**Week 6 – Homework Exercise 1**

In Exercise 1, you work with a data set named **PVA**. It contains data that represents charitable donations made to an American veterans’ association. The data represent the results of a mail campaign to solicit donations. The data set contains the following information:

* a flag to indicate respondents to the appeal and the dollar amount of their donations (**Target Gift Flag** and **Target Gift Amount**)
* respondents’ PVA promotion and giving history
* demographic data of the respondents

#### PVA Metadata Table

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** |  | **Measurement Level** | **Description** |
| **DemAge** |  | Interval | Age |
| **DemCluster** |  | Nominal | Demographic Cluster |
| **DemGender** |  | Nominal | Gender |
| **DemHomeOwner** |  | Binary | Home Owner |
| **DemMedHomeValue** |  | Interval | Median Home Value Region |
| **DemMedIncome** |  | Interval | Median Income Region |
| **DemPctVeterans** |  | Interval | Percent Veterans Region |
| **GiftAvg36** |  | Interval | Gift Amount Average 36 Months |
| **GiftAvgAll** |  | Interval | Gift Amount Average All Months |
| **GiftAvgCard36** |  | Interval | Gift Amount Average Card 36 Months |
| **GiftAvgLast** |  | Interval | Gift Amount Last |
| **GiftCnt36** |  | Interval | Gift Count 36 Months |
| **GiftCntAll** |  | Interval | Gift Count All Months |
| **GiftCntCard36** |  | Interval | Gift Count Card 36 Months |
| **GiftCntCardAll** |  | Interval | Gift Count Card All Months |
| **GiftTimeFirst** |  | Interval | Time Since First Gift |
| **GiftTimeLast** |  | Interval | Time Since Last Gift |
| **ID** |  | Nominal | Control Number |
| **PromCnt12** |  | Interval | Promotion Count 12 Months |
| **PromCnt36** |  | Interval | Promotion Count 36 Months |
| **PromCntAll** |  | Interval | Promotion Count All Months |
| **PromCntCard12** |  | Interval | Promotion Count Card 12 Months |
| **PromCntCard36** |  | Interval | Promotion Count Card 36 Months |
| **PromCntCardAll** |  | Interval | Promotion Count Card All Months |
| **StatusCat96NK** |  | Nominal | Status Category 96NK |
| **StatusCatStarAll** |  | Binary | Status Category Star All Months |
| **TARGET\_B** |  | Binary | Target Gift Flag |
| **TARGET\_D** |  | Interval | Target Gift Amount |

In this exercise, you continue to use the **PVA** data set to build a decision tree to classify those customers who donated. Please use screenshots if you believe it adds value to any of your responses.

1. **Building a Decision Tree in SAS Visual Statistics**
   1. After logging into SAS Viya for Learners, select Explore & Visualize from the software application menu. Then select and open the **PVA** data source.
   2. Add a decision tree to the canvas.
   3. If you did not already do so, in the Measure column, edit **Target Gift Flag** and select **Category** to create a binary target variable for donations.
   4. Add **Target Gift Flag** as the response.
   5. Under Predictors, click **Add**. In the Add Data Items window, select all 28 predictor variables ***except*** for these four:

* **Control Number**
* **Demographic Cluster**
* **Target Gift Amount**
* **Target Gift Amount with Zero**

(You add 24 columns.)

* 1. Review the decision tree.
* How many customers made donations?

53,273 of customer donated.

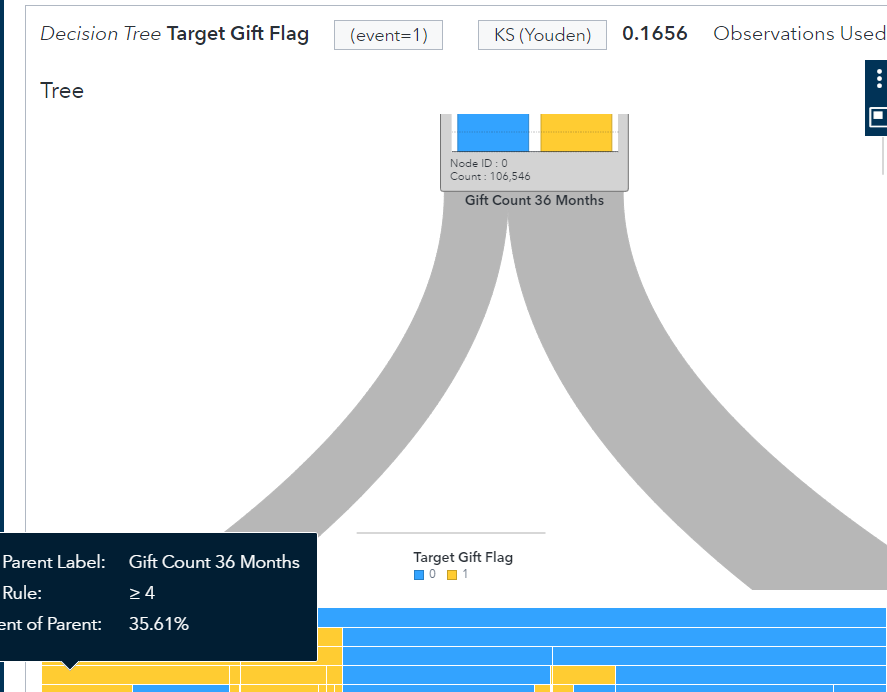
* What proportion of individuals does that represent in the target node?

1/2 or 50percent

* 1. Zoom in toward the root node.
* On what predictor does the top split occur?

Gift Count 36 month

* What is the split point that determines to which branch a customer belongs?

Split point is 4 

* In which branch does the majority of customers fall at this split point?

The Majority of customers fall in no donation

* How many customers were less than this value and belong to Node 2?

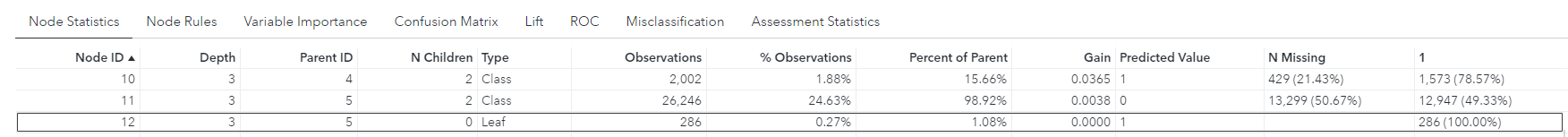
68,607 customer

* What proportion of the customers in Node 2 made donations?

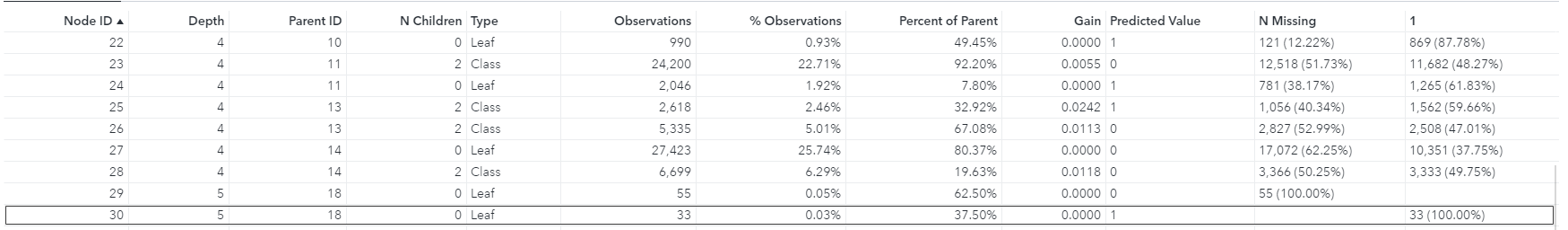
45.17%

1. **Examining Additional Decision Tree Results**
   1. Open the summary table to examine the node statistics.
      1. Examine the last column to see whether there are any 100% donor nodes. If so, which node or nodes?

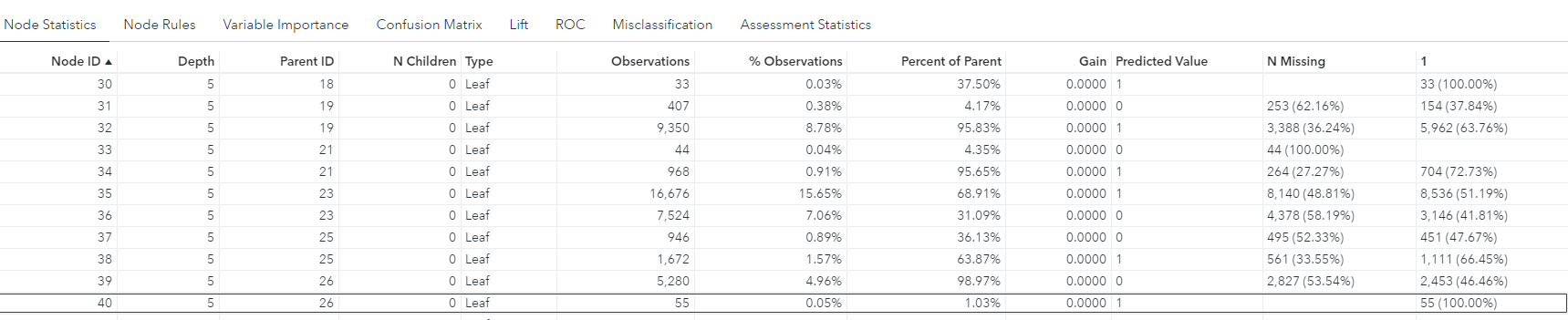
Node ID 12



33

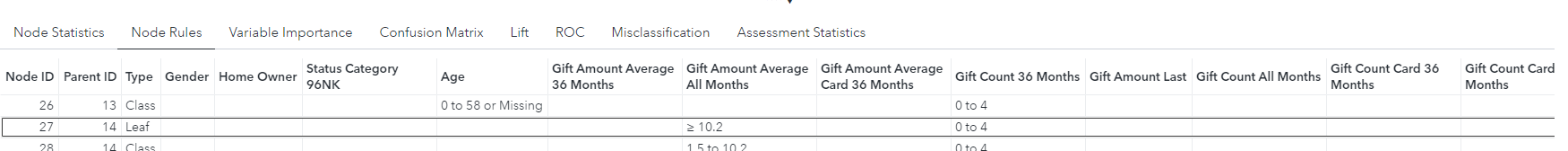


40



* + 1. Click the **Node Rules** tab. Is Node 27 a class node or a leaf node?

Node 27 is Leaf Node

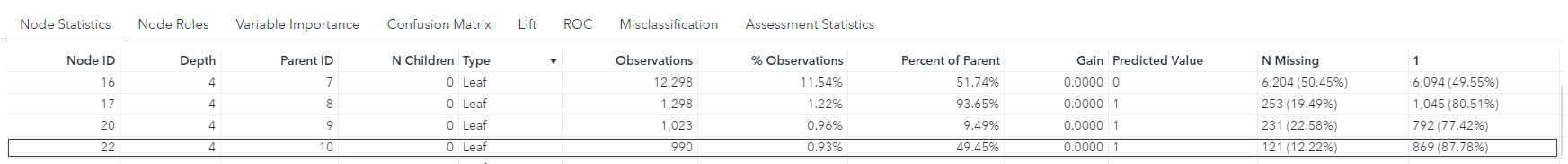


* + 1. Use the node rules to describe the customers in Node 27 in simple business language?

Node 27 is a customer who has gilft amount average all month is bigger or equal to 10.2

* + 1. Examine the Variable Importance table to see whether *media income region* appears to be   
       an important factor when you classify customers who make a donation. Is it important?  
       No, median income region is not important. None of values were measured in section of median income region.
  1. Close the details table.
  2. Maximize the Assessment window and change the variable importance plot to leaf statistics.
     1. Other than the three 100% nodes, which node has the next highest percentage of donors?

Node 22 with 87.87%



* + 1. Which node has the most customers?

In leaf type, node 27 has the most customer

* + 1. Select the node with the most customers and change the chart back to percent   
       to determine the proportion of donors.

37.75% is the proportion of donors

* + 1. Examine the lift chart. What can you determine about the top 10% (percentile)   
       of the data? Explain.

Top 10% about 1.4 tims as many responders as a random ordering of the data, and about 0.6 times fewer responders than perfect ordering of the training data.

* 1. Save your report.

**Week 6 – Homework Exercise 2**

1. Which property setting can you set when modeling decision trees in SAS Visual Statistics?
   1. Seed
   2. Logworth criterion
   3. Variable standardization
   4. Reuse predictors

**Answer: D. Reuse predictors**

*(Simply copy & paste your answer from the options provided)*

1. Which statement is true when creating decision trees in SAS Visual Statistics?
   1. The predictor variables are restricted to measures only
   2. The response variable can be a category or a measure
   3. The predictor variables may include interaction terms
   4. The response variable is limited to category variables only

**Answer: B**

1. To select useful predictors, what type of algorithm do decision trees use?
   1. K-means
   2. Split-search
   3. Principal components
   4. Bootstrapping

**Answer: B**

1. How do classification decision trees identify a good predictor split?
   1. Entropy
   2. Method of least squares
   3. Method of maximum likelihood
   4. F-test

**Answer: A**

1. How do decision trees identify & formulate predictions?
   1. Log-odds
   2. If-Then-Else conditional rules
   3. Sum of squared errors
   4. Convergence criterion

**Answer: A**

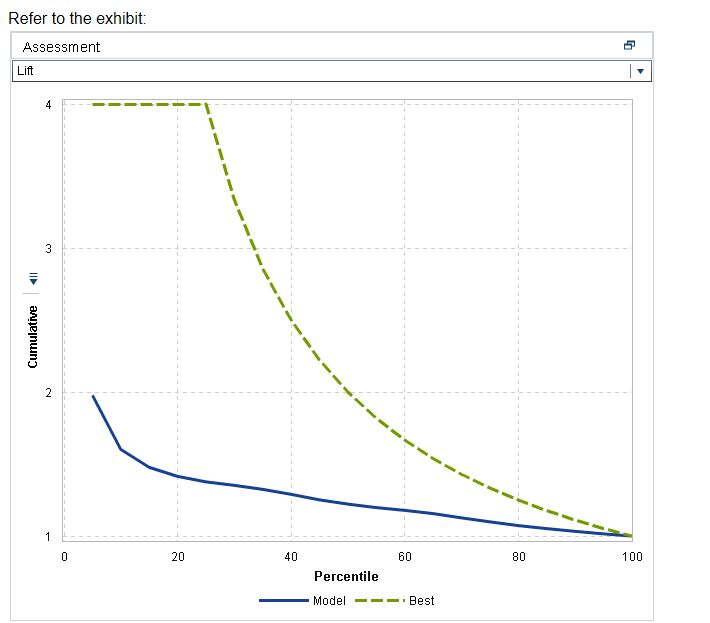
1. In decision trees analysis, why is pruning useful?
   1. Reduces model overfitting
   2. Increases the complexity of the final classifier
   3. Reduces predictive accuracy
   4. Eliminates the need to perform cross validation

**Answer: C**

1. You have created a supervised segmentation using a decision tree. You would like to perform further analysis by building a group-by predictive model using another algorithm (for example, logistic regression) for the segments derived from your decision tree. How would you segment your data for this scenario?
   1. Use the default number of bins
   2. Turn on the rapid growth setting
   3. Export score code
   4. Derive a Leaf ID variable

**Answer: D**

1. Refer to the exhibit below:

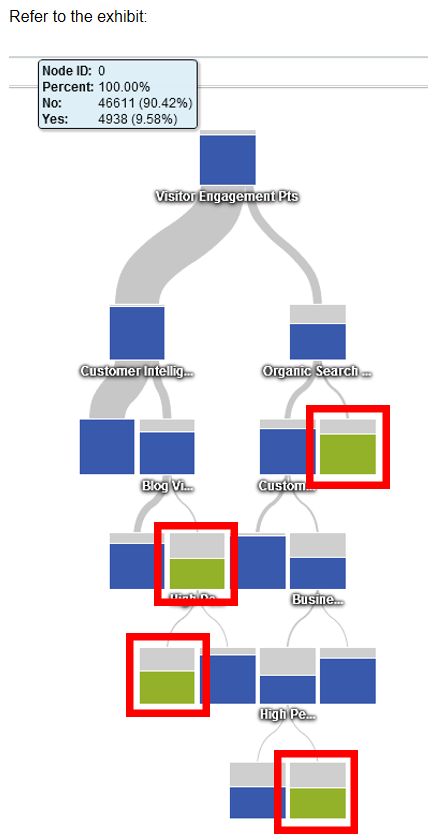


Using the Lift chart above to evaluate a decision tree model where the event level is Purchase, what is the expected performance of the model for the best 20% of cases predicted?

* 1. About 4 times better than a random sample of the same size
  2. Abou**t 1.5 times better than a random sample of the same size**
  3. About 1.1 times better than a random sample of the same size
  4. About 1.0 times better than a random sample of the same size

**Answer: B**

1. Refer to the exhibit below:



What is the correct interpretation of the decision tree nodes that are the color green (highlighted with the red boxes)

* 1. The probability of that an observation is “yes” is 9.58%
  2. The probability of that an observation is “no” is 9.58%
  3. The proportion of observations that are “no” is larger than the observations that are “yes”
  4. The proportion of observations that are “yes” is larger than the observations that are “no”

**Answer: D**

**\*Personal Note : Consider between A or D, I think A is wrong because it said probability.**

1. Refer to the exhibit below:



For this decision tree model with a binary response variable, which outcome highlights the model’s largest number of predictive errors?

* 1. True positives
  2. True negatives
  3. False positives
  4. False negatives

**Answer: C**