Ji Jung Predict 411 SEC 56

Wine Sales Problem

Introduction

This data set contains information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties, with others relating to qualitative ratings by individuals. The target variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling. A large wine manufacture is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If it is possible to predict the number of cases, the manufacture will be able to adjust their wine offerings with the goal to maximize sales.

For this assignment, building Poisson and Negative Binomial models that will predict the target number of cases ordered for each wine will be made. Furthermore, Zero-Inflated Poisson and Negative Binomial will also be constructed and compared to choose the best performing model.

Data Exploration and Preparation

Background check of the dataset is the first step. Overall plan for Exploratory Data Analysis will be;

- Obtain histograms and statistical descriptions for target variable.
    - Examine means and variances to check assumption of equality for Poisson or Negative Binomial distribution.
    - Examine histograms for indications of zero-inflation.
- Obtain histograms for all continuous variables.
    - Examine distributions with potential transformations for modeling stage.
- Obtain frequency counts for all categorical variables for variability.

Table 1 Data Dictionary

VARIABLE NAME	DEFINITION
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average
Alcohol	Alcohol Content
Chlorides	Chloride content of wine
CitricAcid	Citric Acid Content
Density	Density of Wine
FixedAcidity	Fixed Acidity of Wine
FreeSulfurDioxide	Sulfur Dioxide content of wine
LabelAppeal	Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customers don't like the design.

ResidualSugar	Residual Sugar of wine
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor
Sulphates	Sulfate content of wine
TotalSulfurDioxide	Total Sulfur Dioxide of Wine
VolatileAcidity	Volatile Acid content of wine
pH	pH of wine

From examining data dictionary under Table 1, variables LabelAppeal and STARS are suspected to be categorical.

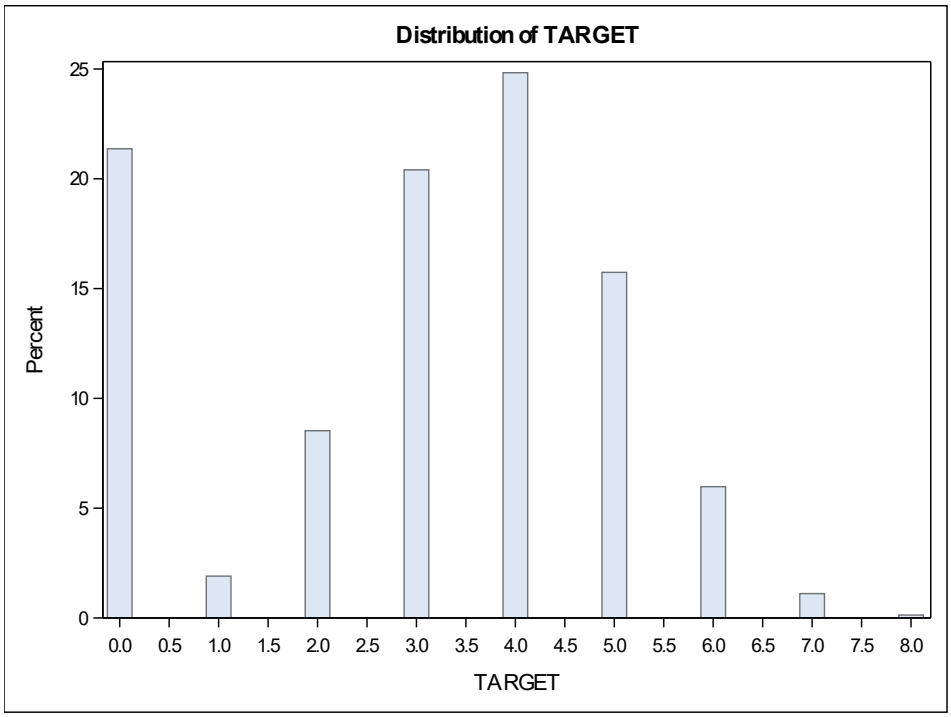
Mean and variance of the target variable is calculated under Table 2.

Table 2 Basic Statistical Measures of Target Variable

Location		Variability	
Mean	3.029074	**Std Deviation**	1.92637
Median	3.000000	**Variance**	3.71089
Mode	4.000000	**Range**	8.00000
		Interquartile Range	2.00000

Since the variance and the mean of Target variable are not equal, the variable is in violation for Poisson distribution. However, the variance is larger than the mean, thus Negative binomial regression will not be in violation.

Figure 1 Histogram of Target



From the histogram of TARGET under Figure 1, the variable passes tests for normality but with signs of zero-inflated. At this stage, modeling approach based on this observation should not be made for all normal OLS regression, Poisson, and Negative Binomial procedures will be examined for differences in performance.

Variables are examined for missing counts for preparation phase under Table 3.

Table 3 Number of Observations and Missing Counts

Variable	N	N Miss	Mean	Variance
AcidIndex	12795	0	7.7727237	1.7527810
Alcohol	12142	653	10.4892363	13.8966348
Chlorides	12157	638	0.0548225	0.1014214
CitricAcid	12795	0	0.3084127	0.7431816
Density	12795	0	0.9942027	0.000704247
FixedAcidity	12795	0	7.0757171	39.9126188
FreeSulfurDioxide	12148	647	30.8455713	22116.02
LabelAppeal	12795	0	-0.0090660	0.7940400
ResidualSugar	12179	616	5.4187331	1139.02
STARS	9436	3359	2.0417550	0.8145785
Sulphates	11585	1210	0.5271118	0.8688650
TotalSulfurDioxide	12113	682	120.7142326	53783.74
VolatileAcidity	12795	0	0.3241039	0.6146783
pH	12400	395	3.2076282	0.4619745

Table 3 shows many variables are missing observations from the dataset. These variables will be imputed with the mean value and create an indicator for missing data. For categorical variables, the mean value will be rounded. Since many variables are missing observations, in order to carry a variable forward into the modeling phase, missing indicator variables will be used as well.

There is one variable as an obvious indication of how much the wine is liked, represented by STARS. However, it is many missing values. Hence, the correlation and frequency values will be studied as the indicator of imputation for stars.

Next, the means statistics are examined for each variable for the count of the target variable. There are differences between many variables from zero to eight cases purchased. Not a single variable stand out as unusable other than the 25% missing observations in STARS. It will not be eliminated for it is logically and theoretically sound to the most valuable variable for indicating desire for purchase.

Two categorical variables, LabelAppeal and STARS, are examined for frequency tables against TARGET and their proportional variations are throughout within range.

From histograms of each variable, followings are noticed;

- All variable passes tests for normality at the significant level.
- Many variables report significant spikes in the center.

- After imputation, STARS is left-centered.

Spikes in the center from continuous variables may be a cause for a log-scaling of the variables. However, scaling is not done at this stage for it adds complexity to interpretations of models.

Simple Pearson correlations between imputed variables are shown under Table 4.

Table 4 Pearson Correlation with TARGET

Pearson Correlation Coefficients, N = 12795 Prob > |r| under H0: Rho=0								
TARGET	i_imp_stars	imp_stars	LabelAppeal	AcidIndex	VolatileAcidity	imp_alcohol	imp_totalsulfurdioxide	FixedAcidity
	-0.57158	0.40013	0.35650	-0.24605	-0.08879	0.06043	0.05010	-0.04901
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

Pearson Correlation Coefficients, N = 12795 Prob > |r| under H0: Rho=0							
TARGET	imp_freesulfurdioxide	imp_chlorides	imp_sulphates	Density	imp_residualsugar	imp_ph	CitricAcid
	0.04269	-0.03724	-0.03691	-0.03552	0.01607	-0.00928	0.00868
	<.0001	<.0001	<.0001	<.0001	0.0691	0.2939	0.3260

From the correlation report, there is a need to throw out any variables without at least an 0.1 level of correlation regardless of transformations. There are high correlations between categorical review variables, LabelAppeal and STARS. The highest correlation is with the indicator of imputation for imp_STARS. Of remaining variables with significant and above 0.1 threshold, continued examinations will be made for the modeling stage. Hence variables AcidIndex, LabelAppeal, imp_STARS, and i_imp_STARS are left.

With one physical measure and two qualitative measures, building the best-fitting model seems straight forward. Further remarks will be made on conclusions regarding variables and potential meanings and strategies to a business owner.

Categorical variables LabelAppeal, STARS and i_imp_STARS on a frequency table against the dependent variable, a linear relationship seems to occur throughout the range.

Model construction will ultimately come down an ordinary least-square regression model. Nonetheless, Poisson and Negative Binomial models will be constructed to assist in selecting parameters. Even with a limited range of variables, different combinations of variables and the respective performance to TARGET will be made.

SAS procedure, Genmod, does not have a method for automatic variable selection. There is a need to narrow down by attempting to explore variables that make sense to incorporate. This may be difficult with the Zero-Inflated model as frequency tables are needed to examine which variables conditionally contribute to the probability that are observed to a zero count with TARGET.

Prior to actual building models, it can be assumed that both Poisson and Negative Binomial models will likely result similarly for the variance is close to the mean value.

Build Models

Poisson

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e$$

Table 5 Poisson Model Analysis Of Maximum Likelihood Parameter

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.3752	0.0476	1.2820	1.4684	836.24	<.0001
AcidIndex		1	-0.0814	0.0045	-0.0902	-0.0726	328.69	<.0001
LabelAppeal	-2	1	-0.6958	0.0424	-0.7789	-0.6126	269.03	<.0001
LabelAppeal	-1	1	-0.4597	0.0250	-0.5086	-0.4107	338.98	<.0001
LabelAppeal	0	1	-0.2702	0.0228	-0.3149	-0.2254	139.87	<.0001
LabelAppeal	1	1	-0.1377	0.0232	-0.1831	-0.0923	35.38	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
imp_stars	1	1	-0.5647	0.0216	-0.6071	-0.5224	682.89	<.0001
imp_stars	2	1	-0.2431	0.0199	-0.2820	-0.2041	149.78	<.0001
imp_stars	3	1	-0.1207	0.0202	-0.1602	-0.0811	35.77	<.0001
imp_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
i_imp_stars	0	1	1.0926	0.0182	1.0569	1.1283	3599.71	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Table 6 Poisson Model Criteria For Assessing Goodness-Of-Fit

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13700.3624	1.0716
Scaled Deviance	13E3	13700.3624	1.0716
Pearson Chi-Square	13E3	11331.6014	0.8863

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Scaled Pearson X2	13E3	11331.6014	0.8863
Log Likelihood		8775.9792	
Full Log Likelihood		-22821.1920	
AIC (smaller is better)		45662.3841	
AICC (smaller is better)		45662.4013	
BIC (smaller is better)		45736.9522	

Table 7 Poisson Model Variables

In Model	In Data
Y is	TARGET
X_1 is	AcidIndex
X_2 is	LabelAppeal
X_3 is	imp_STARS
X_4 is	i_imp_STARS

The exponentiated AcidIndex coefficient is multiplicative used in calculating the estimated target hen it increases by one unit. For categorical variables, the exponentiated coefficient is the multiplicative term to the base level for each variable. The exponentiated intercept is the baseline rate, and all other estimates will be relative to it.

The effect of one unit increase in AcidIndex results in 8% decrease in the expected number of cases purchased.

LabelAppeal has a base level of 2 as the highest rating, following can be interpreted;

- negative two rating: 50% decrease in the expected number of cases purchased
- negative one rating: 36% decrease in the expected number of cases purchased
- zero rating: 23% decrease in the expected number of cases purchased
- positive one rating: 12% decrease in the expected number of cases purchased

Each can be interpreted as decrease from obtaining a 2 rating.

STARS has a base level of 4 as the highest rating, following can be interpreted;

- One rating: 43% decrease in the expected number of cases purchased
- Two rating: 21% decrease in the expected number of cases purchased
- Three rating: 11% decrease in the expected number of cases purchased

Each can be interpreted as decrease from obtaining a 4 rating.

As an example, if a wine is given one STARS rating, our model indicates a 98% increase in the expected number of cases to be purchased.

Negative Binomial

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e$$

Table 8 Negative Binomial Model Analysis Of Maximum Likelihood Parameter

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	1.3752	0.0476	1.2820	1.4684	836.24	<.0001
AcidIndex		1	-0.0814	0.0045	-0.0902	-0.0726	328.69	<.0001
LabelAppeal	-2	1	-0.6958	0.0424	-0.7789	-0.6126	269.03	<.0001
LabelAppeal	-1	1	-0.4597	0.0250	-0.5086	-0.4107	338.98	<.0001
LabelAppeal	0	1	-0.2702	0.0228	-0.3149	-0.2254	139.87	<.0001
LabelAppeal	1	1	-0.1377	0.0232	-0.1831	-0.0923	35.38	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
imp_stars	1	1	-0.5647	0.0216	-0.6071	-0.5224	682.89	<.0001
imp_stars	2	1	-0.2431	0.0199	-0.2820	-0.2041	149.78	<.0001
imp_stars	3	1	-0.1207	0.0202	-0.1602	-0.0811	35.77	<.0001
imp_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
i_imp_stars	0	1	1.0926	0.0182	1.0569	1.1283	3599.71	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		0	0.0000	0.0000	0.0000	0.0000		

Table 9 Negative Binomial Criteria For Assessing Goodness-Of-Fit

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	13E3	13700.3624	1.0716
Scaled Deviance	13E3	13700.3624	1.0716
Pearson Chi-Square	13E3	11331.5923	0.8863
Scaled Pearson X2	13E3	11331.5923	0.8863
Log Likelihood		8775.9792	
Full Log Likelihood		-22821.1920	

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
AIC (smaller is better)		45664.3841	
AICC (smaller is better)		45664.4047	
BIC (smaller is better)		45746.4090	

Table 10 Negative Binomial Model Variables

In Model	In Data
Y is	TARGET
X_1 is	AcidIndex
X_2 is	LabelAppeal
X_3 is	imp_STARS
X_4 is	i_imp_STARS

Negative Binomial model results in almost identical figures as Poisson approach for to the mean and variance being so close. Exception of AICC and BIC, both are same. Therefore, the interpretation for Poisson will apply the same for Negative Binomial model.

Even with varied inputs, neither models are compelling with the parameters, thus further modeling will be made.

Zero-Inflated Poisson

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e$$

Table 11 Zero-Inflated Poisson Model Analysis Of Maximum Likelihood Parameter

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	1.8750	0.0499	1.7773	1.9727	1413.66	<.0001
AcidIndex		1	-0.0223	0.0049	-0.0320	-0.0126	20.34	<.0001
LabelAppeal	-2	1	-0.9652	0.0439	-1.0512	-0.8793	484.26	<.0001
LabelAppeal	-1	1	-0.5995	0.0260	-0.6504	-0.5486	533.13	<.0001
LabelAppeal	0	1	-0.3390	0.0236	-0.3852	-0.2928	206.89	<.0001
LabelAppeal	1	1	-0.1567	0.0238	-0.2032	-0.1101	43.46	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
imp_stars	1	1	-0.4170	0.0230	-0.4622	-0.3718	327.39	<.0001
imp_stars	2	1	-0.2012	0.0199	-0.2403	-0.1621	101.76	<.0001
imp_stars	3	1	-0.1049	0.0202	-0.1445	-0.0653	26.98	<.0001
imp_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
i_imp_stars	0	1	0.1868	0.0196	0.1483	0.2253	90.62	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Table 12 Zero-Inflated Poisson Model Analysis Of Maximum Likelihood Zero Inflation Parameter

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.4731	0.1989	-3.8628	-3.0833	305.03	<.0001
AcidIndex		1	0.4773	0.0247	0.4288	0.5258	372.21	<.0001
i_imp_stars	0	1	-3.6189	0.0919	-3.7990	-3.4388	1550.75	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.

Table 13 Zero-Inflated Poisson Model Criteria For Assessing Goodness-Of-Fit

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		41927.6145	
Scaled Deviance		41927.6145	
Pearson Chi-Square	13E3	6122.3756	0.4790
Scaled Pearson X2	13E3	6122.3756	0.4790
Log Likelihood		10633.3640	
Full Log Likelihood		-20963.8072	
AIC (smaller is better)		41953.6145	

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
AICC (smaller is better)		41953.6429	
BIC (smaller is better)		42050.5530	

Table 14 Zero-Inflated Poisson Model Variables

In Model	In Data
Y is	TARGET
X_1 is	AcidIndex
X_2 is	LabelAppeal
X_3 is	imp_STARS
X_4 is	i_imp_STARS

The exponentiated AcidIndex coefficient is the multiplicative term used to calculate the estimated target when AcidIndex increases by one unit. For the categorical variables, the exponentiated coefficient is the multiplicative term relative to the base level for each variable. The exponentiated intercept is the baseline rate, and all other estimates will be relative to it.

The effect of a one unit increase in AcidIndex is a 2% decrease in the expected number of cases purchased.

Given that LabelAppeal has a base level of 2 as the highest rating, followings can be interpreted;

- Negative two rating: 61% decrease in the expected number of cases purchased
- Negative one rating: 45% decrease in the expected number of cases purchased
- Zero rating: 28% decrease in the expected number of cases purchased
- Positive one rating: 14% decrease in the expected number of cases purchased

Each of which is interpreted as decrease from obtaining a two rating.

Given that STARS has a base level of 4 as the highest rating, followings can be interpreted;

- One rating: 34% decrease in the expected number of cases purchased
- Two rating: 18% decrease in the expected number of cases purchased
- Three rating: 10% decrease in the expected number of cases purchased

Each of which is interpreted as a ‘decrease from from obtaining a 4 rating’.

For the Zero-Inflation Poisson model, this portion refers to the logistic model due to link function, predicting whether the number of cases purchased is zero.

The effect of one unit increase in AcidIndex results in 61% increases in the odds that this wine would belong to the certain zero group for number of cases purchased. The effect of not being rated, or being

omitted from having a star rating, is a 97% increase in the odds that this wine would belong to the certain zero group for number of cases purchased.

Zero-Inflated Negative Binomial

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e$$

Table 15 Zero-Inflated Negative Binomial Model Analysis Of Maximum Likelihood Parameter

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	1.8705	0.0499	1.7726	1.9684	1403.01	<.0001
AcidIndex		1	-0.0214	0.0049	-0.0310	-0.0117	18.71	<.0001
LabelAppeal	-2	1	-0.9704	0.0440	-1.0566	-0.8842	487.27	<.0001
LabelAppeal	-1	1	-0.6029	0.0260	-0.6539	-0.5519	536.36	<.0001
LabelAppeal	0	1	-0.3409	0.0236	-0.3872	-0.2945	207.84	<.0001
LabelAppeal	1	1	-0.1574	0.0238	-0.2041	-0.1106	43.55	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
imp_stars	1	1	-0.4068	0.0230	-0.4519	-0.3618	312.88	<.0001
imp_stars	2	1	-0.1999	0.0200	-0.2391	-0.1606	99.53	<.0001
imp_stars	3	1	-0.1046	0.0203	-0.1444	-0.0648	26.56	<.0001
imp_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
i_imp_stars	0	1	0.1854	0.0197	0.1469	0.2239	88.92	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		0	0.0019	0.0000	0.0019	0.0019		

Table 16 Zero-Inflated Negative Binomial Model Analysis Of Maximum Likelihood Zero Inflation Parameter

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-3.3657	0.1930	-3.7439	-2.9875	304.21	<.0001
AcidIndex		1	0.4637	0.0240	0.4168	0.5107	374.47	<.0001
i_imp_stars	0	1	-3.4689	0.0828	-3.6311	-3.3067	1757.03	<.0001
i_imp_stars	1	0	0.0000	0.0000	0.0000	0.0000	.	.

Table 17 Zero-Inflated Negative Binomial Model Criteria For Assessing Goodness-Of-Fit

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance		41984.4131	
Scaled Deviance		41984.4131	
Pearson Chi-Square	13E3	6016.3408	0.4707
Scaled Pearson X2	13E3	6016.3408	0.4707
Log Likelihood		-20992.2065	
Full Log Likelihood		-20992.2065	
AIC (smaller is better)		42012.4131	
AICC (smaller is better)		42012.4459	
BIC (smaller is better)		42116.8084	

Table 18 Zero-Inflated Negative Binomial Model Variables

In Model	In Data
Y is	target
X_1 is	acidindex
X_2 is	labelappeal
X_3 is	imp_stars
X_4 is	i_imp_stars

The result is similar to that of Zero-Inflated Poisson model, thus it shares the same interpretations.

Select Model

Comparing model coefficients makes logical sense, yet a presence of an analog to the ROC curve we used in logistic regression may have yielded for better selection. As values of mean and variance are similar, it was expected that only minor difference between Poisson and Negative Binomial would be present prior to building models and results support that theory. Since TARGET value is zero-inflated, **Zero-Inflated Negative Binomial Model** is chosen as the best model.

Comparison methods were built using a sub-sampling test procedure to evaluate the performance of the difference models.

Conclusion

Initially, creating Logistic and Poisson models caused complexities with interpretations. Once employed, however, these feels more natural than other techniques throughout the course. Choosing a basis for interpretation when incorporating categorical variables can significantly influence the feel of interpretations.

For the business owner for this assignment, this study would most likely desire using objectively measurable parameters in the model. Choosing the qualitative parameters may provide for better modeling performance. It creates further requirements for future data collection. For example, the business owner may wish to object to the inclusion of STARS because it requires human sampling of the product in a disposable fashion, while LabelApproval can be sampled in a non-destructive fashion.