William Eddy 9/21/19
MIS 6903 – Business Intelligence and Data Analytics Homework 1 – Decision Tree

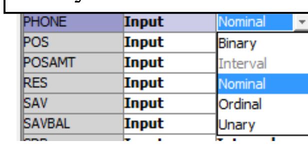
Explanation of Model Creation Process

For data importing, I used R to convert the dataset to CSV, and then imported it into SAS Enterprise Miner (EM). Since this method does remove some of the information included in the SAS data file, I needed to relabel the variables in the correct way. All "Binary" variables below were labelled manually.

Name	Role	Level
ACCTAGE	Input	Nominal
AGE	Input	Nominal
ATM	Input	Binary
ATMAMT	Input	Interval
BRANCH	Input	Nominal
CASHBK	Input	Interval
CC	Input	Binary
CCBAL	Input	Nominal
CCPURC	Input	Nominal
CD	Input	Binary
CDBAL	Input	Interval
CHECKS	Input	Interval
CRSCORE	Input	Nominal
DDA	Input	Binary
DDABAL	Input	Interval
DE	Rejected	Interval
DEP	Input	Interval
DEPAMT	Input	Interval
DIRDEP	Input	Binary
HMOWN	Input	Binary
HMVAL	Input	Nominal
IDNUM	ID	Nominal
ILS	Input	Binary
ILSBAL	Input	Interval

INAREA	Input	Binary
INCOME	Input	Nominal
INS	Target	Binary
INV	Input	Binary
INVBAL	Input	Nominal
IRA	Input	Binary
IRABAL	Input	Interval
LOC	Input	Binary
LOCBAL	Input	Interval
LORES	Input	Nominal
MM	Input	Binary
MMBAL	Input	Interval
MMCRED	Input	Interval
MOVED	Input	Binary
MTG	Input	Binary
MTGBAL	Input	Interval
NSF	Input	Binary
NSFAMT	Input	Interval
PHONE	Input	Nominal
POS	Input	Nominal
POSAMT	Input	Nominal
RES	Input	Nominal
SAV	Input	Binary
SAVBAL	Input	Interval
SDB	Input	Binary
TELLER	Input	Interval
VAR1	Input	Nominal

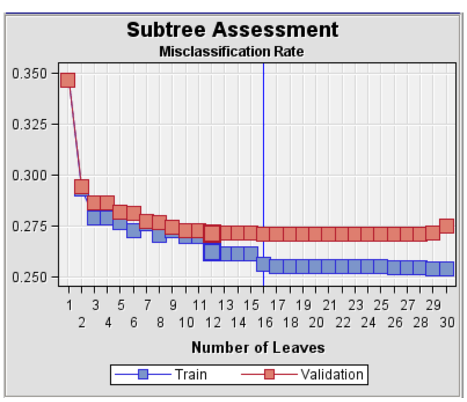
The instructions requested that we change the labels for CCPURC and PHONE to Interval. However, as shown below, EM labeled these as Nominal, and Interval cannot be selected. After relabeling variables to the best of my ability, I commenced with the analysis as normal.

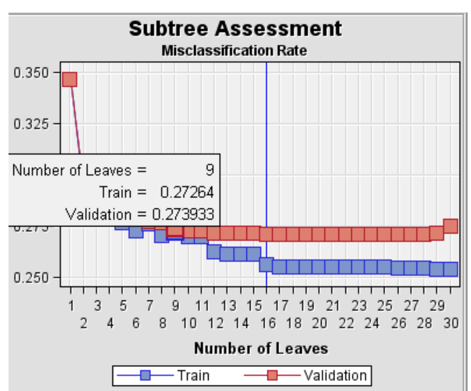


After labelling the variables, I used the data partition node to split the dataset into 70% training data and 30% test data. EM creates the model using the training data, and evaluates its accuracy using the test data.

I set out to use a decision tree model to identify the best variables for predicting whether an individual has insurance. EM can add variables to the decision tree model in order of variable predictive ability. The more accurate the model is, the lower the misclassification rate will be when the model is applied to the test data. I defined "Strongest indicators" as any variables remaining after trimming those which do not substantially reduce the misclassification rate when the model is applied to the test dataset.

Each box on the subtree assessment plot below shows a step in the decision tree where a variable is used to distinguish whether an individual has insurance. For a simplified example, a variable might indicate "yes" if its value is above 100, and "No" if its value is below 100. These steps are called nodes.





The subtree assessment plot for the test data showed the lowest misclassification rate with a minimum of 16 nodes in the model. With 16 nodes, the misclassification rate is .271 – rounding to two decimal places at .27. However, this is too complex. I included 9 nodes, which showed a misclassification rate of .273, also rounding to .27. I chose not to reduce the size of the decision tree further, because with only 8 nodes the misclassification rate is .278, which rounds to .28 – not quite as impressive. The variables used in these 9 remaining nodes are as follows, the most powerful variables for predicting group membership in the training set:

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
SAVBAL		2	1.0000	1.0000	1.0000
MM		1	0.6770	0.6266	0.9255
DDABAL		3	0.5336	0.4676	0.8764
DDA		2	0.5023	0.4331	0.8623
CD		3	0.3430	0.2837	0.8271
BRANCH		1	0.1765	0.0000	0.0000
CHECKS		1	0.1762	0.0000	0.0000
ATMAMT		1	0.1209	0.0000	0.0000
CCBAL		1	0.1000	0.1300	1.2998

The variable importance statistics included in the output above are based on the reduction in sum of squared error a variable contributes. This information is analogous that indicated in the subtree assessment plot. Notice that the importance decreases drastically for each consecutive variable. We want any variable we include to have high variable importance.

Additionally, we want high variable importance in not only the training dataset, but also the test dataset. 3 variables listed above – BRANCH, CHECKS, and ATMAMT, were important in the training data, but not useful for predicting group membership in the test data. This may indicate overfitting of the model. Because of this, I removed these variables from the final decision tree.

The remaining 6 variables are highly predictive of group membership and adding any more does not improve our model's predictive ability substantially. These include SAVBAL, MM, DDABAL, DDA, CD, and CCBAL. Furthermore, as explained in the executive summary below, we are targeting individuals with a high likelihood of not having insurance. Based on the percentage classified as not having insurance displayed in the decision tree model, the most useful variables can be further reduced to just SAVBAL, MM, and DDABAL.

Classification table for the validation dataset:

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	76.4706	84.5668	3211	55.2858
1	0	23.5294	49.1298	988	17.0110
0	1	36.4201	15.4332	586	10.0895
1	1	63.5799	50.8702	1023	17.6136

The classification table above shows that our model produces a total 72.9% accuracy rate. Our model is better at predicting that someone does NOT have insurance, predicting correctly 85% of cases where individuals in the validation dataset do not have insurance – our true negative rate. Conversely, the model correctly predicted 51% of cases where individuals do have insurance – our true positive rate.

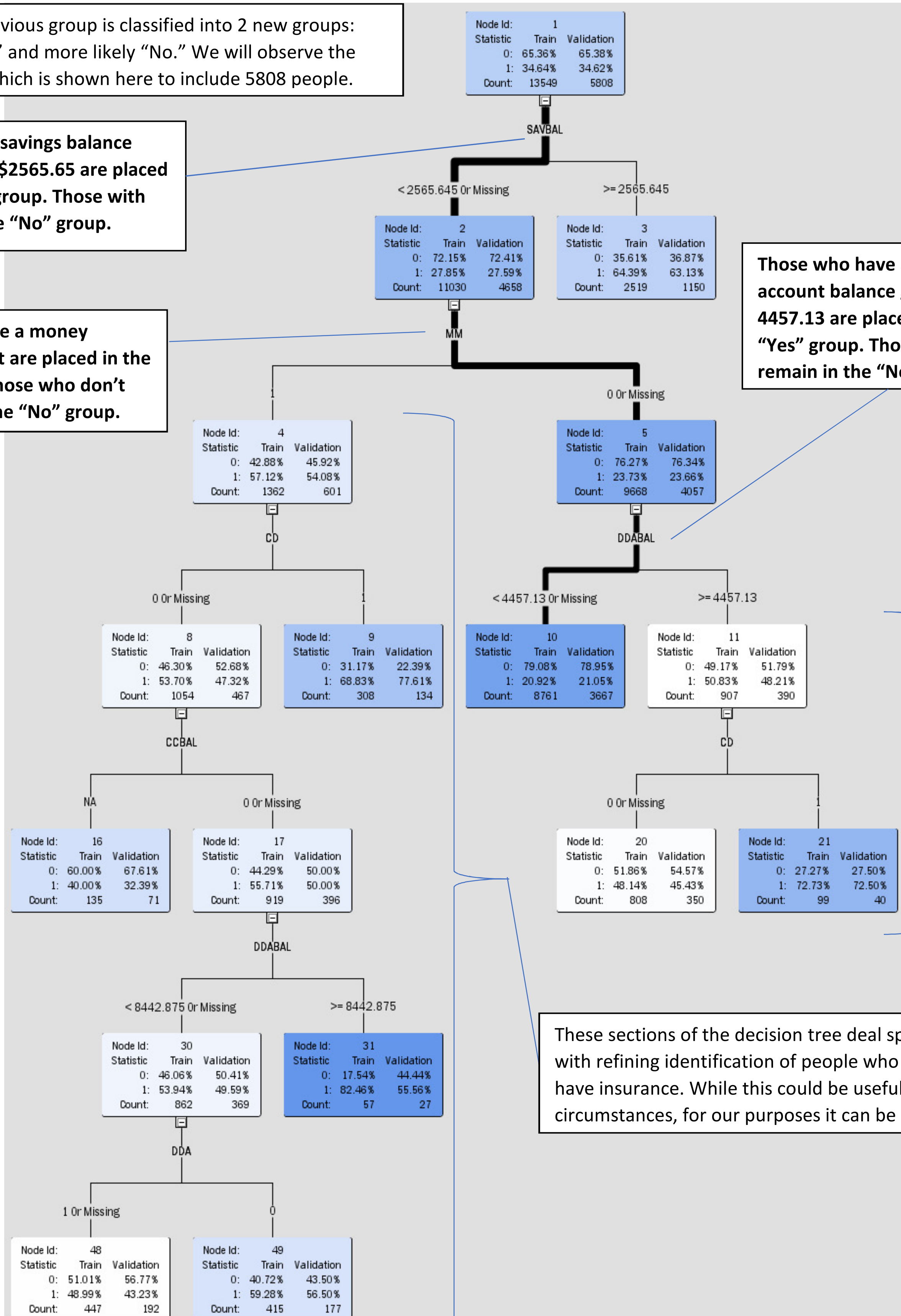
Because we are attempting to sell insurance, this is a useful model. We can identify large numbers of people who do not yet have insurance and target them in our sales efforts.

At each split, previous group is classified into 2 new groups: More likely “Yes” and more likely “No.” We will observe the Validation set, which is shown here to include 5808 people.

Those with a savings balance greater than $2565.65 are placed in the “Yes” group. Those with less are in the “No” group.

Those who have a money market account are placed in the “Yes” group. Those who don’t are placed in the “No” group.

Those who have a DDA account balance greater than 4457.13 are placed in the “Yes” group. Those with less remain in the “No” group.



These sections of the decision tree deal specifically with refining identification of people who already have insurance. While this could be useful in other circumstances, for our purposes it can be ignored.

Executive Summary and Recommendations

I created a decision-tree predictive model with the goal of finding people who do not yet have insurance.

When our model predicts that someone does not have insurance, it is correct 85% of the time. Because we are attempting to sell insurance, this is a useful model. We can identify large numbers of people who do not yet have insurance and target them in our sales efforts.

The best variables for predicting that someone does not currently have insurance are:

1. SAVBAL - Savings account balance, indicating amount in the account in US Dollars
2. MM - Money Market, indicating whether an individual has a money market account
3. DDABAL – Demand Deposit Account balance, indicating amount in the account in US Dollars

Based on the decision tree model, see the following profile of the ideal customer to target:

1. Target people who have a savings balance of less than $2565.65.
2. Target people who do not have a money market account.
3. Target people who have a DDA account balance less than 4457.13.

This model will be most accurate when used to target individuals who have all 3 of these aspects. However, a wider marketing campaign might include individuals who possess any one of these characteristics.

Using this model, we can contact customers who are highly unlikely to already be enrolled in a current insurance plan, maximizing our success rate per contact, and creating the highest return on investment for the marketing campaign.