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 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.

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).

Note – some of the variables have negative numbers when they technically shouldn't (based on laws of chemistry, etc.). Ignore these issues and use the data as you normally would. Basically, we modified the original data for proprietary reasons and made some values negative that should not have been.

DELIVERABLES:

- Your write up in PDF Format. Your write up should have four sections. Each one is described below. **(160 Points)**

- A file that contains all the R (or SAS) code you used in your analysis. I should be able to run this file and get all the output that you got.

- A stand-alone R (or SAS) Data Step that will score new data as it becomes available. This should include all of the transformations and the regression equation from your analysis. **(40 Points)**

- A CSV file which has the scored records values from WINE_TEST. There will be only two columns in this file: INDEX, P_TARGET *[with 3335 rows, not counting header row]*. You will be graded on how your model performs versus my model and those of other students in the class. I will award bonus points for high accuracy in your P_TARGET model. **(50 Points)**

WRITE UP (160 POINTS):

1. DATA EXPLORATION (40 points)

Describe the size and the variables in the WINE data set so that a manager can understand it. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. 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"?

2. DATA PREPARATION (40 Points)

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some 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 (40 Points)

Build at least five different models using the procedures:

- Poisson Regression
- Zero-Inflated Poisson Regression
- Negative Binomial Regression
- Zero-Inflated Negative Binomial Regression
- Multiple Linear Regression (similar to Unit 1 – Moneyball)

In SAS the procedures are:
PROC GENMOD and PROC REG. The five models will be:

- GENMOD with Poisson distribution
- GENMOD with Negative Binomial distribution
- GENMOD with Zero Inflated Poisson distribution
- GENMOD with Zero Inflated Negative Binomial distribution
- REGRESSION (use standard PROC REG and if you wish you may use a variable selection method)

Sometimes Poisson and Negative Binomial models give the same results. If that is the case, comment on that. Consider changing the input variables if that occurs so that you get different models.

You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

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 POISSON model, but a negative effect in my REGRESSION model".

4. SELECT MODELS (40 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 POISSON 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 poisson, negative binomial, Zero Inflated Poisson, or Zero Inflated Negative Binomial model for model deployment.

STAND ALONE SCORING PROGRAM (40 POINTS)

WRITE MODEL DEPLOYMENT CODE (40 Points)

Write a Stand Alone data step that will score new data and predict the number of wine cases that will be sold based upon the qualities of the wine. This data step must be in the Poisson family (you may also use a logistic model / poisson model combination. The variable should be named:

P_TARGET

The data step will need to include:

a. All the variable transformations such as fixing missing values
b. The poisson regression formulas

Note: If you decide to do a LOGISTIC/POISSON model, you might wish to do this with two data steps instead of one.

SCORED DATA FILE (50 POINTS)

SCORE THE WINE_TEST DATA SET (50 Points)

Use the stand alone program that you wrote in the previous section. Score the data file WINE_TEST. 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). This number should be greater than 0.

Your values will be compared against …

- A Perfect Model
- Instructor's Model
- 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 negative points
If your model is not better than generating a RANDOM value, you will receive a LOT of negative points
If your model is not better than the WORST model, then it will be a WHOLE LOT of negative points.

BINGO BONUS:

If you want Bingo 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 Bingo Bonus work, I cannot give you credit. Bingo 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 Bingo Bonus is: ***All Sales are Final !***

- (20 Points) Develop a LOGISTIC / POISSON model (if you like it, you may select this as your champion)
- (20 Points) Use decision tree software such as Angoss or Weka or something else for variable selection or missing value imputation (the more use you make of decision trees, the more points you will receive). Be sure to carefully present your decision tree output so that I can see what you did.
- (20 Points) Build a decision tree model to predict the number of cases sold. You cannot use this as your champion model, but comment on it and compare it to your Poisson models.
- (20 Points) Recreate as much of the program as you can in a second software (e.g., R and SAS)
- (?? 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
- (Lose ?? Points) For anything that I think might annoy your boss !