Kay Quiballo | MSDS 410 | 02.20.2023

Modeling Assignment 8: Modeling Dichotomous Responses

Assignment Overview

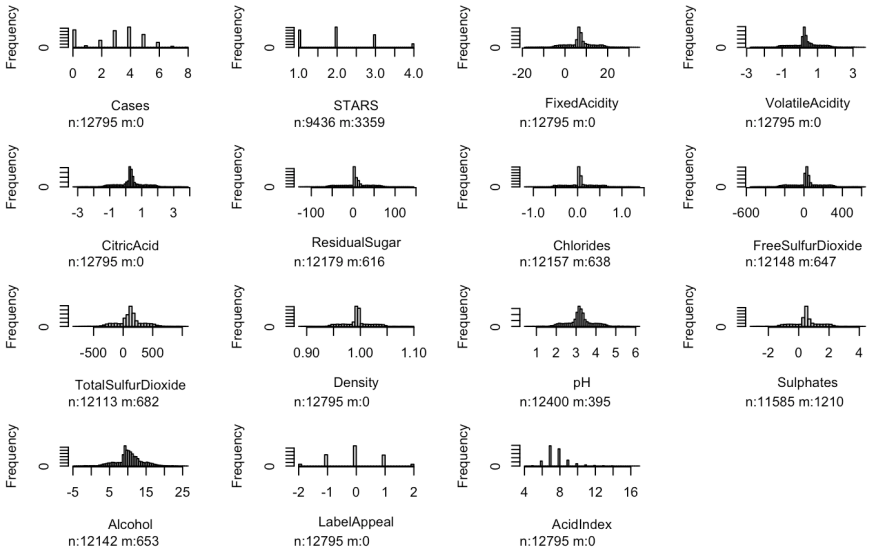
A large wine manufacturer is interested in being able to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales. But, it is also important to understand what influences purchase decisions in the first place, as well as what contributes to the quality of the wine. Your task in this assignment is to model the purchase decision using Logistic Regression models.

This data set contains information on approximately 12,000 commercially available wines. A record can be considered the data associated with a bottle of wine. The explanatory variables are mostly related to the chemical properties of the wine. But, there are other variables as well. For example, the PURCHASE reflects whether or not a purchase was made of that wine. PURCHASE is the response variable for this assignment. The variable CASES then indicates the number of cases purchased. 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 a wine is to be sold at a high end restaurant. Similarly, each wine, when possible, was rated by a panel of experts as to its quality (STARS).

From a statistical perspective, please note the size of the sample: n is approximately 12,000 records. You should immediately be thinking, “I have tons of statistical power.” I have to be careful about statistical significance, as it is not the be all and end all. Again, you can think about randomly splitting the file into a 70% model development dataset, and into 30% validation data set, if you wish.

1. Use your data analysis knowledge to date, to conduct an Exploratory Data Analysis (EDA) for fitting Logistic Regression models to predict the PURCHASE decision.

Histogram of Variables



Here we note that all predictors are numeric with an approximately normal distribution. Although we will not be using Cases in this analysis (as it seems to act more as a response), we note that it also has an approximately normal distribution. All chemical property variables of the wine are continuous (or integers that are treated as continuous like FreeSulfurDioxide and AcidIndex). However, non-chemical variables like STARS and LabelAppeal are discrete ordinal variables. In this analysis, I choose to treat these variables as numeric which preserves the ordered characteristic of the data but adds the assumption that each rating is equidistant from the next. We note summary statistics below.

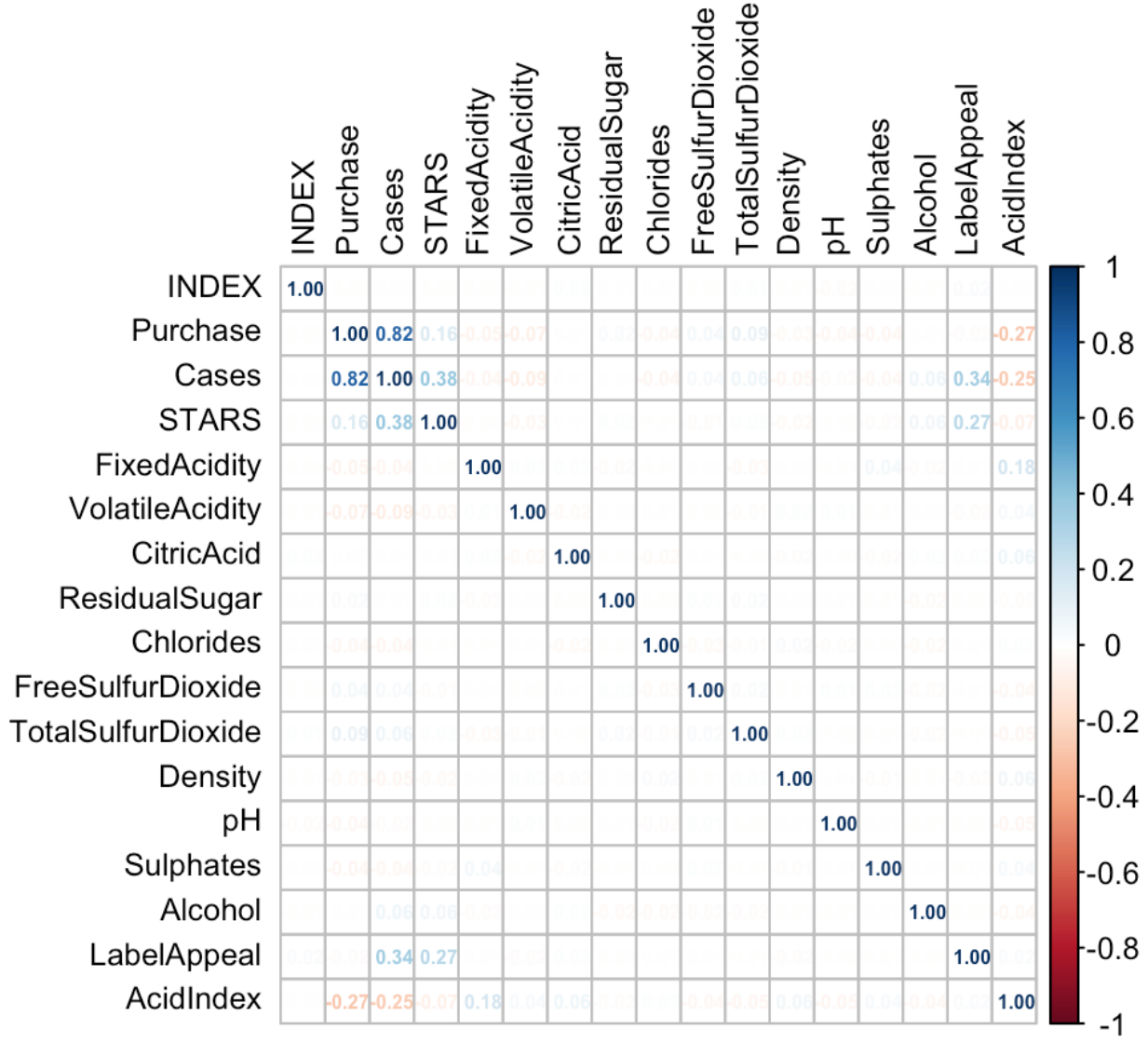
Summary Statistics

```
   Purchase          Cases           STARS        FixedAcidity      VolatileAcidity     CitricAcid      ResidualSugar         Chlorides
Min.   :0.0000   Min.   :0.000   Min.   :1.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710
1st Qu.:1.0000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310
Median :1.0000   Median :3.000   Median :2.000   Median :  6.900   Median : 0.2800   Median : 0.3100   Median :   3.900   Median : 0.0460
Mean   :0.7863   Mean   :3.029   Mean   :2.042   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548
3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:2.042   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530
Max.   :1.0000   Max.   :8.000   Max.   :4.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510
SD     :0.2614   SD     :1.546   SD     :0.771   SD     :  6.281   SD     : 0.7801   SD     : 0.8606   SD     :  33.656   SD     : 0.3166
                                                                                                       NA's   :616        NA's   :638

FreeSulfurDioxide TotalSulfurDioxide    Density             pH          Sulphates          Alcohol       LabelAppeal          AcidIndex
Min.   :-555.00   Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300   Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000
1st Qu.:   0.00   1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000
Median :  30.00   Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000   Median :10.40   Median : 0.000000   Median : 8.000
Mean   :  30.85   Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271   Mean   :10.49   Mean   :-0.009066   Mean   : 7.773
3rd Qu.:  70.00   3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000
Max.   : 623.00   Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400   Max.   :26.50   Max.   : 2.000000   Max.   :17.000
SD     : 148.30   SD     : 228.9     SD     :0.0265   SD     :0.678   SD     : 0.9247   SD     : 3.71   SD     : 0.871290   SD     : 1.191
NA's   :647       NA's   :682                         NA's   :395     NA's   :1210      NA's   :653
```

From the summary statistics, we can note that a few of the variables are missing data valuables. STARS has the most missing data. I chose to reconcile this by replacing NA values with the mean rating value of non-NA STAR ratings. For the other predictors however, those records will be eliminated from the dataset if the model uses that predictor. We can also match up the statistic of central tendency to the histograms above for a more accurate reading. If outliers exist, we will examine them at the end of the analysis when we are in the model selection phase.

Correlation Matrix



In the correlation matrix, we will take note of any potential predictors with a relatively high association with our response, Purchase. The results include: STARS (0.16) and AcidIndex (-0.27). Cases has a high correlation with Purchase because it indicates the number of Cases of wine purchased when Purchase=1 which is why we will treat it as a response. We may also note that although LabelAppeal does not have a high association with Purchase, but it does have a high association with Cases and Stars indicating that better looking bottles of wine tend to sell more and have higher ratings.

2. Produce your best predictive model for the PURCHASE (Y) decision. You have enough data, so you should very seriously consider taking a validation approach to this modeling endeavor. You need to be sure you can interpret your models, have evidence on goodness of fit, and check on assumptions via diagnostics. What criteria are you going to use select your "best" model?

Model	Method	Variables	Max VIF	AIC/BIC	% change deviance	AUC	Nested Chisq test	Precision Sensitivity Specificity
stepwise.lm	Stepwise AIC on all variables	STARS + AcidIndex + TotalSulfurDioxide + LabelAppeal + VolatileAcidity + pH + FreeSulfurDioxide + Sulphates + Chlorides + CitricAcid	1.096316	5704.966 5778.816	9.98083%	0.7099	0.8613	FALSE TRUE 0 80 453 1 39 2018 0.8100386 0.6722689 0.8166734
cor.glm	Variables with high correlation to Purcase	STARS,AcidIndex	1.000073	5791.438 5811.579	8.35880%	0.6654	2.2e-16	FALSE TRUE 0 65 468 1 26 2031 0.8092664 0.7142857 0.8127251
dict.glm	Variables with predicted association to Purchase according to data dictionary	STARS, LabelAppeal	1.073923	6134.449 6154.59	2.92566%	0.4246	2.2e-16	FALSE TRUE 0 0 533 1 0 2057 0.7942085 NA 0.7942085
sci.glm	only scientific variables	AcidIndex, Alcohol, Chlorides, CitricAcid, Density, FixedAcidity, FreeSulfurDioxide, ResidualSugar, Sulphates, TotalSulfurDioxide, VolatileAcidity, pH	1.051706	5858.304 5945.581	7.61591%	0.6709	2.2e-16	FALSE TRUE 0 73 460 1 28 2029 0.811583 0.7142857 0.8127251
full.glm	All variables	[all variables]	1.070610	5711.666 5812.37	10.0014%	0.7103	BASE	FALSE TRUE 0 81 452 1 40 2017 0.8100386 0.6694215 0.8169299

Red indicates the worst performing column value for each model.
Yellow indicates middle-of-the-road performance.
Green indicates relatively high level performance.

Write of description of the technique you used to decide on your final model. Write up your final model. Report the model. Discuss the coefficients in the model, do they make sense? Report on goodness of fit and model diagnostics.

stepwise.lm: Although the model that was chosen through ML stepwise AIC performs the best in terms of training statistics, the testing evaluation did not yield the best results (specifically in sensitivity).

cor.glm: This model performed decently in training statistics – no multicollinearity due to VIF, one of the lowest AIC/BIC values, moderate decrease in deviance, moderate area under the curve, and best performing evaluation of test dataset using precision/sensitivity/specificity. This will be our final model due to its out-of-sample validation and the lowest number of predictors. Although the Chi Squared test indicates that a variable not in the model is statistically significant, we choose not to select additional predictors based on the validation of the other models.

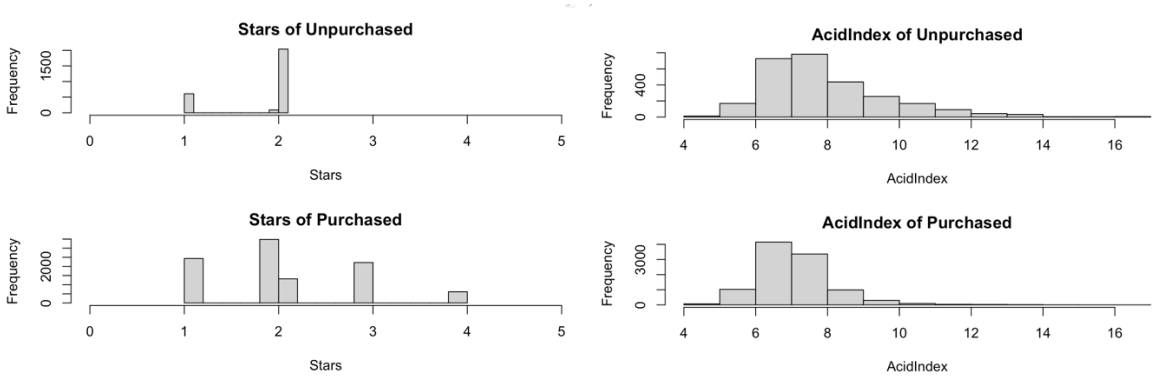
dict.glm: This model has the poorest performance in training data statistics, and predicts all bottles of wine will be purchased which yields a trash model.

sci.glm: This model performed moderately like cor.glm in training statistics and performed well in out-of-sample validation. We choose not to use this as our final model due to the number of variables (high), some of which test insignificantly according to T-tests.

full.glm: The full model performed similarly to the cor.glm model – great training statistics but moderate out-of-sample validation. We choose not to use this as our final model.

A more in depth analysis of the final model's statistics is in the conclusion.

Analysis of Final Model Predictors



On the left, we see that the distribution of bottles purchased tends to favor higher STAR ratings. Bottles that have 3 or 4 stars are much more likely to be purchased over 1 or 2 stars.

On the right, we see that the distribution of bottles purchased tends to *not* favor AcidIndex > 9. Bottles that are too acidic are not as likely to be purchased.

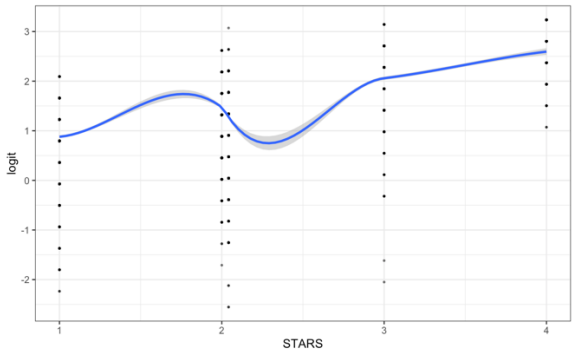
Model Assumptions and Diagnostics

- The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0.

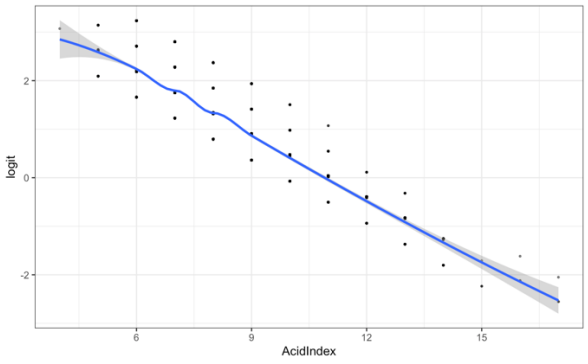
Using the predict function for this model, it yields probabilities ranging from .07 to .96 which satisfies this assumption.

- There is a linear relationship between the logit of the outcome and each predictor variables.

Logit vs Stars

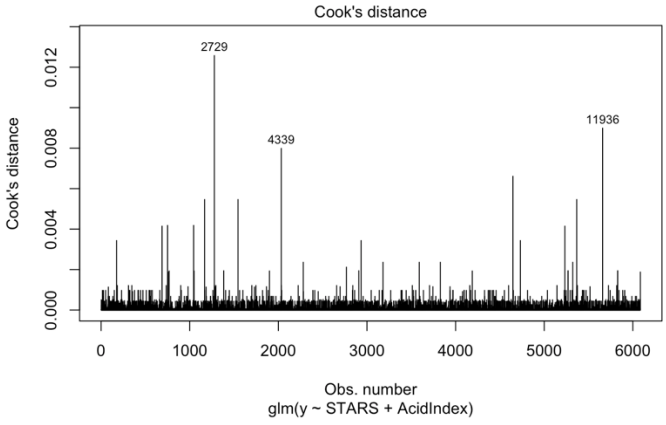


Logit vs AcidIndex



These predictors do have an approximately linear relationship with logit Purchase. Stars tends to increase whereas AcidIndex tends to decrease in Logit value as the predictor grows

- There is no influential values (extreme values or outliers) in the continuous predictors



.std.resid
<dbl>
2.088711
2.164556
1.901390

The 3 most influential points are labeled in the Cook's distance plot, but none of them have standardized residuals > 3. We choose not to remove any datapoints.

- There is no high intercorrelations (i.e. multicollinearity) among the predictors.

The highest VIF value for this model is 1.000073 which is < 5. We conclude there is no multicollinearity.

Now that we have confirmed that all of the logistic regression model assumptions, we can move on to coefficient interpretation.

Model Summary

```
Coefficients:
            Estimate Std. Error z value            Pr(>|z|)
(Intercept)  3.73054    0.21175   17.62 <0.0000000000000002 ***
STARS        0.52538    0.04722   11.13 <0.0000000000000002 ***
AcidIndex   -0.43273    0.02352  -18.39 <0.0000000000000002 ***
```

Coefficient Interpretation

For each additional STAR rating, we estimate the odds of a wine bottle being purchased *increases* by exp(0.52538) – 1 = 69.11013%.
For each additional Acid Index, we estimate the odds of a wine bottle being purchased *decreases* by exp(-0.43273) – 1 = 35.12644%.

Goodness of Fit

Precision (testing dataset): 0.8092664
This model produces positive predictions (Purchased bottles of wine) that are correct 80.92% of the time

Sensitivity (testing dataset): 0.7142857
This model has the ability to identify true positives (Purchased bottles of wine) 71.42% of the time.

Specificity (testing dataset):: 0.8127251
This model has the ability to identify true negatives (Unpurchased bottles of wine) 81.12% of the time.

AUC (training dataset): 0.6654
This indicates that the model does an acceptable job at discriminating between response classes.

Although the model does not have the best AUC out of all the models tested, it does have the highest out-of-sample validation statistics (precision, sensitivity, and specificity) and the lowest number of predictors to avoid overfitting.

3. Conclusion

What conclusions do you draw from having conducted this analysis?

From this assignment, the biggest takeaway I had was that the most complex models are not the best models. In order to predict whether or not a bottle of wine was going to be purchased, I tested 5 different logistic regression models. Some models were highly influenced by statistics such as correlation and AIC values. Other models had nothing to do with statistics but rather the science of wine and assumptions on influential factors based on the data dictionary. After testing all these models, in the end I chose the simplest model with only two predictors based on their association with the response variable. Although this model performed in the middle of the pack in the training phase (AIC/BIC/AUC/deviation), this model had some of the highest out-of-sample validation statistics (precision/specificity/sensitivity) that scored on par with 10+ variable models. With only a fraction of the predictors, this model still performed relatively well and yielded straight forward interpretations of coefficients while mitigating chances for multicollinearity and overfitting.

What did you learn about the wine world through your modeling endeavor?

Wine is very complex and is more than just the look and taste. The dataset we analyzed in this assignment contained a dozen chemical properties of the wine ranging from free sulphur dioxide to fixed acidity. I typically focus on the color of wine, possibly the year it was made, and even the alcoholic content, but I did not know that there were this many different chemical aspects of wine. I also assumed that the LabelAppeal would play a larger role in predicting Purchases, but in the end, the final model mostly relied on experts' ratings and the chemical acidity of the wine.

What actions can you recommend to anyone involved in this field?

I would recommend that anyone who is involved in the intersection of data science and wine sales have some background knowledge in each of these areas. As a data science expert that does not know much about chemistry, wine, or sales, I had to look to outside resources on how these fields may interact. For example, it may be helpful to have a background in chemistry to see how chemical properties influence the taste of a wine. It may also be helpful to have a background in wine sales to see how the different seasons or time of sale influences the purchases of certain bottles. Whether it's a team working together, or a data scientist using outside resources, it's important to grasp the different concepts from the intersections of various areas of expertise.

How did your perspective on modeling change?

When I began this assignment, I automatically assumed that the stepwise AIC model would produce the objectively best fitting model. I was surprised to find that it performed middle-of-the-pack in terms of out-of-sample validation (specifically sensitivity). Despite running chi squared tests on nested models that concluded the need for additional variables, I chose to stick with a simpler model that had great validation stats and a lower number of variables. I learned in context that the stepwise AIC models may be objectively accurate, but they are not necessarily the best choice in terms of interpretation.

4. Supplemental Materials

Data Dictionary

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
Purchase	Indicator Purchase was made (1=yes, 0=no)	None
Cases	Number of Cases Purchased	None
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor	A high number of stars suggests high sales
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	
Sulphates	Sulfate conten of wine	
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	
VolatileAcidity	Volatile Acid content of wine	
pH	pH of wine	