**EXAM on Classification Model Building**

The purpose of this exam is to test student’s capability of building a classification model and utilize this model (80 Points) to find optimal decision on one specific application (20 Points each).

**Data:**

* Training Data: The data set used in this exam is “TRAIN.CSV”. There is one binary target variable in this data set. It is a food data set to test your understanding on classification model building.
* Scoring Data: The scoring data set is “SCORE.CSV”. This data set has all predictors of the training sample except the target variable.
* Data Dictionary: The data dictionary for this study is the file entitled “Data\_Dictionary”.
* Sample Result Data Set: After the completion of your analysis, you can turn in your results using the format which is like the data set “SAMPLE.CSV”. Five columns are in this sample file: “ID”, “Probability” and “PRED\_TARGET”.

**PART I Classification Model Building**

Build a classification model using the training data. The target variable for this model is “TARGET” and the meaning for each predictor is not important and you can refer to the dictionary to understand these predictors. After complete your model, you need to score the “Scoring Data” and report the “Probability” of being Positive (i.e., TARGET = 1). This is an imbalance data set; you might need to perform over-sampling or under-sampling to achieve model with high performance. Your classification model will be graded using the “AUC” (c statistics). The “AUC” will be calculated using the column “Probability” in your resulting file. Assume that the AUC for a typical model for this set of data is in the interval (0.63, 0.65), models built by you will get 80 points if your model AUC is higher than 0.65. All other models will get partial credit calculated using the formula +40.

**PART II Making decision with highest possible precision** (15 Points)

You need to make decision (i.e., the value of PRED\_TARGET be 1 or 0) to obtain the maximum revenue based on the information provided below:

1. The cost of targeting a customer is $1.00 (i.e., the value of “PRED\_TARGET” is 1).
2. The cost of not targeting a customer is $0 (i.e., the value of “PRED\_TARGET” is 0).
3. The revenue of targeting a customer who buy the product is $30.00 (i.e., both the values of “TRUE\_TARGET” and “PRED\_TARGET” are 1).
4. **NOTE: The “TRUE\_TARGET” is only available to the instructor. You need to use the model built in “PART I” to classify the status of “TRUE\_TARGET” as either “+” or “-“ (or either “1” or “-1”).**

Assume that the revenue for a “normal” decision is $100 and the revenue based on your decision is $99, the score for this part of exam is .

**PART III Explanation**

* You should get the following files: (1) “TRAIN.CSV”, the training data set; (2) “SCORE.CSV”, the scoring data set; (3) “SAMPLE.CSV”, the sample data set that needs to turn in for grading; and (4) “Data\_Dictionary.XLSX”.
* You need to turn in one file that has the scoring results from your analysis. This file should be similar to the file, “SAMPLE.CSV”.