

INFO 3236 091 Spring 2021

Homework #4

Building a Logistic Regression Model with SAS Enterprise Miner

(Due Apr. 5, 2021)

Objectives

- to apply logistic regression model to a real-world problem
- to be able to interpret the logistic regression output

Tasks

1. Read the problem description below
2. Follow the instruction in this document to set up the project and data set in SAS Enterprise Miner.
3. Use SAS Enterprise Miner in Apporto. Use Chrome in Apporto to login to Canvas, download the data file (and save it in the shared folder in the virtual machine), and to submit the assignment.
4. Build logistic regression models in SAS Enterprise Miner for the *develop* dataset to **predict the likelihood of customers purchasing a variable annuity (insurance product)**.
5. Write up the answers in the template provided, replace 'ID' in the filename with your *Niner login*.
6. Submit the following files back to Canvas before the due date.
 - a. Your answer in Word document.
 - b. Logistic Regression output file (Output window)
 - c. Your project folder – zipped format (right click your project folder and select compress file to zip format)

Problem Description

A bank seeks to increase sales of a variable annuity product. To do this, the bank will send product offers to existing banking customers. However, to maximize profits, the bank wants to be selective about whom it targets. This selectivity will be achieved by constructing a predictive model.

To achieve the bank's analytic objective, an analysis data set was assembled. A data set was created with 32,264 cases (banking customers) and 47 input variables. The binary target variable, **Ins**, indicates whether the customer bought the variable annuity product. The 47 input variables represent other product usages in a three-month period and demographics. Two of the inputs are nominally scaled. The others are interval or binary.

The data are in a SAS dataset called **develop**. The dataset was assembled from several source tables within the bank's data warehouse. The source tables include the customer

master table, the transaction detail table, the product detail table, and a third-party demographic overlay table.

About half of the variables have some missing values. Many of the variables, especially those relating to monetary amounts, have an extremely large range and highly skewed distribution.

The BRANCH variable, a nominal input with 19 distinct levels, indicates the branch in which the customer's initial account was opened. The RES variable, a nominal input with three distinct levels, classifies the customer's primary residence as rural, suburban, or urban.

The target variable for this analysis, INS, indicates acquisition of the variable annuity over a fixed period of time. While overall acquisition rate is about 2%, the acquisition rate in the raw analysis data is more than 34%. This reflects the separate sampling used to generate the raw data.

These are the variables in the data set:

Ins	purchase variable annuity account (1=yes, 0=no)
AcctAge	age of oldest account in years
DDA	checking account (1=yes, 0=no)
DDABal	checking account balance
Dep	number of checking deposits
DepAmt	amount deposited
CashBk	number of times customer received cash back
Checks	number of checks
DirDep	direct deposit (1=yes, 0=no)
NSF	occurrence of insufficient funds (1=yes, 0=no)
NSFAmt	amount of insufficient funds
Phone	number of times a customer used telephone banking
Teller	number of teller visits
Sav	savings account (1=yes, 0=no)
SavBal	savings balance
ATM	used ATM service (1=yes, 0=no)

This material has been adapted for INFO3236 from SAS Advanced Business Analytics, SAS Institute Inc., Cary, NC, and other online sources. To be used only for INFO3236 class. Please do not copy or distribute outside of this class.

ATMAmt	ATM withdrawal amount
POS	number of point of sale transactions
POSAmt	amount in point of sale transactions
CD	has certificate of deposit (1=yes, 0=no)
CDBal	certificate of deposit balance
IRA	has retirement account (1=yes, 0=no)
IRABal	retirement account balance
LOC	has line of credit (1=yes, 0=no)
LOCBal	line of credit balance
Inv	has investment account (1=yes, 0=no)
InvBal	investment account balance
ILS	has installment loan (1=yes, 0=no)
ILSBal	installment loan balance
MM	has money market account (1=yes, 0=no)
MMBal	money market balance
MMCred	number of money market credits
MTG	has mortgage account (1=yes, 0=no)
MTGBal	mortgage balance
CC	has credit card account (1=yes, 0=no)
CCBal	credit card balance
CCPurc	number of credit card purchases
SDB	has a safety deposit box (1=yes, 0=no)
Income	income in thousands of dollars
HMOwn	owns home (1=yes, 0=no)
LORes	length of residence in years
HMVal	home value in thousands of dollars
Age	age of customer in years

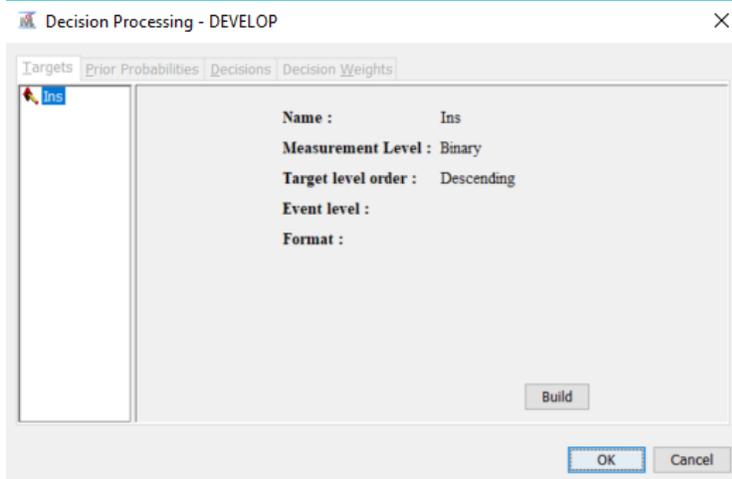
This material has been adapted for INFO3236 from SAS Advanced Business Analytics, SAS Institute Inc., Cary, NC, and other online sources. To be used only for INFO3236 class. Please do not copy or distribute outside of this class.

CRScore	credit score
Moved	recent address change (1=yes, 0=no)
InArea	local address (1=yes, 0=no)
Res	area classification (R=rural, S=suburb, U=urban)
Branch	branch of bank (B1 – B19)

Setting Up a Project for the Annuity Data Set

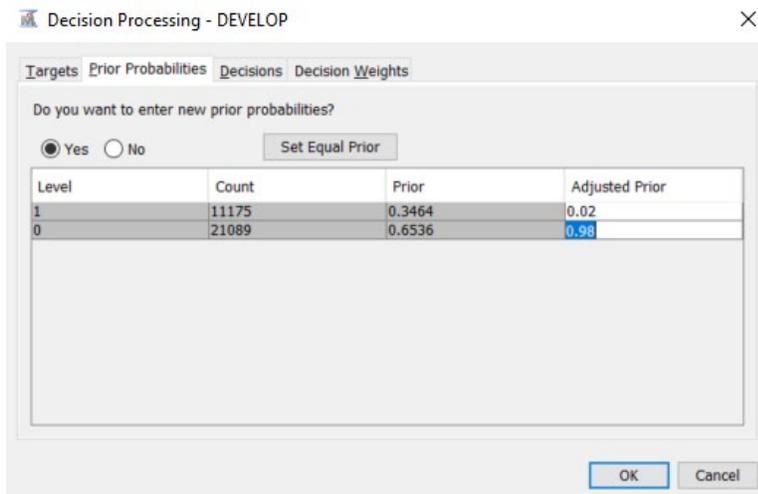
1. Download the **develop** SAS data file from Canvas.
2. Save the file in your library folder (pick your own location)
3. Launch SAS Enterprise Miner. Create a new project and a new library. See the handout posted in Canvas on Creating SAS Project if you forgot how to complete this task.
4. Add a data source (**develop** data file)
 - a. Step 1, select SAS Table and click Next
 - b. Step 2, double click your library and select Develop and click OK. Then you should see *[name of your library].DEVELOP* in the Table field, click Next
 - c. Step 3, click Next
 - d. Step 4, select Advanced and click Next
 - e. Step 5, click Next
 - f. Step 6, you should see 48 columns and 32264 rows. Click Next.
 - g. Step 7, the role is Raw. Click Next.
 - h. Step 8, click Finish.
5. Create a new diagram.
6. Drag and drop the develop data source on your diagram workspace.
7. Select the DEVELOP data source node. View the Variables property by clicking the ellipsis button next to the Variables. Update the roles and levels of the **Ins** variable to Target and binary if needed. Click OK to close the window.
8. Decision Processing: select the DEVELOP data source node still and view the Decision property by clicking the ellipsis button next to the Decisions.

Note: The information given above enables you to create a new assessment statistics for this homework. This window also enables you to adjust for oversampling.



Select the button Build. The tabs are now active.

Select Prior Probabilities ⇒ Yes and enter .02 and .98 for Adjusted Prior.



- Use a partition of 50% for training and 50% for validation. Add a data partition node (Sample and Data Partition). Connect the data partition node to the DEVELOP data source node. Select Data Partition and click the ellipsis button next to the Partitioning Method and select **Stratified** method. Enter 50 for Training and 50 for Validation.

Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Stratified
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	10/6/18 5:01 PM
Run ID	

10. Add a Regression node to the diagram workspace and connect it to the Data Partition node. Use the “Stepwise” method for model selection and “Validation Error” as the selection criteria. After running the model, answer the questions in **HW4_Report_ID.docx**.

Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	

11. Save the output for submission.
12. Save and close your SAS Enterprise Miner project.