

**Modeling Assignment 4: The Wine Study**

**MSDS 410**

***INTRODUCTION***

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 variables are mostly related to the chemical properties of the wine being sold. But data comes from other sources as well. For example, the PURCHASE reflects whether or not a purchase was made of that wine any wine distribution companies after sampling a wine. 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 group of experts as to its quality (STARS).

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. But, it is also important to understand what influences purchase decisions as well as what contributes to the quality of the wine.

For this project, you may choose one of three response variables for which you will build a predictive model. The possible response variables are:

Purchase Decision (PURCHASE)

The rating of the wine (STARS)

The number of cases of wine sold (CASES)

Please note, the variable STARS can also be an explanatory variable for the PURCHASE and CASES variables. You can only use the variables given to you, or variable that you derive from the variables provided. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. You are welcome to use OLS regression, Logistic regression, Poisson Regression, or Zero-Inflated Poisson Regression methods in fitting these models. Be sure you know which one fits your situation.

All topics that have been learned over the term should be employed in this data modeling assignment. This includes: EDA, use of categorical and continuous explanatory variables, model fit statistics and diagnostics, automated selection methods or some other approach to variable selection, validation, sample splitting, and model refinement and cleaning. 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.

As you proceed to the write up for this assignment, please keep in mind that you are writing this assignment for a manager or boss. What you should do for this assignment:

***TASK 1: EXPLORATORY DATA ANALYSIS and DATA PREP***

Use your data analysis knowledge to date, to conduct an Exploratory Data Analysis (EDA). Some suggestions for things that you could do are:

* 1. Histograms for each continuous variable
  2. Means, standard deviations, minimum, maximum, median for all continuous variables
  3. Are variables correlated to the target variable (TARGET\_WINS) or to other possible explanatory variables?
  4. Are any of the variables with missing values that need to be imputed or “fixed”? Fix missing values (maybe with a Mean or Median value or use a decision tree). Are there variables with so many missing values that the entire variable should be eliminated from the analysis?
  5. Do any of the variables have outliers or extreme values? Should these extreme values be replaced? Fix any extreme values that need fixing.
  6. Do any of the variables need a mathematical transformation, such as log or square root? Create new variables with these transformations and add them to the end of the dataset.
  7. Create any new variables that you are interested in.

Please do NOT treat this as a check list of things you must do to complete the assignment. The EDA is your responsibility. You should have your own thoughts about this step based on your prior experiences this class so far.

Write a description of what you did in performing your EDA and data cleaning. Describe what you did and what you found so that a manager can understand it. Consider that too much detail will cause a manager to lose interest. DO NOT DATA DUMP! If you include a graph, you must describe and discuss that graph! Similarly, too little detail will make the manager consider that you aren’t doing your job or that you were not careful. Your reputation is at stake! This description is to be included in your final project write-up.

***TASK 2: MODELING***

There is not one perfectly correct way to approach model building. You are now charged with the task of producing your best predictive model for the WINE data and the response variable that you chose. Be sure that you are selecting the appropriate method for the variable you chose. This is an open-ended task where you are free to do whatever you wish in modeling this data. You may select the variables manually, use an approach such as Forward or Stepwise, do the variable selection by hand, or try to use an automated procedure. 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?

Write of description of the technique you used to decide on your final model. DO NOT DATA DUMP. A manager does not need to see everything that you produce, if though you have to produce the graphs and models to come to your final decision. If you include something in this report like a graph, table, or anything else, you must write about it! If you manually selected a variable for inclusion into the model or exclusion into the model, indicate how this was done.

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.

As you put together your “Final” prediction model for this project, remember that you can do the above and implement validation modeling like you did for Modeling Assignment #3. You can fit your final model on the Development Data and use the model to make predictions in the Validation dataset. You could fit your model on the Development Data and then refit your model on the Total dataset (development plus validation data) or just the validation dataset. Look to see how much the coefficients change, as a means of checking the stability of the model, as well assess its potential exporting success.

This isn’t totally rigorous, but it’s a nice check. Not required. Doesn’t even have to be reported. Just reminding you, that you know how to do this. Whatever validation work you do and want to report, put the evidence in a table so that comparisons can be easily made and discussed.

***TASK 3: CONCLUSIONS AND REFLECTIONS***

What conclusions do you draw from having conducted this analysis? What did you learn about the wine world through your modeling endeavor? What actions can you recommend to anyone involved in this field? How did your perspective on modeling change? Discuss anything else you wish to discuss.

***EXTRA CREDIT***

If you would like to gain extra credit or simply experience the joy of model fitting once again, please feel free to create a predictive model for either or both of the remaining response variables. Most likely, you would not need to do the EDA again, except for the logistic regression model looking for good explanatory variables, so the work would be centered around model fitting. I hope you take advantage of this opportunity. Each variable requires a different type of modeling.

CONGRATULATIONS! You are done with Modeling Assignment 4. You are home free! Well done!