonstrated the best results than the previous models based upon the means comparison in table 21; however, the means error procedure in table 22 demonstrates this model has the lowest error mean than all the other models. An important improvement is using the class level information of Label Appeal, IMP Stars, and M Stars within this model. The specific values of Label Appeal, IMP_Stars, and M_Stars were chosen based upon frequency table procedure results shown in section Model Comparisons.

Model Information	
Data Set	WORK.FIXFILE
Distribution	Zero Inflated Poisson
Link Function	Log
Dependent Variable	TARGET

Class Level Information		
Class	**Levels**	**Values**
LabelAppeal	5	-2 -1 0 1 2
IMP_STARS	4	1 2 3 4
M_STARS	2	0 1

Table(s) 17: Model GENMOD Zero Inflated Poisson

Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance		41121.2589	
Scaled Deviance		41121.2589	
Pearson Chi-Square	13E3	5851.5101	0.4582
Scaled Pearson X2	13E3	5851.5101	0.4582
Log Likelihood		11036.5418	
Full Log Likelihood		-20560.6295	
AIC (smaller is better)		41169.2589	
AICC (smaller is better)		41169.3529	
BIC (smaller is better)		41348.2224	

Table 18: Model GENMOD Zero Inflated Poisson

Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.8884	0.0523	1.7859	1.9910	1302.49	<.0001
AcidIndex		1	-0.0324	0.0049	-0.0420	-0.0228	43.56	<.0001
IMP_Alcohol		1	0.0067	0.0014	0.0038	0.0095	21.56	<.0001
IMP_Chlorides		1	-0.0268	0.0168	-0.0598	0.0063	2.52	0.1122
IMP_FreeSulfurDioxid		1	0.0000	0.0000	-0.0000	0.0001	1.13	0.2876
LabelAppeal	-2	1	-1.0895	0.0462	-1.1801	-0.9989	555.17	<.0001
LabelAppeal	-1	1	-0.6396	0.0256	-0.6899	-0.5894	622.47	<.0001
LabelAppeal	0	1	-0.3501	0.0232	-0.3955	-0.3047	228.32	<.0001
LabelAppeal	1	1	-0.1597	0.0234	-0.2055	-0.1139	46.68	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
IMP_STARS	1	1	-0.3197	0.0222	-0.3631	-0.2762	207.69	<.0001
IMP_STARS	2	1	-0.1958	0.0200	-0.2350	-0.1566	95.72	<.0001
IMP_STARS	3	1	-0.0979	0.0202	-0.1375	-0.0583	23.46	<.0001
IMP_STARS	4	0	0.0000	0.0000	0.0000	0.0000	.	.
M_STARS	0	1	0.1840	0.0198	0.1452	0.2229	86.10	<.0001
M_STARS	1	0	0.0000	0.0000	0.0000	0.0000	.	.
IMP_TotalSulfurDioxi		1	-0.0000	0.0000	-0.0001	0.0000	0.01	0.9192
VolatileAcidity		1	-0.0156	0.0067	-0.0287	-0.0025	5.42	0.0199
Scale		0	1.0000	0.0000	1.0000	1.0000		

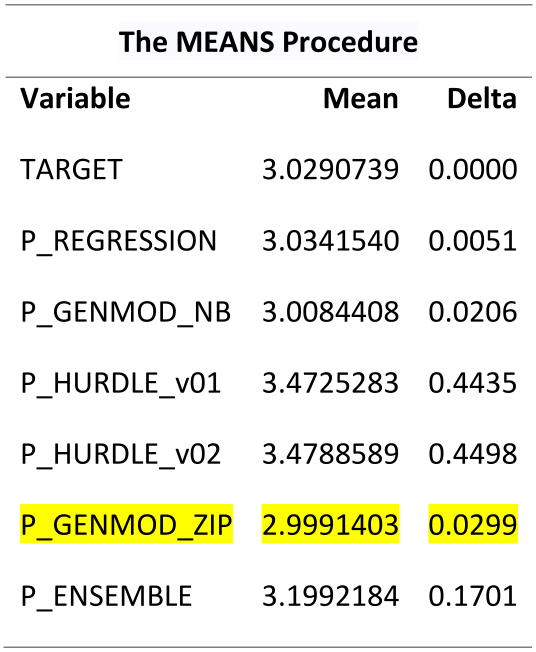
Table 19: Model GENMOD Zero Inflated Poisson

Analysis of Maximum Likelihood Zero Inflation Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-17.4014	0.2061	-17.8053	-16.9975	7130.57	<.0001
IMP_STARS	1	1	23.1841	0.4262	22.3488	24.0194	2959.32	<.0001
IMP_STARS	2	0	19.1728	0.0000	19.1728	19.1728	.	.
IMP_STARS	3	1	0.2419	3761.504	-7372.17	7372.655	0.00	0.9999
IMP_STARS	4	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-2	1	-3.7128	0.4431	-4.5813	-2.8442	70.20	<.0001
LabelAppeal	-1	1	-2.0514	0.2148	-2.4725	-1.6303	91.17	<.0001
LabelAppeal	0	1	-1.2912	0.2086	-1.6999	-0.8824	38.32	<.0001
LabelAppeal	1	1	-0.5541	0.2148	-0.9750	-0.1331	6.66	0.0099
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
M_STARS	0	1	-6.1129	0.4245	-6.9449	-5.2809	207.36	<.0001
M_STARS	1	0	0.0000	0.0000	0.0000	0.0000	.	.

Table 20: Model GENMOD Zero Inflated Poisson

Model Comparisons

The preceding models were scored and compared against the mean of the target value in order to compare performance to the actual target. This comparison using the Means procedure, table 21 may not be utilized as a sole determinate for comparison. It must be utilized in conjunction with the error means procedure in table 22. Based upon those combined results the Zero Inflated Poisson model demonstrates the best combined results based upon using both means procedures in tables 21 and 22. These results are further confirmed having a Kaggle score of 1.30570. The criteria for assessing goodness of fit values of AIC, AICC, and BIC are not being utilized as model comparisons because the values are not useful for comparing non-like models. An example where using goodness of fit values is within the same model, shifting variables in and out of the model.



The MEANS Procedure

Variable	Mean	Delta
TARGET	3.0290739	0.0000
P_REGRESSION	3.0341540	0.0051
P_GENMOD_NB	3.0084408	0.0206
P_HURDLE_v01	3.4725283	0.4435
P_HURDLE_v02	3.4788589	0.4498
P_GENMOD_ZIP	2.9991403	0.0299
P_ENSEMBLE	3.1992184	0.1701

Table 21: Means Procedure Comparison

The MEANS Procedure	
Variable	**Mean**
E_REGRESSION	1.0010942
E_GENMOD_NB	1.0017976
E_HURDLE_V01	1.1335678
E_HURDLE_V02	1.1516217
E_GENMOD_ZIP	0.9568581
E_ENSEMBLE	0.9976553

Table 22: Error Means Procedure Comparison

Class Variables

Adding class variables to modeling using SAS provides further opportunity for model improvement. We have chosen the following variables and corresponding values for the selected model, Zero Inflated Poisson.

Class Level Information		
Class	**Levels**	**Values**
LabelAppeal	5	-2 -1 0 1 2
IMP_STARS	4	1 2 3 4
M_STARS	2	0

One method for choosing which class variable values is to create a frequency table of the variable based upon the target value. This methodology provides an opportunity to measure which values provide the greatest percentage of impact in determining the target value as shown in tables 22, 23, and 24.

Table of M_STARS by TARGET_FLAG			
M_Stars	Target Flag		
	0	1	Total
0	696 5.44 7.38 25.46	8740 68.31 92.62 86.87	9436 73.75
1	2038 15.93 60.67 74.54	1321 10.32 39.33 13.13	3359 26.25
Total	2734 21.37	10061 78.63	12795 100.00

Table 22: Frequency table of M_Stars

Table of IMP_STARS by TARGET_FLAG			
IMP_Stars	Target Flag		
	0	1	Total
1	607 4.74 19.95 22.20	2435 19.03 80.05 24.20	3042 23.77
2	2127 16.62 30.70 77.80	4802 37.53 69.30 47.73	6929 54.15
3	0 0.00 0.00 0.00	2212 17.29 100.00 21.99	2212 17.29
4	0 0.00 0.00 0.00	612 4.78 100.00 6.08	612 4.78
Total	2734 21.37	10061 78.63	12795 100.00

Table 23: Frequency table of IMP_Stars

Table of Label Appeal by TARGET_FLAG			
Label Appeal	Target Flag		
	0	1	Total
-2	102 0.80 20.24 3.73	402 3.14 79.76 4.00	504 3.94
-1	671 5.24 21.40 24.54	2465 19.27 78.60 24.50	3136 24.51
0	1193 9.32 21.24 43.64	4424 34.58 78.76 43.97	5617 43.90
1	660 5.16 21.65 24.14	2388 18.66 78.35 23.74	3048 23.82
2	108 0.84 22.04 3.95	382 2.99 77.96 3.80	490 3.83
Total	2734 21.37	10061 78.63	12795 100.00

Table 24: Frequency table of Label Appeal

Conclusion

This analysis is a comparison five identified models, including multiple scoring attempts utilized for model testing, and optimization. Based upon comparisons how the means performed against the target value, error means procedure, and an Ensemble soring, the best model is the Zero Inflated Poisson model for predicting the target value wine case sales. The Zero Inflated Poisson model scored 1.30570 which is considered a good score based upon beating the benchmark scoring of the Ensemble model of 1.42774, and the benchmark Poisson model of 1.64638. The score does fall short of the Neural Net Model of 1.27465; although it is not a requirement for this project, it does point out opportunity for model improvement.

Near Term Recommendation

We should begin assessing wine case sales based upon using the following 9 variables:

- AcidIndex
- IMP_Alcohol
- IMP_Chlorides
- IMP_FreeSulfurDioxid
- LabelAppeal
- IMP_STARS
- M_STARS
- IMP_TotalSulfurDioxi
- VolatileAcidity

Long Term Recommendation

There is room for improvement upon this predictive model, with the long term goal to reach a model Kaggle score of less than 1.27465, thus improving the predictive accuracy of the model. The methodology utilized to build a long term model will conduct variable selection based upon decision tree analysis using either R, Angoss, or SAS Enterprise Miner. [1, 2]

Appendix

1. Decision Trees for Decision Making, HBR, https://hbr.org/1964/07/decision-trees-for-decision-making

2. Decision Tree, Wikipedia, https://en.wikipedia.org/wiki/Decision_tree

SAS Utilized for Output of Scored File as SAS Data Set

```
************************************************************************;
* Unit 03: Wine {Export .sas7bdat}                          *;
* Eric Lewis Section 55 Spring 2016                             *;
************************************************************************;

proc import datafile='/folders/wine_score_03.csv'
   dbms=csv
   out=scored
   replace;
run;
*proc print data=scored;
data '/folders/wine_score_03';
set scored;
run;
quit;
```

SAS Utilized for Logistic / Poisson Models

```
************************************************************************;
* PROC LOGISTIC                                              *;
************************************************************************;
proc logistic data=&FIXFILE.;
model TARGET_FLAG(ref="0") =
                   acidindex
                imp_alcohol
                imp_chlorides
        imp_freesulfurdioxide
                labelappeal
                imp_stars
                                M_STARS
                imp_totalsulfurdioxide
                volatileacidity;
output out=&FIXFILE. p=X_LOGIT_PROB;
run;

************************************************************************;
* PROC GENMOD 5 {Poisson}                                          *;
************************************************************************;
data &FIXFILE.;
set &TEMPFILE.;
run;
proc genmod data=&FIXFILE.;
class labelappeal imp_stars M_stars;
model TARGET =
                   acidindex
                imp_alcohol
                imp_chlorides
        imp_freesulfurdioxide
                labelappeal
                imp_stars
                                M_STARS
                imp_totalsulfurdioxide
                volatileacidity
                        /link=log dist=zip;
zeromodel IMP_STARS LabelAppeal M_STARS / link=logit;
output out=&FIXFILE. pred=X_GENMOD_ZIP pzero=X_GENMOD_PZERO;
run;
```

```
************************************************************************;
* Scoring Logistic                                                     *;
************************************************************************;

P_LOGIT_PROB =      1.9599                          +
                          AcidIndex                  *(-0.3836)          +
                          IMP_Alcohol                     *(-0.0208)          +
                          IMP_Chlorides                   *(-0.1497)          +
                          IMP_FreeSulfurDioxide      *(0.000592)         +
                          LabelAppeal                     *(-0.4644)          +
                          IMP_STARS                       *(2.5553)           +
                          M_STARS                              *(-4.3686)          +
                          IMP_TotalSulfurDioxide     *(0.000972)         +
                          VolatileAcidity                 *(-0.1822);

if P_LOGIT_PROB > 1000 then P_LOGIT_PROB = 1000;
if P_LOGIT_PROB < -1000 then P_LOGIT_PROB = -1000;
P_LOGIT_PROB = exp(P_LOGIT_PROB) / (1+exp(P_LOGIT_PROB));


************************************************************************;
* Scoring {Poisson}                                                    *;
************************************************************************;

P_ZERO_PROB =       -17.4014                                             +
                          (imp_stars in (1))         *(23.1841)          +
                          (imp_stars in (2))         *(19.1728)          +
                          (imp_stars in (3))         *(0.2419)           +
                          (LabelAppeal in (-2))      *(-3.7128)          +
                          (LabelAppeal in (-1))      *(-2.0514)          +
                          (LabelAppeal in (0))       *(-1.2912)          +
                          (LabelAppeal in (1))       *(-0.5541)          +
                          (M_stars in (0))                *(-6.1129);

if P_ZERO_PROB > 1000 then P_ZERO_PROB = 1000;
if P_ZERO_PROB < -1000 then P_ZERO_PROB = -1000;
P_ZERO_PROB = exp(P_ZERO_PROB) / (1+exp(P_ZERO_PROB));
```

```
************************************************************************;
* Scoring {Poisson}                                                    *;
************************************************************************;
P_GENMOD_ZIP =     1.8884                          +
                            AcidIndex                    *(-0.0324)        +
                            IMP_Alcohol                        *(0.0067)         +
                            IMP_Chlorides                      *(-0.0268)        +
                            IMP_FreeSulfurDioxide        *(0.0000)         +
                            (LabelAppeal in (-2))        *(-1.0895)        +
                            (LabelAppeal in (-1))        *(-0.6396)        +
                            (LabelAppeal in (0))         *(-0.3501)        +
                            (LabelAppeal in (1))         *(-0.1597)        +
                            (imp_stars in (1))           *(-0.3197)        +
                            (imp_stars in (2))           *(-0.1958)        +
                            (imp_stars in (3))           *(-0.0979)        +
                            (M_stars in (0))                   *(0.1840)         +
                            IMP_TotalSulfurDioxide       *(-0.0000)        +
                            VolatileAcidity                    *(-0.0156);

P_GENMOD_ZIP       = exp(P_GENMOD_ZIP);
P_GENMOD_ZIP       = P_GENMOD_ZIP*(1-P_ZERO_PROB);

P_GENMOD_ZIP = round( P_GENMOD_ZIP, 0.01 );
X_GENMOD_ZIP = round( X_GENMOD_ZIP, 0.01 );
```

SAS Utilized for Scoring

```
  ************************************************************************;
* Unit 03: WINE PROJECT  {Score}                                *;
* Eric Lewis Section 55 Spring 2016                              *;
************************************************************************;

%let PATH = /folders/myfolders/Pred411/Data;
%let NAME = WINE;
%let LIB = &NAME..;

libname &NAME. "&PATH.";

%let INFILE = &LIB.WINE_TEST;
%let TEMPFILE = TEMPFILE;

data &TEMPFILE.;
set &INFILE.;

data validate;
   set &TEMPFILE.;

           IMP_ResidualSugar = ResidualSugar;
           IMP_Chlorides = Chlorides;
           IMP_FreeSulfurDioxide = FreeSulfurDioxide;
           IMP_TotalSulfurDioxide = TotalSulfurDioxide;
           IMP_pH = pH;
           IMP_Sulphates = Sulphates;
           IMP_Alcohol = Alcohol;
           IMP_STARS = STARS;
           M_STARS = 0;

           if missing(ResidualSugar) then IMP_ResidualSugar = 5.4187331;
           if missing(Chlorides) then IMP_Chlorides = 0.0548225;
           if missing(FreeSulfurDioxide) then IMP_FreeSulfurDioxide = 30.8455713;
           if missing(TotalSulfurDioxide) then IMP_TotalSulfurDioxide = 120.7142326;
           if missing(pH) then IMP_pH = 3.2076282;
           if missing(Sulphates) then IMP_Sulphates = 0.5271118;
           if missing(Alcohol) then IMP_Alcohol = 10.4892363;
           if missing(STARS) then do; IMP_STARS = 2; M_STARS = 1; end;

           if IMP_TotalSulfurDioxide     < -330 then IMP_TotalSulfurDioxide = -330;
           if IMP_TotalSulfurDioxide     > 630  then IMP_TotalSulfurDioxide = 630;


data score;
   set validate;
```

```
P_ZERO_PROB =       -17.4014                                                +
                            (imp_stars in (1))          *(23.1841)          +
                            (imp_stars in (2))          *(19.1728)          +
                            (imp_stars in (3))          *(0.2419)           +
                            (LabelAppeal in (-2))       *(-3.7128)          +
                            (LabelAppeal in (-1))       *(-2.0514)          +
                            (LabelAppeal in (0))        *(-1.2912)          +
                            (LabelAppeal in (1))        *(-0.5541)          +
                            (M_stars in (0))                    *(-6.1129);

if P_ZERO_PROB > 1000 then P_ZERO_PROB = 1000;
if P_ZERO_PROB < -1000 then P_ZERO_PROB = -1000;
P_ZERO_PROB = exp(P_ZERO_PROB) / (1+exp(P_ZERO_PROB));

P_GENMOD_ZIP =      1.8884                            +
                            AcidIndex                   *(-0.0324)          +
                            IMP_Alcohol                         *(0.0067)           +
                            IMP_Chlorides                       *(-0.0268)          +
                            IMP_FreeSulfurDioxide       *(0.0000)           +
                            (LabelAppeal in (-2))       *(-1.0895)          +
                            (LabelAppeal in (-1))       *(-0.6396)          +
                            (LabelAppeal in (0))        *(-0.3501)          +
                            (LabelAppeal in (1))        *(-0.1597)          +
                            (imp_stars in (1))          *(-0.3197)          +
                            (imp_stars in (2))          *(-0.1958)          +
                            (imp_stars in (3))          *(-0.0979)          +
                            (M_stars in (0))                    *(0.1840)           +
                            IMP_TotalSulfurDioxide      *(-0.0000)          +
                            VolatileAcidity                     *(-0.0156);

P_GENMOD_ZIP        = exp(P_GENMOD_ZIP);
P_GENMOD_ZIP        = P_GENMOD_ZIP*(1-P_ZERO_PROB);
P_TARGET = round( P_GENMOD_ZIP, 1 );

 keep index P_TARGET;

proc print data=score;

proc export data=score
   outfile='/folders/wine_score_03.csv'
   dbms=csv
   replace;

run;
```

SAS Utilized for Analysis

```
************************************************************************;
* Unit 03: Wine Sales PROJECT  {Analysis}                        *;
* Eric Lewis                                 *;
************************************************************************;

%let PATH = /folders/myfolders/Pred411/Data;
%let NAME = P411;
%let LIB = &NAME..;

libname &NAME. "&PATH.";

%let INFILE = &LIB.WINE;
%let TEMPFILE = TEMPFILE;
%let FIXFILE  = FIXFILE;

*proc print data=&INFILE.(obs=5);
*run;
*proc contents data=&INFILE.;
*run;

************************************************************************;
* Find means, missing data                                   *;
************************************************************************;

*proc means data=&INFILE. n nmiss mean std;
*var _numeric_ ;
*run;

************************************************************************;
* Data Exploration: Visual Analysis                             *;
************************************************************************;

* proc univariate data=&INFILE. normal;
*    var Target;
*    histogram;

************************************************************************;
* Impute missing data w/means                                  *;
************************************************************************;

   data &TEMPFILE.;
   set &INFILE.;

   TARGET_FLAG = ( TARGET > 0 );
   TARGET_AMT = TARGET - 1;
```

```
  if TARGET_FLAG = 0 then TARGET_AMT = .;

       IMP_ResidualSugar = ResidualSugar;
       IMP_Chlorides = Chlorides;
       IMP_FreeSulfurDioxide = FreeSulfurDioxide;
       IMP_TotalSulfurDioxide = TotalSulfurDioxide;
       IMP_pH = pH;
       IMP_Sulphates = Sulphates;
       IMP_Alcohol = Alcohol;
       IMP_STARS = STARS;
       M_STARS = 0;

       if missing(ResidualSugar) then IMP_ResidualSugar = 5.4187331;
       if missing(Chlorides) then IMP_Chlorides = 0.0548225;
       if missing(FreeSulfurDioxide) then IMP_FreeSulfurDioxide = 30.8455713;
       if missing(TotalSulfurDioxide) then IMP_TotalSulfurDioxide = 120.7142326;
       if missing(pH) then IMP_pH = 3.2076282;
       if missing(Sulphates) then IMP_Sulphates = 0.5271118;
       if missing(Alcohol) then IMP_Alcohol = 10.4892363;
       if missing(STARS) then do; IMP_STARS = 2; M_STARS = 1; end;

       if IMP_TotalSulfurDioxide    < -330 then IMP_TotalSulfurDioxide = -330;
       if IMP_TotalSulfurDioxide    > 630  then IMP_TotalSulfurDioxide = 630;

keep    TARGET
              TARGET_FLAG
              TARGET_AMT
    acidindex
    citricacid
    density
    fixedacidity
    IMP_ResidualSugar
    IMP_Sulphates
    IMP_pH
    imp_alcohol
    imp_chlorides
    imp_freesulfurdioxide
    labelappeal
    IMP_STARS
              M_STARS
    imp_totalsulfurdioxide
    volatileacidity;

run;



*proc freq data=&TEMPFILE.;
```

```
*table (M_STARS IMP_STARS)*TARGET_FLAG /missing;
*run;

*proc freq data=&TEMPFILE.;
*Table (labelappeal)*TARGET_FLAG /missing;
*run;


* proc univariate data=&TEMPFILE. normal;
*    var IMP_TotalSulfurDioxide;
*    histogram;

*    proc means data=&TEMPFILE. n nmiss mean var;
*    var acidindex
        imp_alcohol
        imp_chlorides
        citricacid
        density
        fixedacidity
        imp_freesulfurdioxide
        labelappeal
        imp_residualsugar
        imp_stars
        imp_sulphates
        imp_totalsulfurdioxide
        volatileacidity
        imp_ph;

************************************************************************;
* Data Preparation: Variable Selection                                 *;
************************************************************************;

*proc reg data=&TEMPFILE.;
*model TARGET =
        acidindex
        imp_alcohol
        imp_chlorides
        citricacid
        density
        fixedacidity
        imp_freesulfurdioxide
        labelappeal
        imp_residualsugar
        imp_stars
        imp_sulphates
        imp_totalsulfurdioxide
        volatileacidity
        imp_ph;
```

```
*/selection=forward;
*/selection=backward;
*/selection=stepwise;
*run;
*quit;

************************************************************************;
* Correlation of all numeric values                           *;
************************************************************************;

*proc corr data=&TEMPFILE.;
*    var TARGET
       acidindex
       imp_alcohol
       imp_chlorides
       citricacid
       density
       fixedacidity
       imp_freesulfurdioxide
       labelappeal
       imp_residualsugar
       imp_stars
       imp_sulphates
       imp_totalsulfurdioxide
       volatileacidity
       imp_ph;


************************************************************************;
* Frequency of the stars and label appeal for model selection {Class}    *;
************************************************************************;

*proc freq data=&TEMPFILE.;
*    tables target*M_stars;

*proc freq data=&TEMPFILE.;
*    tables target*imp_stars;

*proc freq data=&TEMPFILE.;
*    tables target*labelappeal;



************************************************************************;
* Model Building  {}    *;
************************************************************************;

data &FIXFILE.;
```

```
set &TEMPFILE.;
run;

**********************************************************************************;
* PROC REG                                        *;
**********************************************************************************;

proc reg data=&FIXFILE.;
model TARGET =
                acidindex
                imp_alcohol
                imp_chlorides
        imp_freesulfurdioxide
                labelappeal
                imp_stars
                                M_STARS
                imp_totalsulfurdioxide
                volatileacidity;
*                               /selection = stepwise;
        output out=&FIXFILE. p=X_REGRESSION;
run;
quit;

**********************************************************************************;
* PROC GENMOD 1                                        *;
**********************************************************************************;


proc genmod data=&FIXFILE.;
model TARGET =
                acidindex
                imp_alcohol
                imp_chlorides
        imp_freesulfurdioxide
                labelappeal
                imp_stars
                                M_STARS
                imp_totalsulfurdioxide
                volatileacidity
                                /link=log dist=nb;
output out=&FIXFILE. p=X_GENMOD_NB;
run;

**********************************************************************************;
* PROC LOGISTIC                                      *;
**********************************************************************************;

proc logistic data=&FIXFILE.;
```

```
model TARGET_FLAG(ref="0") =
                  acidindex
                imp_alcohol
                imp_chlorides
        imp_freesulfurdioxide
                labelappeal
                imp_stars
                                M_STARS
                imp_totalsulfurdioxide
                volatileacidity;
output out=&FIXFILE. p=X_LOGIT_PROB;
run;

**********************************************************************;
* PROC GENMOD 2   {Hurdle Model 1}                                   *;
**********************************************************************;


proc genmod data=&FIXFILE.;
model TARGET =
                  acidindex
                imp_alcohol
                imp_chlorides
        imp_freesulfurdioxide
                labelappeal
                imp_stars
                                M_STARS
                imp_totalsulfurdioxide
                volatileacidity
                                /link=log dist=nb;
output out=&FIXFILE. p=X_GENMOD_HURDLE_v01;
run;

**********************************************************************;
* PROC GENMOD 3                                                      *;
**********************************************************************;

proc genmod data=&FIXFILE.;
class imp_stars(ref="1");
model TARGET =
                  acidindex
                imp_alcohol
                imp_chlorides
        imp_freesulfurdioxide
                labelappeal
                imp_stars
                                M_STARS
                imp_totalsulfurdioxide
```

```
                    volatileacidity
                                        /link=log dist=nb;
output out=&FIXFILE. p=X_GENMOD_HURDLE_v02;
run;

**********************************************************************;
* PROC GENMOD 4                                           *;
**********************************************************************;

*proc genmod data=&FIXFILE.;
*model TARGET =
                       acidindex
                    imp_alcohol
                    imp_chlorides
          imp_freesulfurdioxide
                    labelappeal
                    imp_stars
                                        M_STARS
                    imp_totalsulfurdioxide
                    volatileacidity
                              /link=log dist=zip;
*zeromodel IMP_STARS LabelAppeal M_STARS / link=logit;
*output out=&FIXFILE. pred=X_GENMOD_ZIP pzero=X_GENMOD_PZERO;
*run;

**********************************************************************;
* PROC GENMOD 5 {Poisson}                                         *;
**********************************************************************;

data &FIXFILE.;
set &TEMPFILE.;
run;
proc genmod data=&FIXFILE.;
class labelappeal imp_stars M_stars;
model TARGET =
                       acidindex
                    imp_alcohol
                    imp_chlorides
          imp_freesulfurdioxide
                    labelappeal
                    imp_stars
                                        M_STARS
                    imp_totalsulfurdioxide
                    volatileacidity
                              /link=log dist=zip;
zeromodel IMP_STARS LabelAppeal M_STARS / link=logit;
output out=&FIXFILE. pred=X_GENMOD_ZIP pzero=X_GENMOD_PZERO;
run;
```

```
*********************************************************************;
* SCORE MODELS                                    *;
*********************************************************************;


data SCOREFILE;
set &FIXFILE.;

*  Regression                         *;

P_REGRESSION =      3.46236                  +
                           AcidIndex                    *(-0.20007)           +
                           IMP_Alcohol                         *(0.01246)            +
                           IMP_Chlorides                       *(-0.11742)           +
                           IMP_FreeSulfurDioxide        *(0.00028171)         +
                           LabelAppeal                         *(0.46626)            +
                           IMP_STARS                           *(0.78030)            +
                           M_STARS                                    *(-2.24712)           +
                           IMP_TotalSulfurDioxide       *(0.00024441)         +
                           VolatileAcidity                     *(-0.09693);

*    GENMOD 1                         *;

P_GENMOD_NB =       1.4480                   +
                           AcidIndex                    *(-0.0804)            +
                           IMP_Alcohol                         *(0.0035)             +
                           IMP_Chlorides                       *(-0.0368)            +
                           IMP_FreeSulfurDioxide        *(0.0001)             +
                           LabelAppeal                         *(0.1587)             +
                           IMP_STARS                           *(0.1882)             +
                           M_STARS                                    *(-1.0246)            +
                           IMP_TotalSulfurDioxide       *(0.0001)             +
                           VolatileAcidity                     *(-0.0312);

P_GENMOD_NB = exp(P_GENMOD_NB);

*  Logistic                               *;

P_LOGIT_PROB =      1.9599                   +
                           AcidIndex                    *(-0.3836)            +
                           IMP_Alcohol                         *(-0.0208)            +
                           IMP_Chlorides                       *(-0.1497)            +
                           IMP_FreeSulfurDioxide        *(0.000592)           +
                           LabelAppeal                         *(-0.4644)            +
                           IMP_STARS                           *(2.5553)             +
                           M_STARS                                  *(-4.3686)          +
                           IMP_TotalSulfurDioxide       *(0.000972)           +
                           VolatileAcidity                        *(-0.1822);

if P_LOGIT_PROB > 1000 then P_LOGIT_PROB = 1000;
if P_LOGIT_PROB < -1000 then P_LOGIT_PROB = -1000;
P_LOGIT_PROB = exp(P_LOGIT_PROB) / (1+exp(P_LOGIT_PROB));

*   GENMOD 2                                  *;


P_GENMOD_HURDLE_v01 =          1.4480                  +
                           AcidIndex                    *(-0.0804 )           +
                           IMP_Alcohol                            *(0.0035)             +
                           IMP_Chlorides                          *(-0.0368)            +
                           IMP_FreeSulfurDioxide        *(0.0001 )            +
                           LabelAppeal                            *(0.1587)             +
                           IMP_STARS                              *(0.1882)             +
                           M_STARS                                  *(-1.0246)          +
                           IMP_TotalSulfurDioxide       *(0.0001)             +
                           VolatileAcidity                        *(-0.0312);
P_GENMOD_HURDLE_v01 = exp(P_GENMOD_HURDLE_v01);

*   GENMOD 3                                  *;

P_GENMOD_HURDLE_v02 =          1.5568                 +
                           AcidIndex                    *(-0.0795)            +
                           IMP_Alcohol                            *(0.0038)             +
                           IMP_Chlorides                          *(-0.0386)            +
                           IMP_FreeSulfurDioxide        *(0.0001)             +
                           LabelAppeal                            *(0.1591)             +
                           (IMP_STARS=2)                          *(0.3227)             +
                           (IMP_STARS=3)                          *(0.4417)             +
                           (IMP_STARS=4)                          *(0.5567)             +
                            M_STARS                               *(-1.0904)            +
                           IMP_TotalSulfurDioxide       *(0.0001)             +
                           VolatileAcidity                        *(-0.0307);

P_GENMOD_HURDLE_v02 = exp(P_GENMOD_HURDLE_v02);

P_HURDLE_v01 = P_LOGIT_PROB * (P_GENMOD_HURDLE_v01+1);
P_HURDLE_v02 = P_LOGIT_PROB * (P_GENMOD_HURDLE_v02+1);


*    GENMOD 4   {Bottom Parameter Estimates Chart                          *;

*P_ZERO_PROB =       2.5001                        +
                           LabelAppeal                            *(0.7497)             +
                           IMP_STARS                         *(-4.1348)          +
                           M_STARS                                 *(6.2279);

*if P_ZERO_PROB > 1000 then P_ZERO_PROB = 1000;
*if P_ZERO_PROB < -1000 then P_ZERO_PROB = -1000;
*P_ZERO_PROB = exp(P_ZERO_PROB) / (1+exp(P_ZERO_PROB));

*    GENMOD 4   {Top Parameter Estimates Chart                    *;

*P_GENMOD_ZIP =      1.2780                  +
                           AcidIndex                   *(-0.0329)          +
                           IMP_Alcohol                       *(0.0065)           +
                           IMP_Chlorides                     *(-0.0262)          +
                           IMP_FreeSulfurDioxide       *(0.0000)           +
                           LabelAppeal                       *(0.2331)           +
                           IMP_STARS                         *(0.1057)           +
                           M_STARS                                 *(-0.1816)          +
                           IMP_TotalSulfurDioxide      *(-0.0000)          +
                           VolatileAcidity                   *(-0.0157);

*P_GENMOD_ZIP       = exp(P_GENMOD_ZIP);
*P_GENMOD_ZIP       = P_GENMOD_ZIP*(1-P_ZERO_PROB);

*P_GENMOD_ZIP = round( P_GENMOD_ZIP, 0.01 );
*X_GENMOD_ZIP = round( X_GENMOD_ZIP, 0.01 );


*    GENMOD 5   {Poisson}                       *;

P_ZERO_PROB =        -17.4014                                          +
                           (imp_stars in (1))          *(23.1841)          +
                           (imp_stars in (2))          *(19.1728)          +
                           (imp_stars in (3))          *(0.2419)           +
                           (LabelAppeal in (-2))       *(-3.7128)          +
                           (LabelAppeal in (-1))       *(-2.0514)          +
                           (LabelAppeal in (0))        *(-1.2912)          +
                           (LabelAppeal in (1))        *(-0.5541)          +
                           (M_stars in (0))                  *(-6.1129);

if P_ZERO_PROB > 1000 then P_ZERO_PROB = 1000;
if P_ZERO_PROB < -1000 then P_ZERO_PROB = -1000;
P_ZERO_PROB = exp(P_ZERO_PROB) / (1+exp(P_ZERO_PROB));

*    GENMOD 5   {Top Parameter Estimates Chart                    *;

P_GENMOD_ZIP =       1.8884                    +
                           AcidIndex                   *(-0.0324)          +
                           IMP_Alcohol                       *(0.0067)           +
                           IMP_Chlorides                   *(-0.0268)          +
                           IMP_FreeSulfurDioxide     *(0.0000)           +
                           (LabelAppeal in (-2))     *(-1.0895)          +
                           (LabelAppeal in (-1))     *(-0.6396)          +
                           (LabelAppeal in (0))      *(-0.3501)          +
                           (LabelAppeal in (1))      *(-0.1597)          +
                           (imp_stars in (1))        *(-0.3197)          +
                           (imp_stars in (2))        *(-0.1958)          +
                           (imp_stars in (3))        *(-0.0979)          +
                           (M_stars in (0))                *(0.1840)           +
                           IMP_TotalSulfurDioxide    *(-0.0000)          +
                           VolatileAcidity                 *(-0.0156);

P_GENMOD_ZIP        = exp(P_GENMOD_ZIP);
P_GENMOD_ZIP        = P_GENMOD_ZIP*(1-P_ZERO_PROB);

P_GENMOD_ZIP = round( P_GENMOD_ZIP, 0.01 );
X_GENMOD_ZIP = round( X_GENMOD_ZIP, 0.01 );

P_ENSEMBLE = (P_REGRESSION + P_GENMOD_NB + P_HURDLE_v01 + P_HURDLE_v02 +
P_GENMOD_ZIP)/5;

P_REGRESSION        = round(P_REGRESSION      , 1);
P_GENMOD_NB         = round(P_GENMOD_NB             , 1);
P_HURDLE_v01        = round(P_HURDLE_v01      , 1);
P_HURDLE_v02        = round(P_HURDLE_v02      , 1);
P_ENSEMBLE          = round(P_ENSEMBLE              , 1);
P_GENMOD_ZIP        = round(P_GENMOD_ZIP      , 1);

run;

*proc print data=SCOREFILE(obs=25);
*var P_ZERO_PROB X_GENMOD_PZERO ;
*run;

*proc print data=SCOREFILE(obs=25);
*var P_GENMOD_ZIP X_GENMOD_ZIP P_ZERO_PROB X_GENMOD_PZERO;
*run;

*proc print data=SCOREFILE(obs=25);
*var TARGET P_REGRESSION P_GENMOD_NB P_HURDLE_V01 P_HURDLE_V02 P_GENMOD_ZIP
P_ENSEMBLE ;
*run;

proc means data=SCOREFILE mean;
var TARGET P_REGRESSION P_GENMOD_NB P_HURDLE_V01 P_HURDLE_V02 P_GENMOD_ZIP
P_ENSEMBLE ;
run;
```

```
data SCOREFILE;
set SCOREFILE;

if TEST_FLAG = 0 then delete;

        E_REGRESSION        = abs(TARGET - P_REGRESSION);
        E_GENMOD_NB         = abs(TARGET - P_GENMOD_NB);
        E_HURDLE_V01        = abs(TARGET - P_HURDLE_V01);
        E_HURDLE_V02        = abs(TARGET - P_HURDLE_V02);
        E_GENMOD_ZIP        = abs(TARGET - P_GENMOD_ZIP);
        E_ENSEMBLE          = abs(TARGET - P_ENSEMBLE);

run;

proc means data=SCOREFILE mean;
var E_REGRESSION E_GENMOD_NB E_HURDLE_v01 E_HURDLE_v02 E_GENMOD_ZIP E_ENSEMBLE ;
run;


proc univariate data=SCOREFILE;
var E_ENSEMBLE;
histogram;
run;

************************************************************************;
* End                                                                  *;
************************************************************************;

*
```