DATA 621 - Homework #5

Claudio, Mauricio

2022-05-15

Overview

In this homework assignment, you will explore, analyze and model a data set containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales. Your objective is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine.

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET	Number of Cases Purchased	None
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 customes don't like the design.	Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales.
ResidualSugar	Residual Sugar of wine	
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor	A high number of stars suggests high sales
Sulphates	Sulfate conten of wine	
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	
VolatileAcidity	Volatile Acid content of wine	
pH	pH of wine	

Data Exploration and Preparation

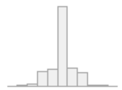
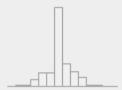
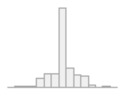
The training dataset consists of 12,795 instances, 14 predictor variables and one target varible. We note the following gaps in the data:

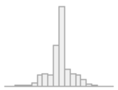
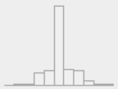
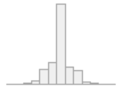
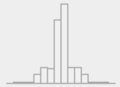
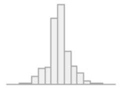
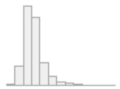
- Missing values in predictor variables ResidualSugar (4.8%), Chlorides (5.0%), FreeSulfurDioxide (5.1%), TotalSulfurDioxide (5.3%), pH (3.1%), Sulphates (9.5%), Alcohol (5.1%) and STARS (26.3%)

To prepare the data for model building, we effect the following transformations:

- Impute the attribute mean to the missing values in predictors ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates and Alcohol.
- Impute a zero value to the missing values in predictor STARS.
- Convert categorical variables to factors

We do not test the data for colinearity, linearity or outlier/leverage points because our aim is prediction rather than inference and those tests are more appropriate in the context of a particular regression model or set of models. Likewise, we wish to avoid the potential for overfitting and diminished predictive performance that variable elimination due to colinearity or observation deletion could effect. The transformed dataset, ready for model building, is summarized below.

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	TARGET [integer]	Mean (sd) : 3 (1.9) min ≤ med ≤ max: 0 ≤ 3 ≤ 8 IQR (CV) : 2 (0.6)	0 : 2734 (21.4%) 1 : 244 (1.9%) 2 : 1091 (8.5%) 3 : 2611 (20.4%) 4 : 3177 (24.8%) 5 : 2014 (15.7%) 6 : 765 (6.0%) 7 : 142 (1.1%) 8 : 17 (0.1%)		12795 (100.0%)	0 (0.0%)
2	FixedAcidity [numeric]	Mean (sd) : 7.1 (6.3) min ≤ med ≤ max: -18.1 ≤ 6.9 ≤ 34.4 IQR (CV) : 4.3 (0.9)	470 distinct values		12795 (100.0%)	0 (0.0%)
3	VolatileAcidity [numeric]	Mean (sd) : 0.3 (0.8) min ≤ med ≤ max: -2.8 ≤ 0.3 ≤ 3.7 IQR (CV) : 0.5 (2.4)	815 distinct values		12795 (100.0%)	0 (0.0%)
4	CitricAcid [numeric]	Mean (sd) : 0.3 (0.9) min ≤ med ≤ max: -3.2 ≤ 0.3 ≤ 3.9 IQR (CV) : 0.5 (2.8)	602 distinct values		12795 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
5	ResidualSugar [numeric]	Mean (sd) : 5.4 (32.9) min ≤ med ≤ max: -127.8 ≤ 4.9 ≤ 141.2 IQR (CV) : 14 (6.1)	2078 distinct values		12795 (100.0%)	0 (0.0%)
6	Chlorides [numeric]	Mean (sd) : 0.1 (0.3) min ≤ med ≤ max: -1.2 ≤ 0 ≤ 1.4 IQR (CV) : 0.1 (5.7)	1664 distinct values		12795 (100.0%)	0 (0.0%)
7	FreeSulfurDioxide [numeric]	Mean (sd) : 30.8 (144.9) min ≤ med ≤ max: -555 ≤ 30.8 ≤ 623 IQR (CV) : 59 (4.7)	1000 distinct values		12795 (100.0%)	0 (0.0%)
8	TotalSulfurDioxide [numeric]	Mean (sd) : 120.7 (225.6) min ≤ med ≤ max: -823 ≤ 120.7 ≤ 1057 IQR (CV) : 164 (1.9)	1371 distinct values		12795 (100.0%)	0 (0.0%)
9	Density [numeric]	Mean (sd) : 1 (0) min ≤ med ≤ max: 0.9 ≤ 1 ≤ 1.1 IQR (CV) : 0 (0)	5933 distinct values		12795 (100.0%)	0 (0.0%)
10	pH [numeric]	Mean (sd) : 3.2 (0.7) min ≤ med ≤ max: 0.5 ≤ 3.2 ≤ 6.1 IQR (CV) : 0.5 (0.2)	498 distinct values		12795 (100.0%)	0 (0.0%)
11	Sulphates [numeric]	Mean (sd) : 0.5 (0.9) min ≤ med ≤ max: -3.1 ≤ 0.5 ≤ 4.2 IQR (CV) : 0.4 (1.7)	631 distinct values		12795 (100.0%)	0 (0.0%)
12	Alcohol [numeric]	Mean (sd) : 10.5 (3.6) min ≤ med ≤ max: -4.7 ≤ 10.5 ≤ 26.5 IQR (CV) : 3.1 (0.3)	402 distinct values		12795 (100.0%)	0 (0.0%)
13	LabelAppeal [factor]	1. -2 2. -1 3. 0 4. 1 5. 2	504 (3.9%) 3136 (24.5%) 5617 (43.9%) 3048 (23.8%) 490 (3.8%)		12795 (100.0%)	0 (0.0%)
14	AcidIndex [integer]	Mean (sd) : 7.8 (1.3) min ≤ med ≤ max: 4 ≤ 8 ≤ 17 IQR (CV) : 1 (0.2)	14 distinct values		12795 (100.0%)	0 (0.0%)
15	STARS [numeric]	Mean (sd) : 1.5 (1.2) min ≤ med ≤ max: 0 ≤ 1 ≤ 4 IQR (CV) : 2 (0.8)	0 : 3359 (26.3%) 1 : 3042 (23.8%) 2 : 3570 (27.9%) 3 : 2212 (17.3%) 4 : 612 (4.8%)		12795 (100.0%)	0 (0.0%)

Generated by summarytools 1.0.0 (R version 4.1.0)
2022-05-15

// Model Building and Selection

| Count Models

We first build two models appropriate for count data – Poisson and Negative Binomial – through backward selection at 95% confidence (p-value < 0.05) and test them for over-dispersion and zero inflation. On over-dispersion, the tests show that the data is not over-dispersed. If anything the data is slightly under-dispersed with a dispersion paramater of about 0.85. On zero-inflation, the tests show that the data is significantly zero inflated and crucially, that the models underfit the zero values. The AIC for Poisson and Negative Binomial models are 46667 and 46670, respectively. The two models are nearly identical in performance and fit. They are summarized below:

	Poisson		Negative Binomial	
Predictors	*Incidence Rate Ratios*	*p*	*Incidence Rate Ratios*	*p*
(Intercept)	2.374	**6e-52**	2.374	**6e-52**
VolatileAcidity	0.967	**2e-07**	0.967	**2e-07**
Chlorides	0.959	**1e-02**	0.959	**1e-02**
FreeSulfurDioxide	1.000	**4e-04**	1.000	**4e-04**
TotalSulfurDioxide	1.000	**3e-04**	1.000	**3e-04**
pH	0.984	**4e-02**	0.984	**4e-02**
Sulphates	0.988	**3e-02**	0.988	**3e-02**
LabelAppeal-1	1.296	**8e-12**	1.296	**8e-12**
LabelAppeal [0]	1.543	**9e-32**	1.543	**9e-32**
LabelAppeal [1]	1.696	**1e-44**	1.696	**1e-44**
LabelAppeal [2]	1.880	**4e-50**	1.880	**4e-50**
AcidIndex	0.916	**1e-84**	0.916	**1e-84**
STARS	1.366	**0e+00**	1.366	**0e+00**
Observations	12795		12795	
R^2 Nagelkerke	0.566		0.566	

We test the predictive performance of the two models by calculating their Root Mean Square Error (RMSE) via Monte Carlo cross-validation with a 50% train / 50% test data split and $\sqrt{n}$ or 113 iterations. We use the RMSE because it is the most conservative estimate, sensitive to the large errors that which wish to minimize in our predictions. For replicability we set the seed to `set.seed(i)` where i is the iteration. This method assures both that each iteration results in a different random sample test/train split and replicability. The RMSE for each model is show below:

RMSE: Poisson model	RMSE: Negative Binomial model	cross-validations
2.59657	2.59657	113

We note that the RMSE for both models is identical, approximately 2.60.

| Zero-inflated Count Models

Due to data zero-inflation, model underfitting of zeros and mediocre predictive performance, we build four zero-inflated models – two mixture models and two Hurdle models – through backward selection at 95% confidence (p-value < 0.05). The mixture models, fitted with the `zeroinfl()` function, model the zeros as a combination of a binary Bernoulli process and a count Poisson process. The zeros may come from one of two processes, the Bernoulli process or from the Poisson process. By constrast, the Hurdle models, fitted with the `hurdle()` function, model the zeros as arising strictly from a single binary Bernoulli process. Once the *hurdle* of a non-zero value is crossed in the Bernoulli process, a Poisson process takes over and models the non-zero, positive counts. In a Hurdle model, the Poisson count process is zero-truncated, that is, it accounts for no zeros and only positive counts, unlike a zero-inflated model. The four zero-inflated models fitted and their AIC are as follows:

- Zero-inflated Poisson (ZIP): AIC 40785
- Zero-inflated Negative Binomial (ZINB): AIC 40787
- Hurdle Poisson (HP): AIC 40679
- Hurdle Negative Binomial (HNB): 40681

They are summarized below:

	Zero-inflated Poisson		Zero-inflated Negative Binomial		Hurdle Poisson		Hurdle Negative Binomial	
Predictors	*Incidence Rate Ratios*	*p*	*Incidence Rate Ratios*	*p*	*Incidence Rate Ratios*	*p*	*Incidence Rate Ratios*	*p*
count_(Intercept)	1.6145	**2e-17**	1.6148	**2e-17**	1.4249	**2e-08**	1.4247	**2e-08**
count_Alcohol	1.0071	**8e-07**	1.0071	**8e-07**	1.0076	**3e-07**	1.0076	**3e-07**
count_LabelAppeal-1	1.5562	**6e-27**	1.5558	**6e-27**	1.7149	**2e-27**	1.7152	**2e-27**
count_LabelAppeal0	2.0812	**2e-74**	2.0808	**3e-74**	2.3248	**4e-67**	2.3252	**3e-67**
count_LabelAppeal1	2.5207	**3e-113**	2.5203	**3e-113**	2.8364	**3e-99**	2.8370	**3e-99**
count_LabelAppeal2	2.9530	**9e-126**	2.9523	**1e-125**	3.3374	**8e-114**	3.3381	**7e-114**
count_AcidIndex	0.9813	**9e-05**	0.9813	**9e-05**	0.9831	**5e-04**	0.9831	**5e-04**
count_STARS	1.1059	**9e-84**	1.1059	**9e-84**	1.0984	**6e-71**	1.0984	**6e-71**
Zero-Inflated Model								
zero_(Intercept)	0.0018	**1e-42**	0.0018	**1e-42**	73.8814	**7e-54**	73.8814	**7e-54**
zero_VolatileAcidity	1.2234	**4e-06**	1.2233	**5e-06**	0.8296	**4e-07**	0.8296	**4e-07**
zero_FreeSulfurDioxide	0.9992	**1e-03**	0.9992	**1e-03**	1.0006	**2e-03**	1.0006	**2e-03**
zero_TotalSulfurDioxide	0.9990	**1e-10**	0.9990	**1e-10**	1.0009	**1e-11**	1.0009	**1e-11**

zero_pH	1.2384	**3e-05**	1.2386	**3e-05**	0.8305	**1e-05**	0.8305	**1e-05**
zero_Sulphates	1.1431	**6e-04**	1.1431	**6e-04**	0.8991	**1e-03**	0.8991	**1e-03**
zero_Alcohol	1.0285	**3e-03**	1.0286	**3e-03**	0.9791	**8e-03**	0.9791	**8e-03**
zero_LabelAppeal-1	4.8612	**5e-06**	4.8648	**5e-06**	0.6096	**4e-04**	0.6096	**4e-04**
zero_LabelAppeal0	10.6671	**5e-12**	10.6739	**5e-12**	0.4008	**2e-11**	0.4008	**2e-11**
zero_LabelAppeal1	21.2483	**2e-18**	21.2560	**2e-18**	0.2373	**3e-23**	0.2373	**3e-23**
zero_LabelAppeal2	31.4226	**2e-18**	31.4533	**2e-18**	0.1715	**8e-16**	0.1715	**8e-16**
zero_AcidIndex	1.5474	**2e-66**	1.5475	**1e-66**	0.6765	**3e-75**	0.6765	**3e-75**
zero_STARS	0.0928	**0e+00**	0.0928	**0e+00**	7.8015	**0e+00**	7.8015	**0e+00**
Observations	12795		12795		12795		12795	
R^2 / R^2 adjusted	0.826 / 0.826		0.826 / 0.826		0.824 / 0.824		0.824 / 0.824	

We test the predictive performance of the three zero-inflated models with the lowest AIC scores: Zero-inflated Poisson, Hurdle Poisson and Hurdle Negative Binomial. We test them via Monte Carlo cross-validation with 50% train / 50% test data splits and $\sqrt{n}$ or 113 iterations, and calculate the RMSE for each model. The model performance summary is show below:

RMSE: Zero-inflated Poisson model	RMSE: Hurdle Poisson model	RMSE: Hurdle Negative Binomial model	cross-validations
1.267746	1.264539	1.264539	113

We note the RMSE of the zero-inflated models is significantly lower than that of the initial two models. The RMSE of the zero-inflated models, about 1.27, is less than half the value of the non zero-inflated models, 2.60. We note also that among the zero-inflated models, the Hurdle models edge out, however marginally, the zero-inflated Poisson model. Between the two Hurdle models, it's a toss-up. **Hence, on the toss-up we select the Hurdle Poisson model as our model for prediction purposes.**

Prediction

We transform the evaluation dataset in the same manner that we transformed the training dataset earlier and predict the response based on our selected model, **Hurdle Poisson**. The predicted probabilities and predictions for the 3,335 observations in the evaluation data set are summarized below. A .csv file of the predictions is available for download.

Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid
prediction [integer]	Mean (sd) : 3.1 (1.5) min ≤ med ≤ max: 0 ≤ 3 ≤ 7 IQR (CV) : 2 (0.5)	0 : 73 (2.2%) 1 : 584 (17.5%) 2 : 474 (14.2%) 3 : 847 (25.4%) 4 : 810 (24.3%) 5 : 401 (12.0%) 6 : 122 (3.7%) 7 : 24 (0.7%)		3335 (100.0%)

Generated by summarytools 1.0.0 (R version 4.1.0)
2022-05-15