Decision Trees with Claim Fraud Dataset/ SAS Enterprise Miner

Didem B. Aykurt

Colorado State University Global

MIS530; Predictive Analytics

Dr.Jennifer Catalano

February 19, 2023

Decision Trees

Decision trees are the most popular predictive and descriptive-analytic; they are easy to create and understand why they are most helpful. A decision tree considers at least one categorical or continuous target variable. The model uses algorithms to split decisively by variables that create branches like a tree structure. The decision tree method makes an if-then-else statement to split the data into smaller segments called nodes. If the node doesn't succeed in breaking, it refers to the leaf. The root node includes all the data. A decision tree can be a significant step in beginning predictive analytics to understand input variables on the target variable, mainly used for market and customer segmentation like mortgage or loan decisions by credit rating. The model can handle missing values evaluated by statistically significant test results like Chi-square or F-test. The strange or considerable impact is whether the input and target values have a strong relationship and whether they should be combined. Deville and Neville report the resulting guideline for the connection.

Confidence	Strength of the relationship
0.001	Extremely good
0.01	Good
0.05	Pretty good
0.10	Not so good
0.15	Extremely weak

Note: From Richard V. McCarthy, Mary M. McCarthy; Wendy Ceccucci, 2022. *Applying Predictive Analytics Finding Value in Data*. Second edition.

Decision trees have two different types of models. One is classification tree models; if the target variable is categorical, the model mainly uses clustering algorithms to split data as Gini impurity and Chi-square. On the other hand, regression tree models have interval target variables and use an F-distribution and average square error to break in the leaves.

Creating a Decision Tree Using SAS Enterprise Miner

I will apply two decision trees to the claim fraud data and compare which decision tree result provides the best predictive result. The first one is the SEMMA model in the Decision Tree node in the Model tab.

Figure 1: Decision Tree node connection in SAS Enterprise Miner.

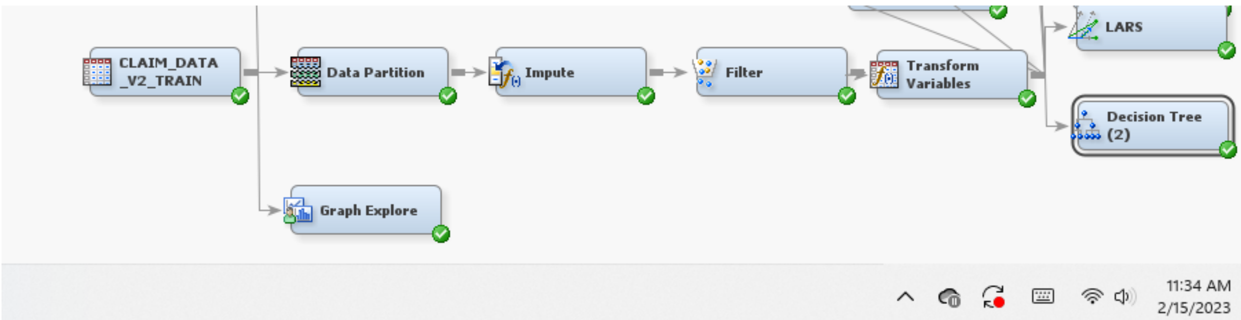


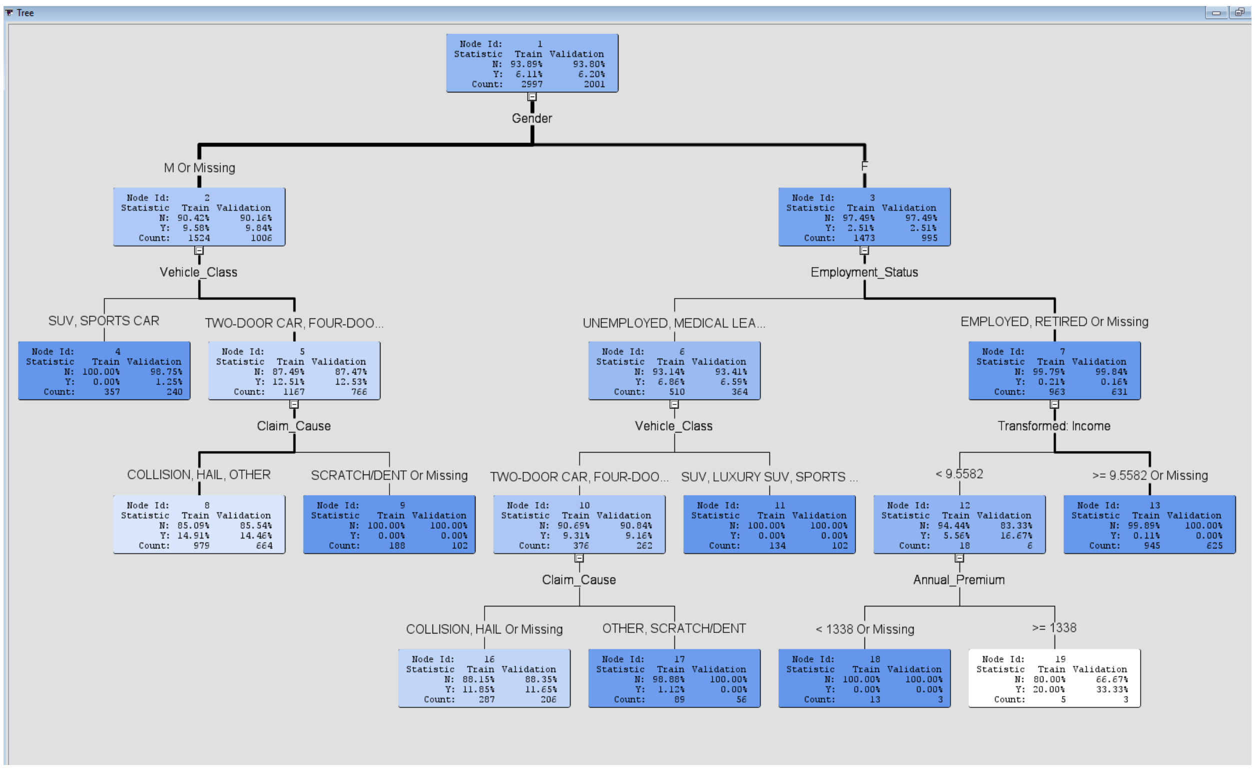
Figure 2: Decision tree properties.

General	
Node ID	Tree
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	Gini
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importa	
Observation Based Importa	No
Number Single Var Importar	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustm	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Interactive Sample	
Create Sample	Default
Sample Method	Random
Sample Size	10000
Sample Seed	12345
Performance	Disk
Score	
Variable Selection	Yes
Leaf Role	Segment
Report	
Precision	4
Tree Precision	4
Class Target Node Color	Percent Correctly Classified
Interval Target Node Color	Average
Node Text	...
Status	
Create Time	2/15/23 6:20 PM

The result of the Decision Tree node Tree window shows the decision tree itself with color differences in the nodes. The darker color has a more decisive influence, and the white node specifies a weaker effect. All nodes have a probability of each outcome for both the test and validation datasets.

Let’s look at the claim fraus dataset result; the female input probability of fraudulent female customers is 2.51% in the test and validation datasets. The darker the line specifies the volume of observations that passed the path, the thicker, the higher the number of observations.

Figure 3: Decision Tree diagram results



The Variable Importance window shows a list of input variables used in the decision tree and the number of splits obtained within those variables. The importance of statistics for the training dataset shows how the input variables fit the ree. The decision tree Variable Importance result shows that the six input variables are vehicle class, claim cause, gender, employment status, transformed income, and annual premium. The vehicle class affects the entire tree, and gender is the second variable that affects the tree. Validation importance is the observation of each variable for the validation dataset. The Ratio of Validation to Training Importance shows the ratio between validation dataset importance statistics and training dataset importance statistics as a minor result that input was used in overly optimistic splitting rules.

Figure 4: Variable Importance window.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Vehicle_Class	Vehicle_Class	2	1.0000	1.0000	1.0000
Claim_Cause	Claim_Cause	2	0.9139	0.9123	0.9982
Gender	Gender	1	0.8537	0.9593	1.1237
Employment_Status	Employment_Status	1	0.5363	0.5719	1.0665
Annual_Premium	Annual_Premium	1	0.1677	0.2536	1.5121
LOG_Income	Transformed: Income	1	0.1011	0.1762	1.7432
Claim_Report_Type	Claim_Report_Type	0	0.0000	0.0000	.
Claim_Amount	Claim_Amount	0	0.0000	0.0000	.
IMP_Outstanding_Balan...	Imputed: Outstanding_B...	0	0.0000	0.0000	.
Monthly_Premium	Monthly_Premium	0	0.0000	0.0000	.
State_Code	State_Code	0	0.0000	0.0000	.
Months_Since_Policy_In...	Months_Since_Policy_In...	0	0.0000	0.0000	.
IMP_Education	Imputed: Education	0	0.0000	0.0000	.
IMP_Location	Imputed: Location	0	0.0000	0.0000	.
Marital_Status	Marital_Status	0	0.0000	0.0000	.
Vehicle_Model	Vehicle_Model	0	0.0000	0.0000	.
Months_Since_Last_Cla...	Months_Since_Last_Cl...	0	0.0000	0.0000	.
Claim_Date	Claim_Date	0	0.0000	0.0000	.
Vehicle_Size	Vehicle_Size	0	0.0000	0.0000	.

11:34 AM
2/15/2023

The average square error at 0.053507 statistically result helps to compare the predictive model. As I worked on Chapter 4 and showed a PLS regression result of 0.054662, the decision tree model was slightly better than the predictive model I applied in Chapter 4 because of the lower average square error.

Figure 5: Decision tree result- Fit Statistics

Fit Statistics

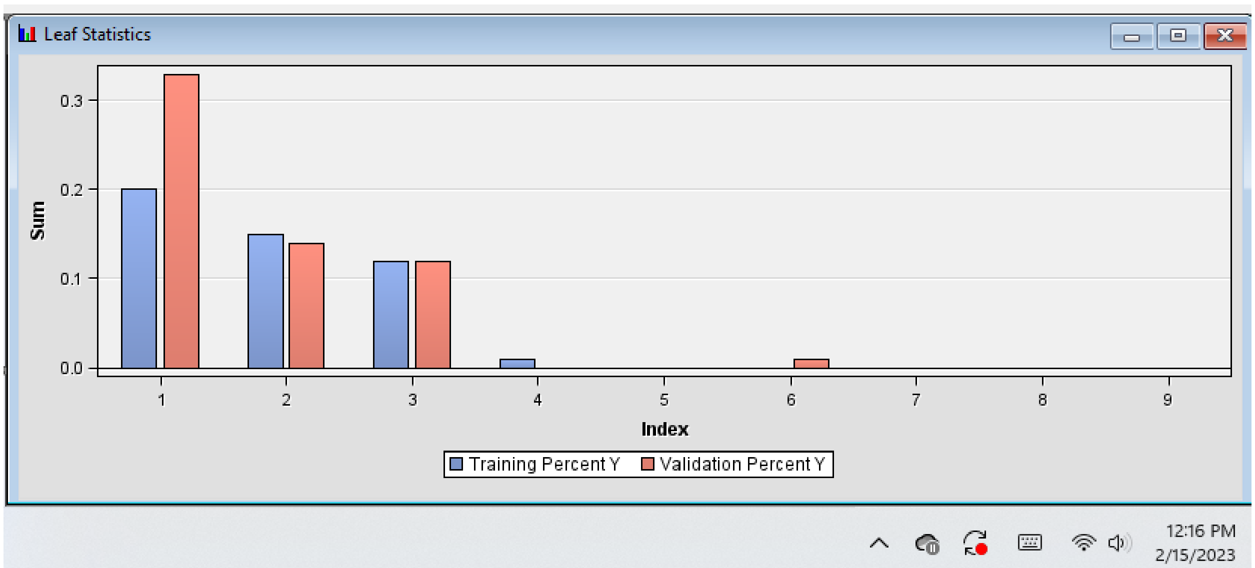
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudulent_Claim	Fraudulent_Claim	_NOBS_	Sum of Frequencies	2997	2001	.
Fraudulent_Claim	Fraudulent_Claim	_MISC_	Misclassification R...	0.061061	0.061969	.
Fraudulent_Claim	Fraudulent_Claim	_MAX_	Maximum Absolute...	0.998942	1	.
Fraudulent_Claim	Fraudulent_Claim	_SSE_	Sum of Squared Er...	313.9732	214.1334	.
Fraudulent_Claim	Fraudulent_Claim	_ASE_	Average Squared E...	0.052381	0.053507	.
Fraudulent_Claim	Fraudulent_Claim	_RASE_	Root Average Squa...	0.228869	0.231315	.
Fraudulent_Claim	Fraudulent_Claim	_DIV_	Divisor for ASE	5994	4002	.
Fraudulent_Claim	Fraudulent_Claim	_DFT_	Total Degrees of Fr...	2997	.	.

Tree

12:04 PM
2/15/2023

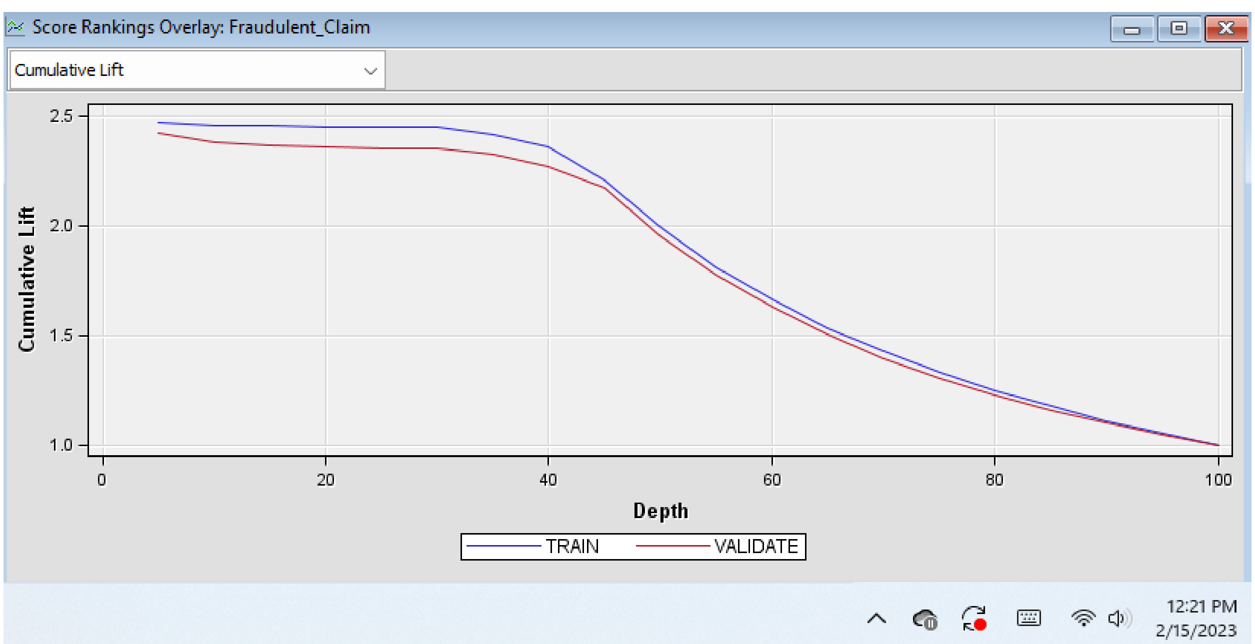
The Leaf Statistics window shows the number of leaves within the tree in case six in the claim fraud decision tree.

Figure 6: Leaf Statistics histogram.



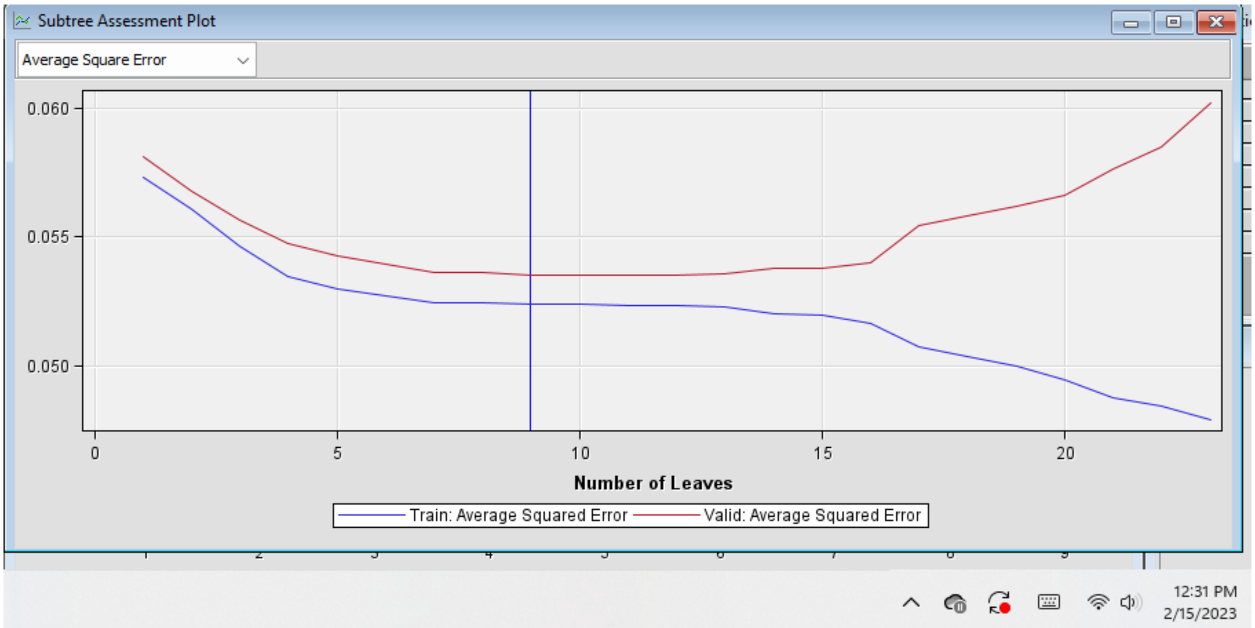
In the first step of 25% of data, the cumulative lift is over 2.4, which provides a signal of the strangeness of this decision tree.

Figure 7: Cumulative lift chart.



The Subtree Assessment plot helps to know if the model has been overfitting. If it fits the data, it becomes less comprehensive.

Figure 8: Subtree Assessment Plot.

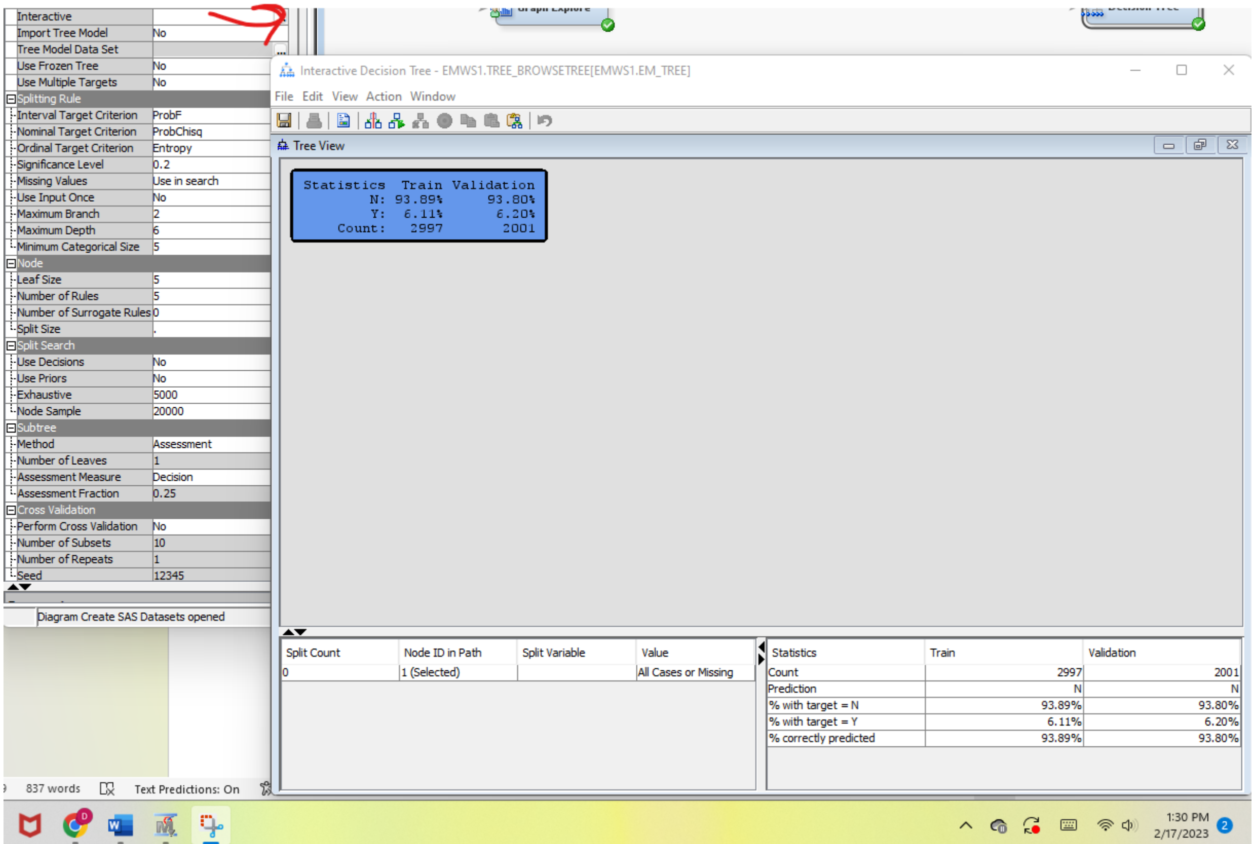


Creating an Interactive Decision Tree

The decision tree node has an excellent property to create a decision tree that controls the variables and values of the variables split which information is needed. One of the most remarkable abilities is to show input variables and how input variables divide to create an optimal tree. Some input variables don't need to be split; however, a decision tree can break unnecessary variables. For example, the input variable age splits on values like 23.7, 38.9, and 51.2; these values might be accurate and can create an interactive decision tree. Additionally, the user can't add variables for analysis. Select the interactive tree property; a new window will show only the root node.

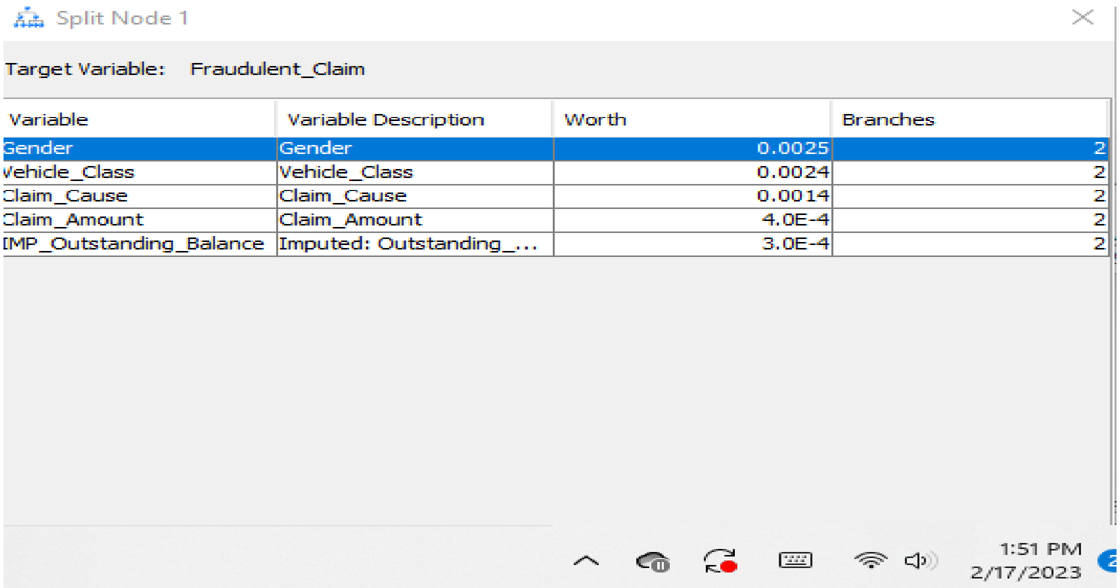
The window displays three panels: one shows the leaf for the root node, and the bottom left shows the split count. The bottom right shows the statistics pane, which provides the observation number and percentage of values, both the train and validation dataset, and predicted results, such as the percentage of target variable N of 93.89% for train data. Right-click on the root node enables a split to occur on the root node.

Figure 9: Property of Interaction Decision Tree Output.



Select Split Node; a Split Node 1 window will appear, showing all the variables that could be split as presented in the default selection. Next, select IMP_Outstanding_Balance and click Edit Rule for the specific split I need.

Figure 10: Split node Variable Selection



I split four levels of the outstanding balance; type in the New split point box and click Add Branch, then Apply, OK.

Figure 11: Edit Rule section

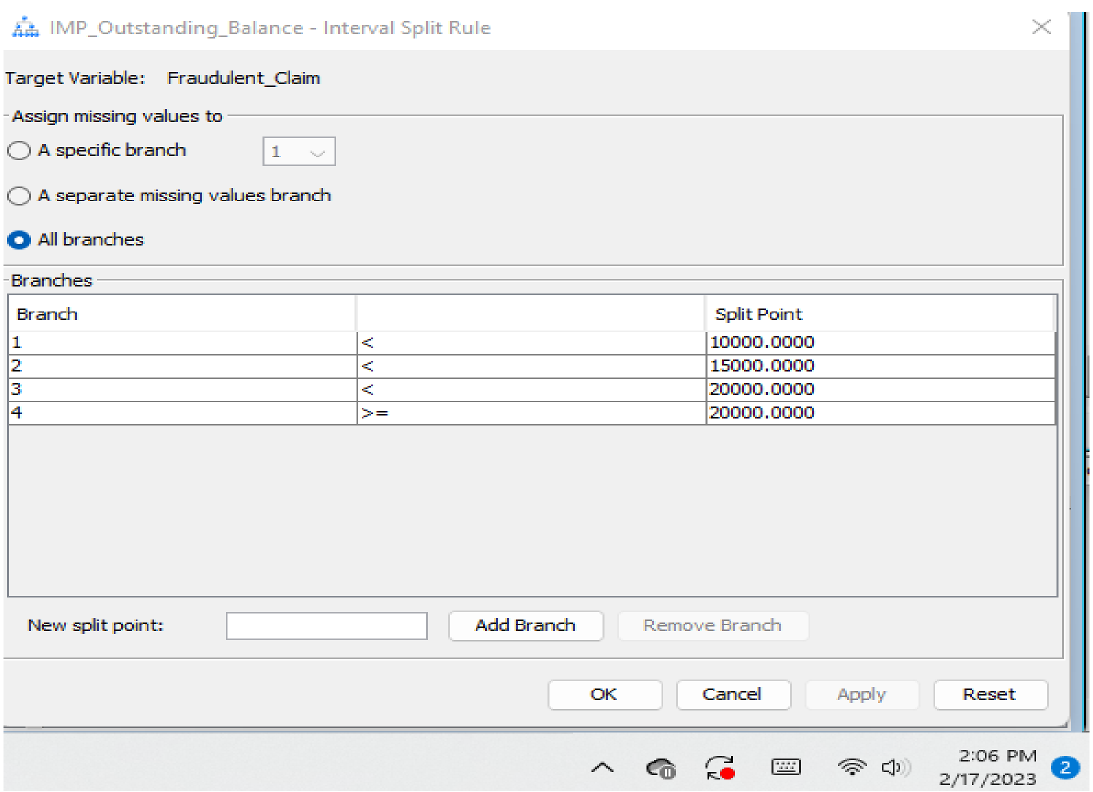
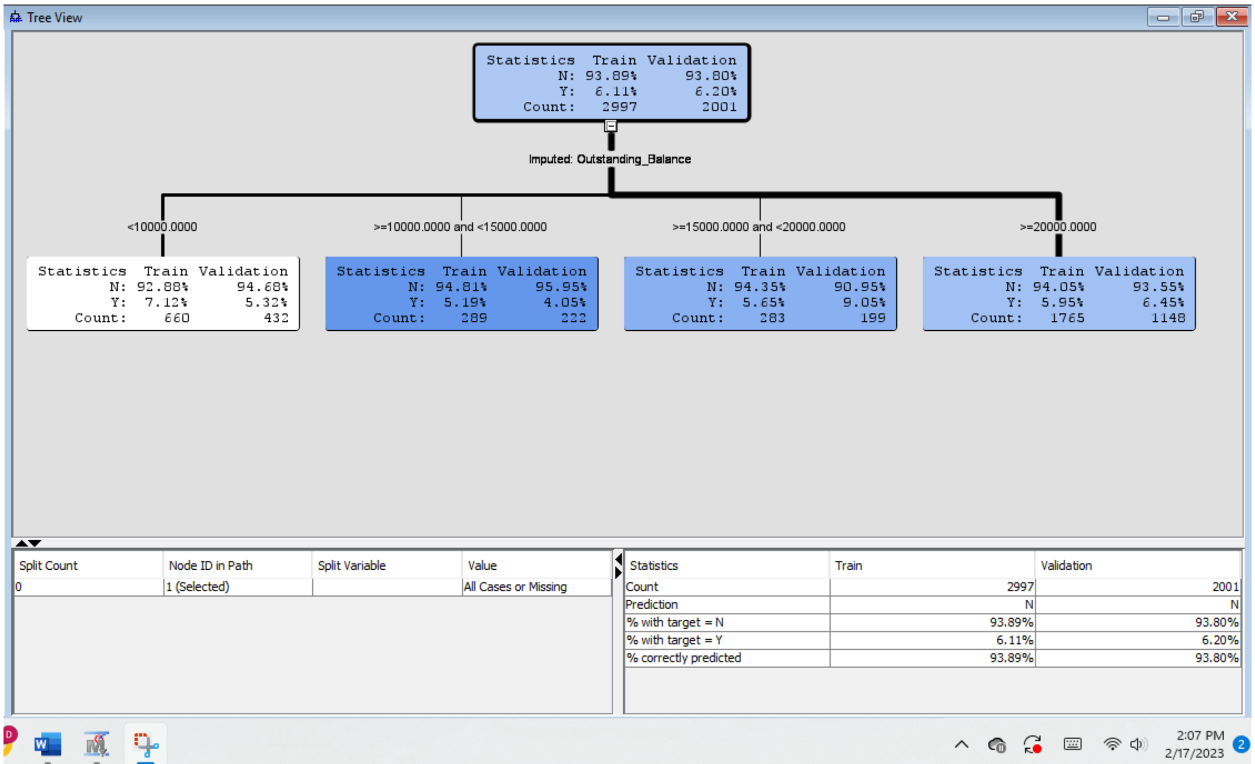


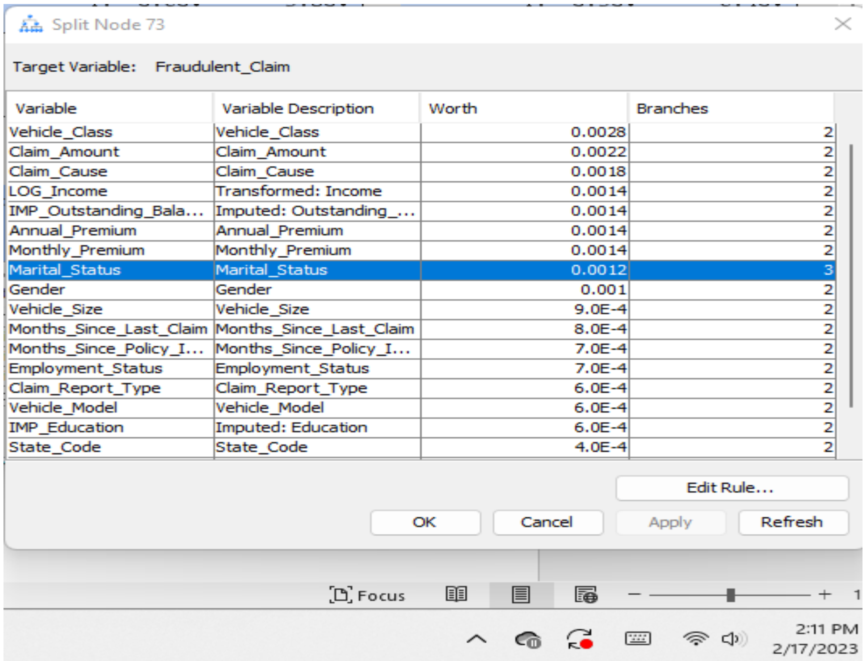
Figure 12, the Tree View window shows four nodes that more logically separate the data for the extraordinary decision between specific values.

Figure 12: The First node is Outstanding Balance split four nodes- Interactive decision tree



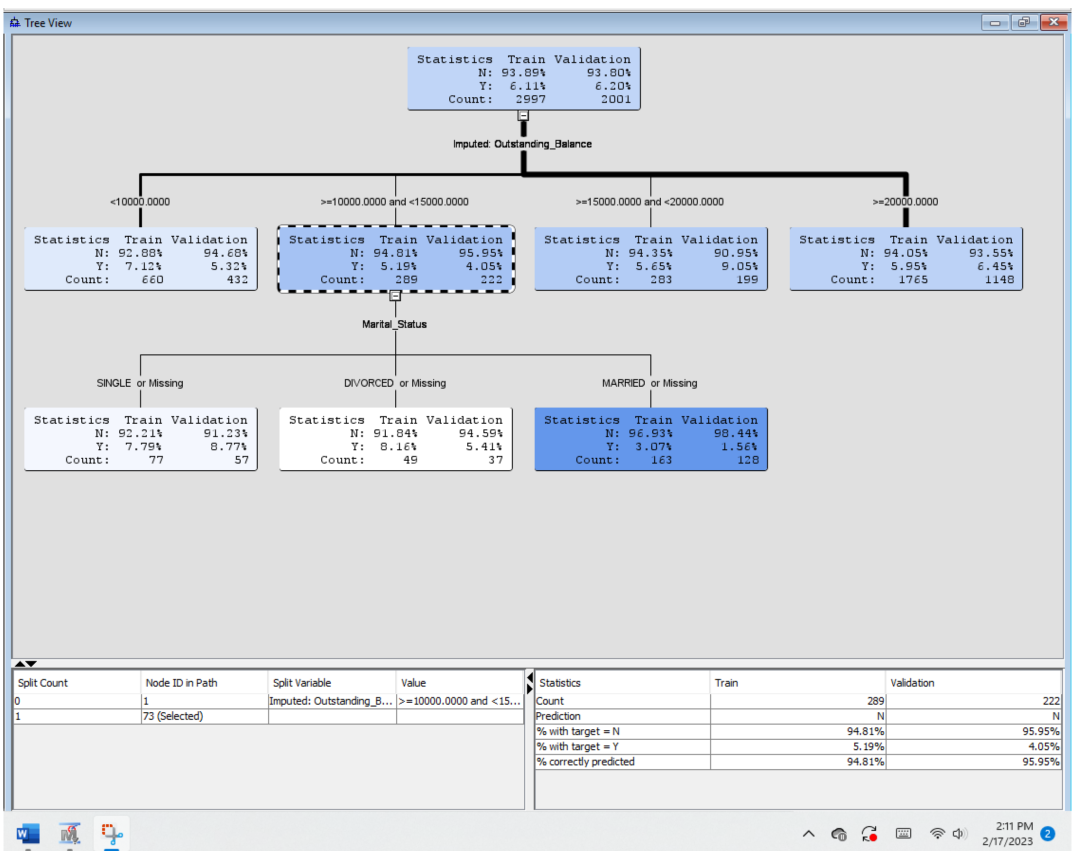
We can add a new variable to grow an interaction decision tree. I added a second split of the Marital Status variable into an Outstanding Balance between 10,000 and 15,000. Right-click on the second node, select Split Node, and choose Marital_Status and Edit Rule.

Figure 13: The second node is the Marital status split into three nodes selection and edit.



The interactive decision tree, the Tree View, shows the split of variables specifically of interest.

Figure 14: An interactive decision tree- second node split.



The result of the interactive decision tree Fit Statistic table has information on the average square error of 0.0581. Compare the development of the middle square error system-generated tree and interactive decision tree; the system-generated tree is slightly better than the interactive decision tree.

Figure 15: Fit Statistics of the Interactive Decision Tree

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Fraudulent_Claim	Fraudulent_Claim	_NOBS_	Sum of Frequencies	2997	2001
Fraudulent_Claim	Fraudulent_Claim	_MISC_	Misclassification Rate	0.061061	0.061969
Fraudulent_Claim	Fraudulent_Claim	_MAX_	Maximum Absolute Error	0.969325	0.969325
Fraudulent_Claim	Fraudulent_Claim	_SSE_	Sum of Squared Errors	343.1091	232.5174
Fraudulent_Claim	Fraudulent_Claim	_ASE_	Average Squared Error	0.057242	0.0581
Fraudulent_Claim	Fraudulent_Claim	_RASE_	Root Average Squared Error	0.239253	0.24104
Fraudulent_Claim	Fraudulent_Claim	_DIV_	Divisor for ASE	5994	4002
Fraudulent_Claim	Fraudulent_Claim	_DFT_	Total Degrees of Freedom	2997	.

The result of cumulative lift for the interaction decision tree is that 15% of data has a charge greater than 1. Thus, the system generated a decision tree to produce a better fit; one thing to do better is to choose and control specific variables and values with an interactive decision tree.

Figure 16: Interactive decision tree Cumulative lift

Creating a Maximal Decision Tree Using

The previous topic is that the interactive decision tree helps to focus each variable that will be effectively used to select the criteria for splitting nodes; another alternative way is creating the maximal tree. The maximal tree creates a large tree structure and uses a starting point, although the decision tree can crop a tree for using predictive modeling.

Maximal tree in SAS Enterprise Miner, click Interactive ellipse in interactive tree property from the Decision tree node, then right-click on the root node and select Train Node. The tree will grow further than the close interactive decision tree page, run the system decision tree, and display the result. I used the claim fraud dataset and the same variables but different splits.

Figure 17: Maximal tree M value of gender node split side.

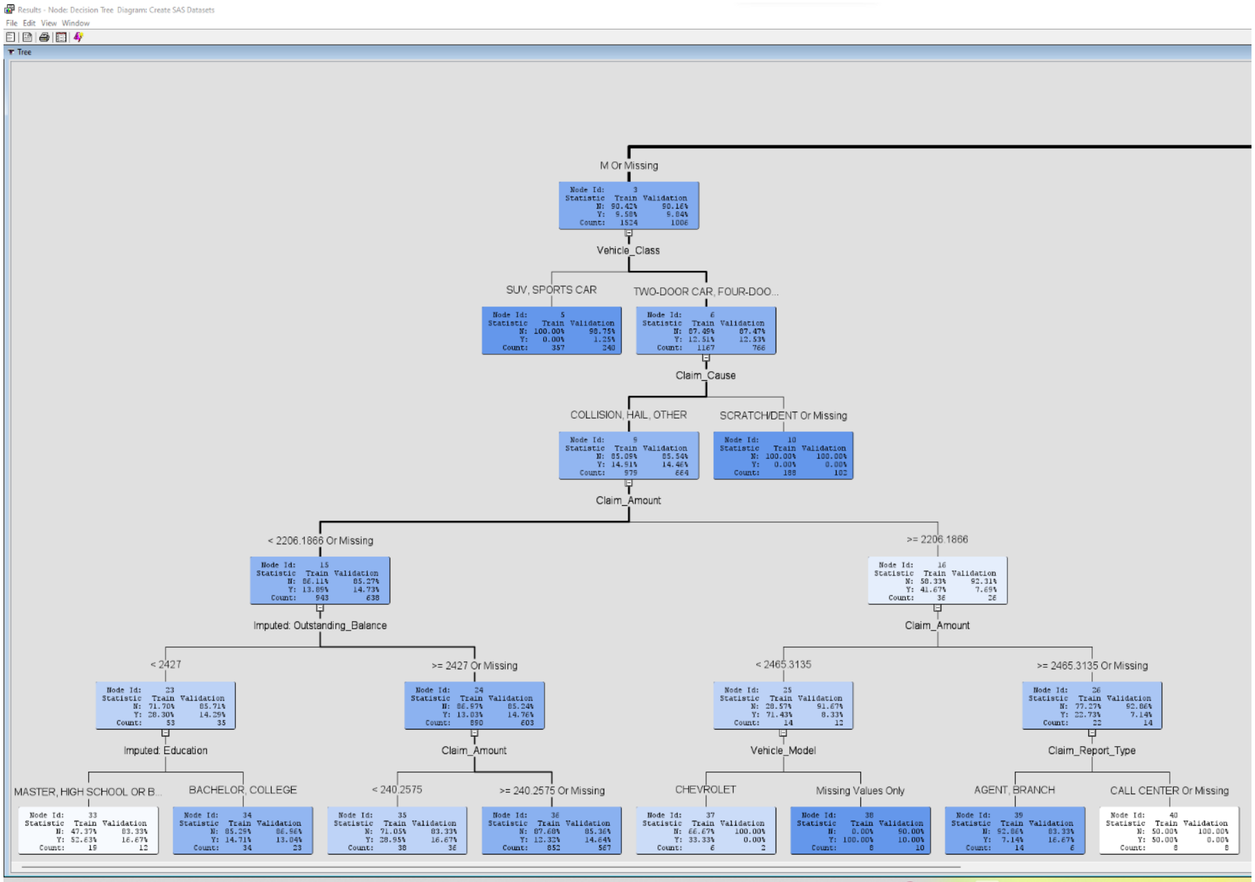
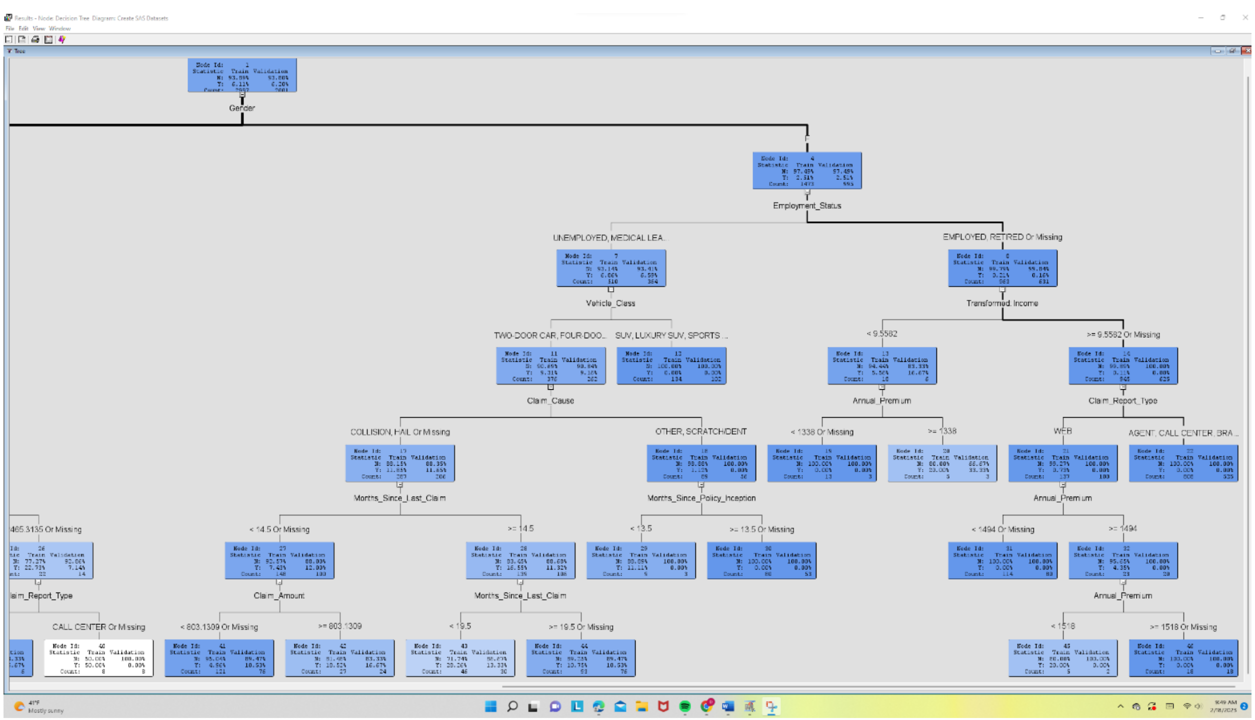
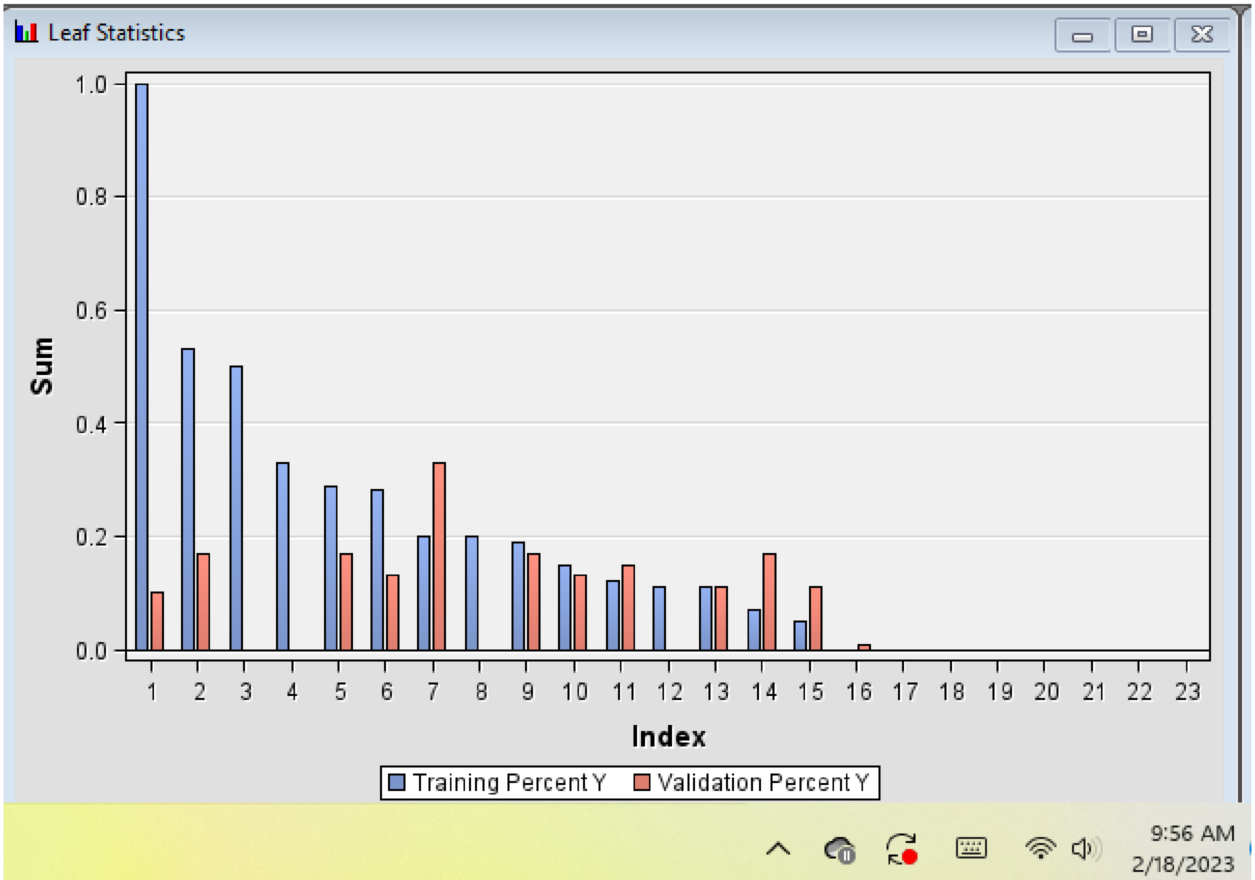


Figure 18: Maximal tree F value of Gender node split side.



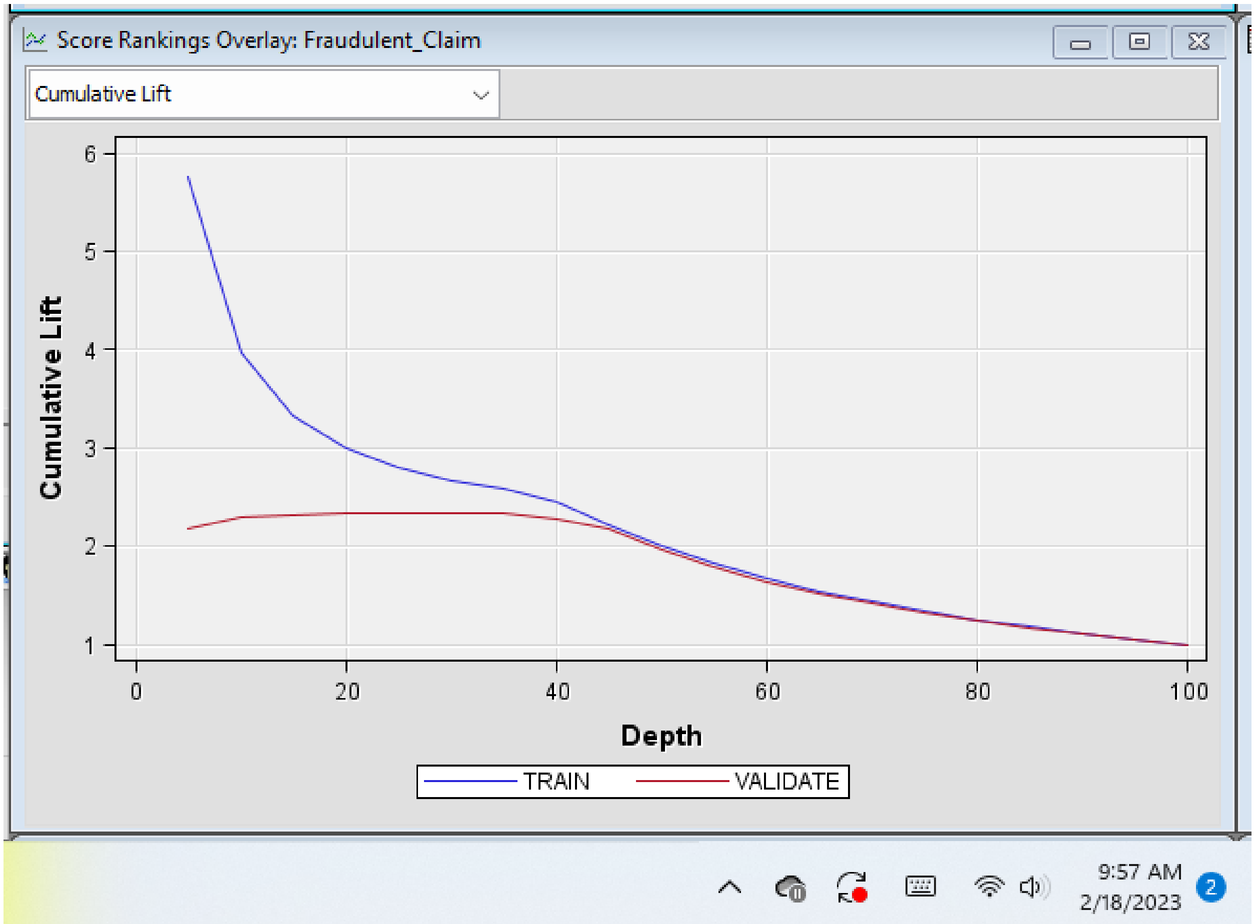
The result of the maximal tree Leaf Statistic window shows 11 branches and 27 leaves.

Figure 19: Maximal tree-leaf Statistic



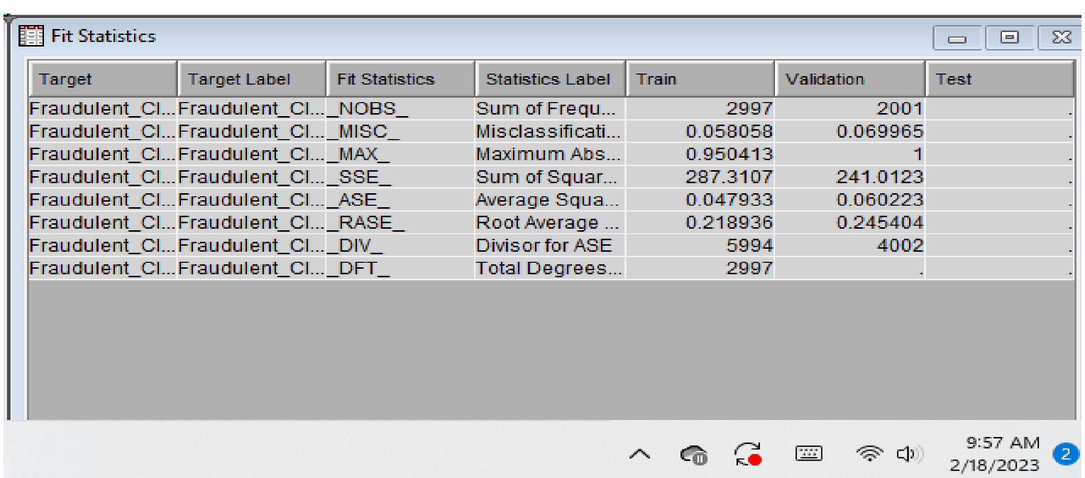
The maximal tree Cumulative Lift results show that 25% of cases have 2.3 cumulative lifts. The result shows the incremental lift of the maximal tree lower than the system-generate decision tree, although it causes continued charge over a large percentage of cases.

Figure 20: Maximal tree Cumulative Lift window.



The fit statistic of the maximal tree shows an average square error of 0.0602223 and a high difference between the train and validation average square error.

Figure 21: Maximal tree Fit Statistics window.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Fraudulent_Cl...	Fraudulent_Cl...	_NOBS_	Sum of Frequ...	2997	2001	.
Fraudulent_Cl...	Fraudulent_Cl...	_MISC_	Misclassificati...	0.058058	0.069965	.
Fraudulent_Cl...	Fraudulent_Cl...	_MAX_	Maximum Abs...	0.950413	1	.
Fraudulent_Cl...	Fraudulent_Cl...	_SSE_	Sum of Squar...	287.3107	241.0123	.
Fraudulent_Cl...	Fraudulent_Cl...	_ASE_	Average Squa...	0.047933	0.060223	.
Fraudulent_Cl...	Fraudulent_Cl...	_RASE_	Root Average ...	0.218936	0.245404	.
Fraudulent_Cl...	Fraudulent_Cl...	_DIV_	Divisor for ASE	5994	4002	.
Fraudulent_Cl...	Fraudulent_Cl...	_DFT_	Total Degrees...	2997	.	.

9:57 AM
2/18/2023

Conclusion

"Decision trees are useful when you have data sets with many input variables, especially when there are nominal variables. They can be useful for segmenting the insignificant nominal variables and ranges that can be easily pruned." (McCarthy,2022)

The three trees are easy to create, and each tree differs from other properties used; however, all the results did not help. Therefore, the optional tree should be used when the analyst needs a specific value.

Model Name	Average square error	Depth	Cumulative Lift
System-generated Decision Tree	0.053507	25%	2.4
Interactive Decision tree	0.0581	15%	Over 1
Maximal Tree	0.0602223	25%	2.3

Thus, the result of the system-generated decision tree's average square error is lower than others and lower than the previous chapter's PLS regression result. Therefore, the system-generated decision tree is the best-fit model for the claim fraud dataset.

Reference

Richard V. McCarthy, Mary M. McCarthy; Wendy Ceccucci, 2022. *Applying Predictive Analytics Finding Value in Data*. Second edition.