## INFS 762 Project

**Due: 10/12/2021 + a one-week grace period**

**You can work on this project individually or in a group of two people.**

**Case description**

A supermarket is offering a new line of organic products. The supermarket's management wants to determine which customers are likely to purchase these products.

The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of the loyalty program participants and collected data that includes whether these customers purchased any of the organic products.

The ORGANICS data set contains 13 variables and over 22,000 observations.

**Step 1.** You need to download the dataset “organics.csv” in D2l-> Content-> Project folder and import it to SAS.

The variables in the data set are shown below with the appropriate roles and levels:

#### Now you want to remove two variables “DemCluster” and “TargetAmount”. The id of a customer shouldn’t be considered in data mining; however you don’t need to remove it. In the table below, the measurement level “nominal” means that the variable is categorical, and “Interval” means the variable is continuous. The target or dependent variable is “targetBuy”.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Type | Model role | Description |
| ID | Nominal | ID | Customer loyalty ID |
| DemAffl | Interval | Input | Affluence grade on a scale from 1 to 30 |
| DemAge | Interval | Input | Age |
| DemCluster | Nominal | Rejected | Type of residential neighborhood |
| DemClusterGroup | Nominal | Input | Neighborhood group |
| DemGender | Nominal | Input | M, F, U |
| DemReg | Nominal | Input | Geo region |
| DemTVReg | Nominal | Input | TV region |
| PromClass | Nominal | Input | Loyalty status, tin, silver, gold or platinum |
| PromSpend | Interval | Input | Total amount spent |
| PromTime | Interval | Input | Time as a member |
| TargetBuy | Binary | Target | Organics purchased? 1 = yes 0 = no |
| TargetAmt | Interval | Rejected | Number of organic products purchased |

**STEP 2:** You to do a quality check. In this dataset, we do not have false/unreasonable values. You need to tell me: 1) which variables have a skewed distribution, and 2) which variables have missing values.

**For the continuous variables, you need to print the histogram for each variable and also check the extreme values.**

The sas code for creating histograms is below

libname target 'C:\Users\jliu2188\AAEM71-ZIP'; //you need to create your own library

goptions reset=global

gunit=pct

hsize= **10.625** in

vsize= **8.5** in

htitle=**4**

htext=**3**

vorigin=**0** in

horigin= **0** in

cback=white border

ctext=black

colors=(black blue green red yellow)

ftext=swiss

lfactor=**3**;

**proc** **univariate** data=target.organics noprint;

histogram promtime; //Please replace the variable name here

title 'histogram for promtime';

**run**;

Based on the histogram, you can determine if a continuous variable has a skewed distribution. Although it is possible to get a skewedness measure for each variable, histogram is more straightforward.

The sas code for checking the extreme and missing values is below.

ODS SELECT EXTREMEVALUES;

ODS select MissingValues;

**PROC** **UNIVARIATE** Data=target.organics NEXTRVAL=**10**;

VAR promtime;

**RUN**;

For each categorical variable, you need to use proc freq to find out how many different categories and how many missing each categorical/nominal variable (except ID) has.

**PROC** **freq** Data=target.organics;

table demreg;

**RUN**;

**STEP 3:** As I discussed in one of the lecture recordings, you are not required to do categorical/nominal variable dummy coding if you use SAS, but in this step, I will ask you to write SAS code for creating dummy variables. Please create dummy variables for the variable PromClass **(**Please remember you need to create k-1 dummies. PromClass include 4 categories; you hence need to create 3 dummy variables). You may choose to drop the variable “PromClass” after you create dummy variables for it. If you choose to keep the original variable, please remember not to include it when you fit your models. You don’t need to do dummy coding for the other categorical/nominal variables

**STEP 4.** You need to do missing value imputation. You want to replaced missing values for the interval (continuous) inputs with the input median (You need to use a SAS procedure called **proc means** to compute the median of the variable), and added unique imputation indicators for each input with missing values. For a categorical variable with missing values, you create a separate category for the missing values.

**STEP 5:** You need to randomly select 60% of the data for training and 40% for validation. Please refer to my SAS tutorial slides to find out how to create a random sample.

**STEP 6:** You need to use the stepwise logistic regression method for variable selection and tell me what variables have been selected.

Now you are done with variable processing, you can export your sas dataset using the following code (Now you have a training dataset and a validation dataset. Please remember to export both).

**proc** **export** data=target.organics

outfile='C:\Users\jliu2188\organics.csv'//you need to specify the output file name here.

dbms=csv replace;

**run**;

**STEP 7:** Based on the selected variables, you fit two additional models (e.g, neural network, random forest or SVM, but please not use a decision tree model). You need to use logistic regression with the selected variables as your baseline model and compare it with the two models you selected. Please create a table to show the precision/recall/accuracy of the three models including logistic regression and the two models you selected. You need to show the measures based on the validation dataset. When you fit the two additional models, remember to just use the variables that has been select in step 6. You can use Weka to fit your models (download: http://www.cs.waikato.ac.nz/ml/weka/downloading.html). You can watch my Weka tutorial and you can also find the Weka tutorial online such as https://www.youtube.com/watch?v=m7kpIBGEdkI. The dataset used in the tutorial is of the ARFF format. You can also load a cvs file to Weka. Please tell me which model you think is the best model for the problem.

### Step 8. In step 2, you have identified some variables with skewed distribution. Such distributions create high leverage points that can distort an input’s association with the target. Let’s now modify these variables before fitting the stepwise regression. You can take a natural log of the variables with the skewed distribution. However, logarithm of zero, log(0), is not defined. You can always do log(x+1), where x is the variable value you want to transform.

**Step 9**: Please use the transformed variables and use the stepwise method for variable selection and tell me what variables have been selected. (If you have transformed a variable using log, you should consider just the log transformed variable, and the original one should be ignored in this run of variable selection)

**Step 10**. Please use the variable you selected in Step 9 to fit the same three models you used in step 7. Please create a table to show the precision/recall/accuracy of the three models. Please tell me which model works the best, and if transforming the variables helps achieve better results.

**Below is a summary of things you need to submit:**

1. All SAS code
2. Histograms
3. Tell me the variables that have missing values and variables that have skewed distribution.
4. Tell me which variables you selected in the first round of variable selection.
5. Use the variables you selected in the previous step to do model fitting. Tell me which three models (including logistic regression) you used.
6. Submit a table that shows the precision/recall/accuracy of the three models.
7. After you transformed the variables with skewed distribution, please do a variable selection again and tell me which variables you selected. Fit the three models and submit a table that shows the precision/recall/accuracy of the three models.
8. Please tell if log transformation results in better model performance.
9. Nothing else.

**If you worked on the project in a group, both of you need to submit the same project doc to the project dropbox (You will get the same score for the project document). Meanwhile, you also need to submit your peer evaluation in a separate word document. Your peer evaluation two short paragraphs. Please tell me what you have done in the first paragraph, and what your partner has done in the second paragraph. The peer evaluation accounts for 5 points.**