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.

Assignment Tasks

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. Some suggestions for things that you could do are:
    - Histograms for each continuous variable
    - Means, standard deviations, minimum, maximum, median for all continuous variables
    - Are variables correlated to the target variable (TARGET_WINS) or to other possible explanatory variables?
    - 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?
    - Do any of the variables have outliers or extreme values? Should these extreme values be replaced? Fix any extreme values that need fixing.
    - 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.

- 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 in this class so far. Remember, the old adage: Garbage In = Garbage Out!

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 or corporate executive can understand it. Consider that too much detail will cause a manager to lose interest, not enough will cause them to question your credibility. DO NOT DATA DUMP! If you include a graph or summary statistics, you must describe and discuss that graph or table of statistics! You DO NOT need to include everything you do as part of your EDA.

2. 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 PURCHASE (Y) decision. This is an open-ended modeling task. You may select the variables manually, or use an automated approach such as Forward or Stepwise. You may use continuous or categorical variables as part of the explanatory variable set. You have enough data, so you should very seriously consider taking a validation approach to this modeling endeavor, though it is not required. 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. 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.

3. 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.

Assignment Document

Results should be presented, labeled, and discussed in the numerical order of the questions given. Please use MS-WORD or some other text processing software to record and present your answers and results. The report should not contain unnecessary results or information. Tables are highly effective for summarizing data across multiple models. The document you submit to be graded MUST be submitted in pdf format. Please use the naming convention: ModelAssign8_YourLastName.pdf.