INSURANCE LOGISTIC REGRESSION PROJECT (250 Points)

This data set contains approximately 8000 records. Each record represents a customer at an auto insurance company. Each record has two target variables. The first target variable, TARGET_FLAG, is a 1 or a 0. A "1" means that the person was in a car crash. A zero means that the person was not in a car crash.

The second target variable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero. Your second part of this assignment is to come up with a model to predict what it will cost you if a person does crash their car. This model is not a major component of this assignment, so you should not spend a lot of time on this part. You can use a simple average or some other simple model. It is not critically important to this exercise. Just make a reasonable estimate of losses (in other words, don't use a random number generator … use something that makes sense).

For this project, most of the points will be awarded for the LOGISTIC MODEL, so that is where you should spend your time.

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 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 SAS code you used in your analysis. I should be able to run this file and get all the output that you got.

- A SAS file, sas7bdat, which has the scored records values from INSURANCE_TEST. There will be only three columns in this file: INDEX, P_TARGET_FLAG and P_TARGET_AMT. You will be graded on how your LOGISTIC REGRESSION 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_AMT model. **(50 Points)**

WRITE UP (200 POINTS):

1. DATA EXPLORATION (40 points)

Describe the size and the variables in the INSURANCE 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 do. These are just ideas.

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

2. DATA PREPARATION (40 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 (40 Points)

Build at least three different LOGISTIC REGRESSION (or PROBIT REGRESSION if you really want to go wild!) using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, 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. Show all of your models and the statistical significance of the input variables.

Discuss the coefficients in the model, do they make sense? For example, if a person has a lot of traffic tickets, you would reasonably expect that person to have more car crashes. If the coefficient is negative (suggesting that the person is a safer driver), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why?

4. SELECT MODELS (40 Points)

Decide on the criteria for selecting the "Best Model". Will you use a metric such as LOG LIKELIHOOD, AIC, or ROC CURVE? Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model. **IMPORTANT: For all three models, you must generate a ROC CURVE and provide the KS STATISTIC.**

It helps if you put the results in a table showing the metrics vs the models for comparison.

5. WRITE MODEL DEPLOYMENT CODE (40 Points)

Write a Stand Alone SAS data step that will score new data and predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. Make sure this is a section in your report. The variables should be named:

P_TARGET_FLAG
P_TARGET_AMT

The SAS data step will need to include:

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

Note: You might wish to do this with two data steps instead of one.

Late Addition to the assignment:

It is OK to copy and paste the SAS output showing the model parameters and bypass the data step. Just make sure you explain the model and the parameters.

SCORED DATA FILE (50 POINTS)

Use the stand alone program that you wrote in the previous section. Score the data file INSURANCE_TEST. Create a file that has only THREE variables for each record:

INDEX
P_TARGET_FLAG
P_TARGET_AMT

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_FLAG is the probability that a person will crash their car. It is a number between 0 and 1. The third number is the insurance damage assuming that a person does crash their car. This number should be greater than 0.

Your LOGISTIC 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.

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 !***

- (10 Points) Once you select a champion model in Step 4, use PROC GENMOD to do the LOGISTIC Regression. Are the results the same? Are there any differences?
- (5 Points) Use at least one PROBIT MODEL when building your logistic models.
- (?? 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 !