- EXERCISE -

Decision Tree

Goal: In this exercise, we will focus on healthcare. The data set for this exercise is made available by the Department of Health of the state of New York. It includes deidentified patient discharge data from multiple health facilities in 2015. (Each row in the data set pertains to one patient. There is a total of 10,552 patients in the data set). Our goal is to predict how long a patient will stay in a facility after they are admitted. For the binary classification task, we will try to predict whether a patient will stay 5 or more days (coded as 1), or not (coded as 0). This is important, because most health facilities try to reduce average length of stay since this is an important performance metric. In this task, we use 5 days as a benchmark because it is the national average.

Start SAS Enterprise Miner:

1) Start SAS Enterprise Miner and open your project called "Exercises and Assignments".

2) We will use the "**Hospital**" data set to complete this exercise. To see whether this data set is already in this project, click on the plus sign next to "Data Sources" and check the list of available data sets in this project. If the data set does not exist, please follow the instructions in the "Importing a New Data Set" section to import this data set.

Description of the variables (columns) in the data set:

3) Please see the following table for the variables and their descriptions.

Name	Description	Role	Level
APR_DRG_Code	The APR-DRG Classification Code	INPUT	NOMINAL
APR_DRG_Description	The APR-DRG Classification Code Description in Calendar Year 2011, Version 28 of the APR-DRG Grouper. http://www.health.ny.gov/statistics/sparcs/sysdoc/appy.htm	REJECTED	NOMINAL
APR_MDC_Code	All Patient Refined Major Diagnostic Category (APR MDC) Code. APR-DRG Codes 001-006 and 950-956 may group to more than one MDC Code. All other APR DRGs group to one MDC category.	INPUT	NOMINAL
APR_MDC_Description	All Patient Refined Major Diagnostic Category (APR MDC) Description.	REJECTED	NOMINAL
APR_Medical_Surgical_Description	The APR-DRG specific classification: Medical, Surgical, Not Applicable.	REJECTED	NOMINAL
APR_Risk_of_Mortality	All Patient Refined Risk of Mortality (APR ROM): Minor, Moderate, Major, Extreme.	INPUT	NOMINAL
APR_Severity_of_Illness_Code	The APR-DRG Severity of Illness Code: 1, 2, 3, 4	INPUT	INTERVAL
APR_Severity_of_Illness_Descript	All Patient Refined Severity of Illness (APR SOI) Description. Minor (1), Moderate (2), Major (3), Extreme (4).	REJECTED	NOMINAL
Abortion_Edit_Indicator	A flag to indicate if the discharge record contains any indication of abortion ("N" = No; "Y" = Yes).	INPUT	NOMINAL
Age_Group	Age in years at time of discharge. Grouped into the following age groups: 0 to 17, 18 to 29, 30 to 49, 50 to 69, and 70 or older.	INPUT	NOMINAL
Birth_Weight	The neonate birth weight in grams; rounded to nearest 100g.	INPUT	INTERVAL
CCS_Diagnosis_Code	AHRQ Clinical Classification Software (CCS) Diagnosis Category Code. More information on the CCS system may be found at the direct link: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp	INPUT	NOMINAL
CCS_Diagnosis_Description	AHRQ Clinical Classification Software (CCS) Diagnosis Category Description. More information on the CCS system may be found at the direct link: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp	REJECTED	NOMINAL

CCS_Procedure_Code	AHRQ Clinical Classification Software (CCS) ICD-9 Procedure Category Code. More information on the CCS system may be found at the direct link: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp	INPUT	NOMINAL
CCS_Procedure_Description	AHRQ Clinical Classification Software (CCS) ICD-9 Procedure Category Description. More information on the CCS system may be found at the direct link: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp	REJECTED	NOMINAL
Emergency_Department_Indicator	The Emergency Department Indicator is set based on the submitted revenue codes. If the record contained an Emergency Department revenue code of 045X, the indicator is set to "Y", otherwise it will be “N”.	INPUT	NOMINAL
Ethnicity	Patient ethnicity: Spanish/Hispanic Origin, Not of Spanish/Hispanic Origin, Multi, Unknown.	INPUT	NOMINAL
Facility_Id	Permanent Facility Identifier. Blank for abortion records.	ID	NOMINAL
Facility_Name	The name of the facility where services were performed based on the Permanent Facility Identifier (PFI), as maintained by the NYSDOH Division of Health Facility Planning. For abortion records ‘Abortion Record – Facility Name Redacted’ appears.	INPUT	NOMINAL
Gender	Patient gender: (M) Male, (F) Female, (U) Unknown.	INPUT	NOMINAL
Health_Service_Area	A description of the Health Service Area (HSA) in which the hospital is located. Blank for abortion records. Capital/Adirondack, Central NY, Finger Lakes, Hudson Valley, Long Island, New York City, Southern Tier, Western NY.	INPUT	NOMINAL
Hospital_County	A description of the county in which the hospital is located. Blank for abortion records.	INPUT	NOMINAL
ID	Patient ID	ID	NOMINAL
Length_of_Stay	The total number of patient days at an acute level and/or other than acute care level (excluding leave of absence days) (Discharge Date - Admission Date) + 1.	REJECTED	INTERVAL
LOS_gte5	Whether the patient days are greater than or equal to 5 days (we consider 5 days as a generally accepted national benchmark). 1=Yes, 0=No.	TARGET	BINARY
Patient_Disposition	The patient's destination or status upon discharge.	INPUT	NOMINAL
Payment_Typology_1	A description of the type of payment for this occurrence.	INPUT	NOMINAL
Payment_Typology_2	A description of the type of payment for this occurrence.	INPUT	NOMINAL
Payment_Typology_3	A description of the type of payment for this occurrence.	INPUT	NOMINAL
Race	Patient race: Black/African American, Multi, Other Race, Unknown, White. Other Race includes Native Americans and Asian/Pacific Islander.	INPUT	NOMINAL
Type_of_Admission	A description of the manner in which the patient was admitted to the health care facility: Elective, Emergency, Newborn, Not Available, Trauma, Urgent.	INPUT	NOMINAL
Zip_Code_3_digits	The first three digits of the patient's zip code. Blank for: - population size less than 20,000 - abortion records, or - cell size less than 10 on population classification strata. “OOS” are Out of State zip codes.	INPUT	NOMINAL

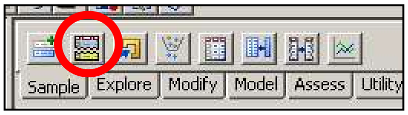
4) Right-click on the data set, select "Edit Variables" and set the levels of the variables as shown above using the dropdown menus. Select OK and close the Variables view when you are done.

Decision Tree for Binary Classification:

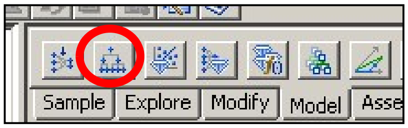
5) Right click on the "Diagrams" and select "Create Diagram". For "Diagram name" type "Decision Tree Classification".

6) Click and drag the data source onto the diagram.

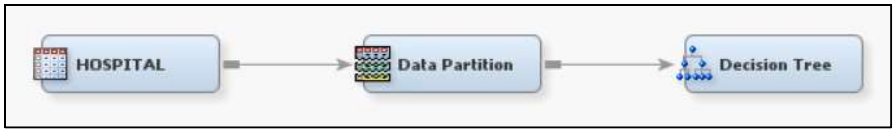
7) Click on the "Sample" tab on the toolbar and click and drag "Data Partition" (the second from left) onto the diagram.



8) Click on the "Model" tab on the toolbar and click and drag the "Decision Tree" node onto the diagram.



9) Connect the nodes as follows:



10) Click on the Data Partition node to open its properties in the left pane. Make sure training, validation, and test data set allocations are 40, 30, and 30 respectively.

11) Right click on the Decision Tree node and select "Run".

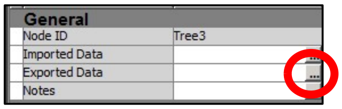
12) When the dialog box appears (after the nodes run), click on OK to close the Run Status window.

Let's check the predictions:

13) Before we move further, let's see the predictions made by the model:

a. Click on the “Decision Tree” node on the diagram.

b. In the properties window on the left, click on the ellipsis for “Exported Data”.



General

Node ID	Tree3
Imported Data	...
Exported Data	...
Notes	...

c. In the window that opens, click on the row for “VALIDATE”. Then click on Browse.

d. The window that opens shows the validation data set. Each row represents one patient in the validation data set. Scroll toward right.

e. Locate the following columns in this window for the first row:

i. “Predicted: LOS_gte5=0” : ______
(This is the probability that this observation belongs to 0)

ii. “Predicted: LOS_gte5=1” : _______
(This is the probability that this observation belongs to 1)

iii. “From: LOS_gte5” : ______
(This is the original classification of the observation)

iv. “Into: LOS_gte5” : ______
(This is what the model predicts as the new classification)

14) Now, let’s make sense of these values:

a. The values in (i) and (ii) should add up to 1 (or 100%).

b. Using a cut-off value of 0.5 (or 50%), the classification is determined based on which probability is greater than 0.5

c. If (iii) is different from (iv), it means the observation is predicted incorrectly. Otherwise, the prediction is correct.

d. The column called “Node” provides the node number in the decision tree that makes this prediction.

15) Close the table view and close the “Exported Data” window.

Now, Let’s Check the Results:

16) Right click on the Decision Tree node and select "Results…".

17) Let's identify the misclassification rate:
(Note: accuracy can be calculated by subtracting the misclassification rate from 1)

    a. Look at the "Fit Statistics" window.

    b. Under "Fit Statistics" column, find "_MISC_" (this is the misclassification rate)

    c. What are the misclassification rates of:

        i. Training set: _______

        ii. Validation set: _______

        iii. Test set: _________

18) Is there any evidence of overfitting? (Remember: overfitting exists when the model performs well on training, but not so well in validation/test sets) _________

19) Close the Results window. Let's identify the baseline misclassification rate:

    a. Click on the data node (i.e., Hospital) on the diagram to open its properties in the left pane.

    b. Click on the ellipsis for "Variables"

    c. Click on the variable name "LOS_gte5" and click "Explore".

    d. In the “Sample Properties” window, click on the dropdown for “Fetch Size” and change it to “Max,” then click “Apply”. (This will make sure that all results are displayed. Otherwise, this view shows up to 10,000 records only)

    e. The bar chart (called " LOS_gte5") shows the number of patients who stayed less than or equal to 4 days (coded as 0) vs. 5 days or more (coded as 1).

    f. Hovering the mouse over each bar shows the number of patients in each category.

    g. How many "1s" are there? ______

    h. How many "0s" are there? ______

    i. Identify the "majority" class (i.e., the class that has the most observations) ______

    j. Calculate the percentage of majority class in the data set: _____

    k. This percentage is your baseline accuracy. (Because, if you “naively” predicted all observations to be in this category, you would be correct/accurate by this much.)

    l. Remember, if you subtract baseline accuracy from 1 (or 100%), you get the baseline misclassification rate. So, what is the baseline misclassification rate? _________

    m. Close the "Explore" window and the "Variables" window.

20) Compare the misclassification rate of the decision tree (for validation) with baseline misclassification. Is the model better/worse? _________

Let’s examine the lift chart:

21) Open the results window again: Right-click on the decision tree node, and select "Results…"

22) Look at the window called "Score Rankings Overlay: LOS_gte5". This is the cumulative lift chart of the model.

    a. This chart can also be used to determine whether there is overfitting or not. Is there any evidence of overfitting? __________

Let's look at the misclassification matrix:

23) Scroll down in the "Output" window and find the section "Event Classification Table". Look at the section for "Data Role=VALIDATE" (this section is for the validation data set).

24) This is the classification matrix. You can observe the model's effectiveness by looking at the classifications. What are the following values?

    a. False negative: ___

    b. False positive: ___

    c. True positive: ___

    d. True negative: ___

Let's observe the tree:

25) Maximize the window called "Tree". Right click anywhere in the window, select "View" then "75%". This helps us view the tree better. (You can go back to 100% using the same menu if you want.)

26) The first node is called the root node. The first splitting variable is "APR_Severity_of_Illness_Code". There are two branches:

    a. if APR_Severity_of_Illness_Code is >=2.5,

    b. if APR_Severity_of_Illness_Code is <2.5 or Missing.

27) From each branch, there are other branches based on other variables.

Let's classify a new patient (manually):

28) If a new patient has the following values: APR_Severity_of_Illness_Code = 4, Patient_Disposition = "HOME OR SELF CARE"; in which node would the patient be? (Provide the node id) _________

29) Is this a "leaf" (or terminal) node, or a "decision" node? __________

30) Since this is a "leaf" node, it can be used for making predictions. Remember: we use the "majority rule" to make predictions. Use the Train column in this node: which class is the majority, 0 or 1? ________ (this would be the prediction for the new patient)

31) If a new patient has the following values: APR_Severity_of_Illness_Code = 2 and APR_DRG_Code=144; what would be the prediction for this patient (hint: follow the nodes in the tree view; use the majority class in the train column for prediction)? ____

Let's look at the rules:

32) Click on "View" and select "Model" and "Node Rules". Look at the section that starts as "Node 7":

    a. What does the rule state (until "then")? _____________________

    b. How many observations are there in this node? _________

    c. What is the value for "Predicted: LOS_gte5=1"? (This would be the percentage of patients whose length of stay is 5 or more days in this node) __________

    d. What is the value for "Predicted: LOS_gte5=0"? (This would be the percentage of patients whose length of stay is 4 or fewer days in this node) _________

    e. Which type of patient is in the majority? ________

    f. (Your answer to the previous question indicates that if a new patient satisfies this rule, the patient will be predicted as such.)

Let's change the default values:

33) Close the Results window.

34) Click on the decision tree node. In the properties pane on the left: change the "Maximum Depth" to 2. This determines how far the tree can go down. (So, you are pruning the tree after 2 levels.) Rerun the node and view the results.

35) In the Results window, maximize the "Tree" window. You should see a much smaller tree because the new setting prematurely pruned the tree.

36) Close the "Tree" window.

37) What are the misclassification rates for:

    a. Training set: ________

    b. Validation set: _______

    c. Test set: ___________

38) Is the misclassification rate of the validation set better than baseline? ____________

39) For the validation set: Is the misclassification rate better than that of the previous model? ______

40) Close the Results window.

41) Change the value of "Maximum Depth" back to 6.

42) Now, change the value of "Maximum Branch" to 3. (This allows a three-way split instead of two where needed.) Rerun the node and view the results.

43) What are the misclassification rates of:

    a. Training set: ________

    b. Validation set: _______

    c. Test set: ___________

44) Is the value for the validation set better/worse than the earlier model? __________

45) Maximize the "Tree" window and observe the tree. How many splits are there from the root node? __

Note: To increase the accuracy of your decision tree, you can change the following settings:

- Maximum depth: the number of levels the tree can go down (be careful: increasing this number can cause overfitting)
- Leaf size: the minimum number of training observations that are allowed in a leaf node. Permissible values are integers greater than or equal to 1. The default setting is 5.
- Split size: the smallest number of training observations that a node must have before it can be split further. Permissible values are integers greater than or equal to 2.

Variable Importance

46) Click on View, Model, and Variable Importance. This shows which variables were used in the decision tree.

47) Identify the variable whose "Importance" value is 1. _____________

48) The variables that have a zero for "number of splitting rules" are not used in the decision tree at all. Therefore, these variables have an "Importance" of zero. You can use this for variable selection and eliminate these variables.

49) Close the Results window.

Model Comparison:

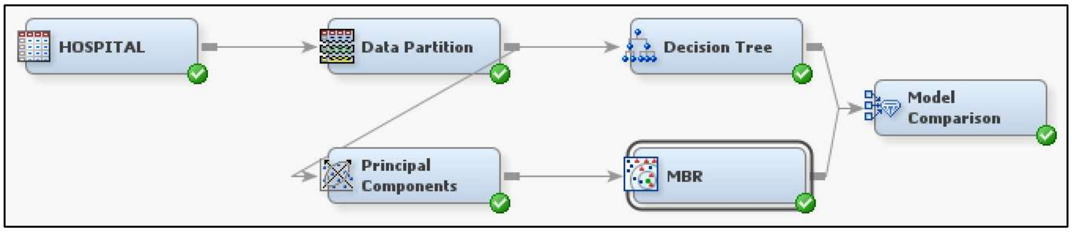
50) Create an MBR model as follows:

   a. Click on the “Modify” tab on the toolbar and click and drag the “Principal Components” node to the diagram.

   b. Click on the "Model" tab on the toolbar and click and drag the “MBR” node to the diagram.

   c. Click on the “Assess” tab on the toolbar and click and drag the “Model Comparison” node to the diagram.

51) Then create the following process flow:



52) Right click on the Model Comparison node and click Run. View the results when it is complete.

53) Which model did SAS select for you? (Hint: the selected model is shown with "Y" under the column "Selected Model"). ______

54) SAS uses the validation misclassification rate to make this decision. In the Fit Statistics window, look at the column labeled "Selection Criterion: Valid: Misclassification Rate". Verify whether this value is lower than the values of other models. ________

55) Close the Results window.

Decision Tree for Regression (Predicting a Continuous Variable):

56) Add a new drawing: right click on the "Diagrams" and select "Create Diagram". For "Diagram name" type "Decision Tree Regression".

57) Click and drag the Hospital data source onto the diagram.

58) We'll have to set a new target variable. In this case, we want to predict the actual “length of stay,” which is an interval (continuous) value. Therefore:

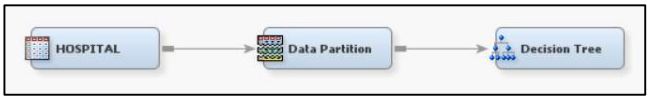
   a. Click on the data set on the diagram. In the properties window on the left, click on the ellipsis for "Variables"

   b. Change the Role of “LOS_gte5” to Rejected.

   c. Change the Role of “Length of Stay” to Target.

   d. Click OK to close the Variable view.

59) Create a model using this data source as follows:



60) Right click on the decision tree node and select "Run".

61) When the dialog box appears (after the nodes run), click on OK to close the Run Status window.

Let's check the predictions:

62) Before we move further, let's see the predictions made by the model: Click on the Decision Tree node, click on the ellipsis of Exported Data (in the left pane), click on Validate, and click on Browse to bring up the validation data set.

63) Look at the first observation (i.e., first row). Scroll all the way to right. What are the values for:

    a. Predicted: Length_of_stay = __________ (this is the predicted value)

    b. Length_of_stay = __________ (this is the actual value)

    c. Residual: Length_of_stay = __________ (this is the difference between actual and predicted)

64) Close the window for the validation data set. Then, close the “Exported Data – Decision Tree” window to go back to the diagram.

Let’s Check the Results:

65) Right click on the Decision Tree node and select "Results…".

66) Look at the Fit Statistics window. Under "Fit Statistics" column, find "_ASE_" (this is the Average Squared Error value). What is the average squared error for validation? _________

Predict the Length of Stay for a new patient (manually):

67) Maximize the window called "Tree". (If you want, you can set the view to 75% to see the tree better).

68) What is the first splitting variable from the root node? _______________

69) If a new patient has the following value, APR_Severity_of_Illness_Code=5, APR_DRG_Code=710, GENDER=M in which node would the patient be? (Provide the node id) _________

70) We use the "average" value to predict the target variable. Use the Train column in the node you identified in the previous step: what is the value of Average? ________ (this would be the value of the prediction for this new observation)

Variable Importance

71) Click on View, Model, and Variable Importance. This shows which variables were used in the decision tree.

72) How many variables are "Important" (in other words, how many variables have an "Importance" value greater than zero? ______ (This means, you can drop the remaining unimportant variables from future analyses.)

73) Close the Results window.

Model Comparison

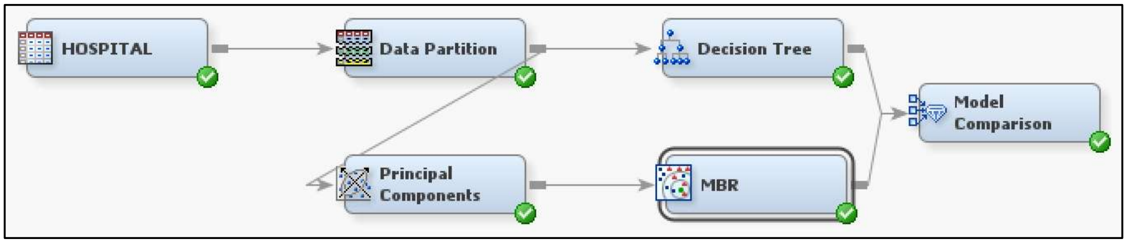
74) Create an MBR model as follows:

    a. Click on the “Modify” tab on the toolbar and click and drag the “Principal Components” node to the diagram.

    b. Click on the "Model" tab on the toolbar and click and drag the “MBR” node to the diagram.

    c. Click on the “Assess” tab on the toolbar and click and drag the “Model Comparison” node to the diagram.

75) Then create the following process flow:



76) Right click on the Model Comparison node and click Run. View the results when it is complete.

77) Which model did SAS select for you? (Hint: the selected model is shown with "Y" under the column "Selected Model"). _______

78) SAS uses the average squared error of the validation set to make this decision. In the Fit Statistics window, look at the column labeled "Selection Criterion: Valid: Average Squared Error". Verify whether this value is lower than the values of other models. _________

79) Close the Results window.

80) Close SAS Enterprise Miner.