WINE SALES PROJECT (250 Points)

This data set contains information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The target 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 a wine is 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.

Build a model to predict the number of cases of wine that will be sold given certain properties of the wine.

HINT: Sometimes, the fact that a variable is missing is actually predictive of the target.

You can only use the variables given to you (or variable that you derive from the variables provided).

DELIVERABLES

- Your write up in PDF Format. Your write up should have five sections. Each one is described below. Each section should have enough detail so that I can follow your logic and someone else can replicate your work. **(200 Points)**

- A file that contains all the Python code you used in your analysis. I should be able to run this file and get all the output that you did.

- A csv file which has the scored records values from test.csv. There will be only two columns in this file: INDEX, P_TARGET. You will be graded on how your model performs versus my model and those of other students in the class.

WRITE UP (200 POINTS):

1. DATA EXPLORATION (50 points)

Describe the size and the variables in the WINE data set so that I am convinced you understand it.

Use my shell code as a start to explore the data. Apply your creativity and go from there.

If you know how to do pivot tables in Excel, it is a great tool for Exploratory Data Analysis (EDA).

Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas:

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?
e. Don't delete records, fix them

2. DATA PREPARATION (50 Points)

This is a critical area for building great models. It is the reason we give you messy data sets to work with. We want you to have a real world experience so you will be able to build better models in the real world.

Describe how you have transformed the data by changing the original variables or creating new variables. If you don't show how you transformed the data I will not be able to see why your model performed good or bad. If you did transform the data or create new variables, tell me why.

Possible transformations:

a. Fix missing values (maybe with a Mean or Median value or use a decision tree)
b. Create flags to suggest if a variable was missing.
c. Transform data by putting it into buckets
d. Mathematical transforms such as log or square root
e. Combine variables (such as ratios or adding or multiplying) to create new variables

3. BUILD MODELS (60 Points)
Build at least three different models. The three models will be:

- Negative Binomial distribution
- Poisson distribution
- Regression
- Hurdle (optional)

You have a lot of choices for this assignment. See the shell code and the week 8 reading regarding a hurdle method plus the R code provided with the assignment. Python does not have code for the hurdle model. Be careful, some of these modeling methods over fit.

You may select the variables manually or use some other method. Describe the techniques you used. If you selected a variable for inclusion or exclusion indicate why.

Show all of your models and the statistical significance of the input variables.

Discuss the coefficients in the model, do they make sense? In this case, about the only thing you can comment on is the number of stars and the wine label appeal. However, you might comment on the coefficient and magnitude of variables and how they are similar or different from model to model. For example, you might say "pH seems to have a major positive impact in my regression model, but a negative effect elsewhere".

4. SELECT MODELS (20 Points)
Decide on the criteria for selecting the "Best Model". Will you use a metric such as AIC or Average Squared Error? Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model.

NOTE: This is a zero-inflated modeling exercise. If you happen to like the standard regression model the best, then that is OK. Please say that you like it the best and why you like it. HOWEVER, you MUST select a model for grading.

5. WRITE A FORMULA FOR YOUR MODEL (20 Points)
Write an equation for your model that will allow someone else to implement it. They should be able to score new data and predict the number of wine cases that will be sold based upon the qualities of the wine. The variable should be named:

P_TARGET

The model equation will need to include:

a. All the variable transformations such as fixing missing values

b. The model formulas

SCORED DATA FILE (50 POINTS)

Score the data file wine_test.csv. Create a file that has only TWO variables for each record:

INDEX
P_TARGET

The first variable, INDEX, will allow me to match my grading key to your predicted value. If I cannot do this, you won't get a grade. So please include this value. The second value, P_TARGET is the number of cases of wine that will be sold. It can be either an integer (i.e. 0, 1,2,3,etc.) or it can be continuous (i.e. 3.14159, 2.7, 4.567, etc).

Your values will be compared against …

- A Perfect Model
- Shell Code
- Performance of Other Students
- Predict the Average value for everybody (MEAN)
- Random Model
- Worst Possible Model

If your model is not better than simply using an AVERAGE value, you will receive 0 points

BONUS

If you want Bonus Points, write a brief section at the top of your Write Up document and tell me exactly what you did and how many points you are attempting. If I cannot see your Bonus work, I cannot give you credit. Bonus is difficult to grade and I don't have time to go back looking for it. If you don't tell me it's there, I cannot give you points.

The policy with Bonus is: ***All Sales are Final !***

- (?? Points) Roll the dice … think of something creative and run with it. I might give you points.

PENALTY BOX

- (Lose 10 Points) If you don't have PDF format
- (Lose 10 Points) If you don't have a GOOD Introduction
- (Lose 10 Points) If you don't have a GOOD Conclusion
- (Lose 10 Points) If you don't put your NAME in the file names of any files you hand in
- (Lose 10 Points) If you don't put your NAME inside of the files you hand in