Homework #1: Due Friday (by 5:00 pm), September16

AA502 – Logistic Regression

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.

A variable annuity is a contract between you and an insurance company / bank, under which the insurer agrees to make periodic payments to you, beginning either immediately or at some future date. You purchase a variable annuity contract by making either a single purchase payment or a series of purchase payments.

A variable annuity offers a range of investment options. The value of your investment as a variable annuity owner will vary depending on the performance of the investment options you choose. The investment options for a variable annuity are typically mutual funds that invest in stocks, bonds, money market instruments, or some combination of the three. If you are interested in more information, see:

http://www.sec.gov/investor/pubs/varannty.htm

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_hw**) and validation (**not shown**). 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 continuous predictors (the continuous predictors have been binned) with labels that should serve as a data dictionary. The binned variables are explained on the next page.

For this assignment, use only the training data set. Use an $\boldsymbol{\alpha = 0.001}$ for the entire assignment.

Contigency Table Analysis:

- Create a table that displays all of the comparisons between each of the variables and our target variable. Your table should include only the test statistics for each comparison as well as the p-value. Rank these variables from smallest to largest p-value. Make sure you use the appropriate hypothesis tests! (Disregard the variable branches for this assignment).

- Create a separate table similar to the above one with only the significant variables.

- Create a table for the odds ratios from **only the significant binary relationships** with whether the customer bought our insurance product. Rank these variables from strongest to weakest relationship. Give an interpretation of the odds ratio for the strongest relationship.

- For all of the significant, non-binary relationships, calculate the appropriate statistic to measure the strength of an association between the variables and whether the customer bought our insurance product.

The marketing department thinks that an interaction might exist with other bank investment products (such as money market accounts and saving accounts).

- Conduct a stratified analysis on the relationship between having a savings account and buying our insurance product **controlling for** whether the customer has a money market account. Is there a significant relationship between having a savings account and whether or not individuals buy our insurance product? Calculate the adjusted (common) odds ratio and compare this to your previous results. Does a common odds ratio appear to be appropriate here?

- For the above stratified analysis, calculate the Breslow-Day-Tarone test statistic. Based on this analysis, is there any evidence of interaction between having a money market account and a savings account? What would an interaction between these two binary variables imply in terms of our problem?

Building the Logistic Model:

- Build a logistic regression model. Use reference coding for all categorical variables. Binary variables should have "0" as the reference level, while categorical variables that are not binary should have the last category (in alpha-numeric order) as the reference level. (Use various techniques, such as Backward Elimination [slstay=0.001], Forward Selection [slentry=0.001], etc…..don't forget model hierarchy).

    - Make sure you check variables and interactions (if any) for quasi-complete separation and adjust accordingly. This may change your model so adjust accordingly.

    - Select the best model.

Categories of Binned Variables (continued on next page):

DDABAL_bin (checking account balance binned)

DDABAL = 0	*DDABAL_Bin = 1*
DDABAL > 0 & DDABAL <=100	*DDABAL_Bin = 2*
DDABAL > 100 & DDABAL <=300	*DDABAL_Bin = 3*
DDABAL > 300 & DDABAL <=750	*DDABAL_Bin = 4*
DDABAL > 750 & DDABAL <=1250	*DDABAL_Bin = 5*
DDABAL > 1250 & DDABAL <=2250	*DDABAL_Bin = 6*
DDABAL > 2250 & DDABAL <=6000	*DDABAL_Bin = 7*
DDABAL > 6000	*DDABAL_Bin = 8*

ACCTAGE_bin (age of oldest account)

ACCTAGE = 20	*ACCTAGE_Bin = 1*
ACCTAGE > 20	*ACCTAGE_Bin = 2*
ACCTAGE = .	*ACCTAGE_Bin = 3*

DEPAMT_bin (amount deposited in checking account)

DEPAMT = 2	*DEPAMT_Bin = 1*
DEPAMT > 2 & DEPAMT <=700	*DEPAMT_Bin = 2*
DEPAMT > 700 & DEPAMT <=2200	*DEPAMT_Bin = 3*
DEPAMT > 2200 & DEPAMT <=6500	*DEPAMT_Bin = 4*
DEPAMT > 6500	*DEPAMT_Bin = 5*

CHECKS_bin (Number of checks written)

CHECKS = 0	*CHECKS_Bin = 1*
CHECKS > 0 & CHECKS <=2	*CHECKS_Bin = 2*
CHECKS > 2 & CHECKS <=4	*CHECKS_Bin = 3*
CHECKS > 4	*CHECKS_Bin = 4*

PHONE_bin (Number of telephone banking transactions)

PHONE = 0	*PHONE_Bin = 1*
PHONE > 0 & PHONE <=1	*PHONE_Bin = 2*
PHONE > 1	*PHONE_Bin = 3*
PHONE = .	*PHONE_Bin = 4*

TELLER_bin (Number of teller banking transactions)

TELLER = 0	*TELLER_Bin = 1*
TELLER > 0 & TELLER <=3	*TELLER_Bin = 2*
TELLER > 3	*TELLER_Bin = 3*

SAVBAL_bin (balance in savings account)

SAVBAL = 0	*SAVBAL_Bin = 1*
SAVBAL > 0 & SAVBAL <=50	*SAVBAL_Bin = 2*
SAVBAL > 50 & SAVBAL <=250	*SAVBAL_Bin = 3*
SAVBAL > 250 & SAVBAL <=1250	*SAVBAL_Bin = 4*
SAVBAL > 1250 & SAVBAL <=3000	*SAVBAL_Bin = 5*
SAVBAL > 3000 & SAVBAL <=8000	*SAVBAL_Bin = 6*
SAVBAL > 8000	*SAVBAL_Bin = 7*

ATMAMT_bin (amount withdrawn from ATM)

ATMAMT ≤ 5	*ATMAMT_Bin = 1*
ATMAMT > 5 & ATMAMT <=3750	*ATMAMT_Bin = 2*
ATMAMT > 3750	*ATMAMT_Bin = 3*

CDBAL_bin (amount in CD)

CDBAL ≤ 500	*CDBAL_Bin = 1*
CDBAL > 500 & CDBAL <=9200	*CDBAL_Bin = 2*
CDBAL > 9200	*CDBAL_Bin = 3*

IRABAL_bin (Number of IRA accounts)

IRABAL ≤ 1	*IRABAL_Bin = 1*
IRABAL > 1	*IRABAL_Bin = 2*

INVBAL_bin (Amount in Investments)

INVBAL ≤ 1000	*INVBAL_Bin = 1*
INVBAL > 1000 & INVBAL <=10000	*INVBAL_Bin = 2*
INVBAL > 10000	*INVBAL_Bin = 3*
INVBAL = 10000	*INVBAL_Bin = 4*

MMBAL_bin (Amount in money market)

MMBAL ≤ 1000	*MMBAL_Bin = 1*
MMBAL > 1000	*MMBAL_Bin = 2*