AA502 – Logistic Regression

Homework 2

You work for a major bank in the retail department. You are in charge of predicting which customers will buy our variable rate annuity insurance product. This analysis is a continuation of your previous one.

We have a sample of 10,619 customers who have been offered the product in the data set under the variable **INS** – 1 if they bought, 0 if not. The data was split into two pieces – training (**Insurance_hw1**) and validation (**Insurance_hw2_valid**). We have 18 other variables that describe the customers' attributes **before** they were offered the new insurance product. There is a mix of categorical and ordinal predictors with labels that should serve as a data dictionary. Be sure to include Branch of Bank in this analysis.

Develop a report that tries to model the probability of whether the customer purchased our insurance product. Use an $\boldsymbol{\alpha = 0.001}$ for the entire assignment. Make sure that your report addresses the following issues:

- Using all of the variables (including Branch of Bank) find the best model possible on the **training data set.** Use information from your first assignment, as well as ROC curves (do not go above 2-way interactions).

- Show me the ROC curve for your final model. (Make sure the axes and titles look professional.)

- ROC curves give us insight into which probability cut-off value may be the best choice for prediction in a model. There are many techniques to determine which cut-off value should be used (cost structures should always be taken into consideration if the information is possible). We will use the Youden Index to help us determine which cut-off should be used to determine if a person is thought to have a high enough probability to buy our new product. Use the final model from your exploration. Use the following model statement in PROC LOGISTIC (with your model put in for the "…" piece):

    ```
    model INS(event="1") = ... / outroc=ROC;
    ```

  The OUTROC= option in the model statement creates a data set (named ROC in the example above) that contains the following variables.

    - **_PROB_**, the estimated probability of an event. These estimated probabilities serve as cutpoints for predicting the response. Any observation with an estimated event probability that exceeds or equals

PROB is predicted to be an event; otherwise, it is predicted to be a nonevent.

  - **_POS_**, the number of correctly predicted event responses
  - **_NEG_**, the number of correctly predicted nonevent responses
  - **_FALPOS_**, the number of falsely predicted event responses
  - **_FALNEG_**, the number of falsely predicted nonevent responses
  - **_SENSIT_**, the sensitivity, which is the proportion of event observations that were predicted to have an event response
  - **_1MSPEC_**, one minus specificity, which is the proportion of nonevent observations that were predicted to have an event response

Create another data set named Youden that contains all of these variables along with a new variable J defined as:

  - *J* = Sensitivity + Specificity – 1

This is Youden's J Statistic (Youden Index). The maximum of this J statistic provides the optimal cut-off point for the probability of the event. Look through your data set. What is the maximum value of the Youden Index? What is the corresponding probability cut-off value?

- Based on the cut-off level that you defined in the previous part, create a classification table of your results. How many false positives and false negatives do you have?

- Build a classification table of your results using your **validation data set**. How many false positives and false negatives do you have? What is your misclassification error rate? This should be the first number you report in your results section!