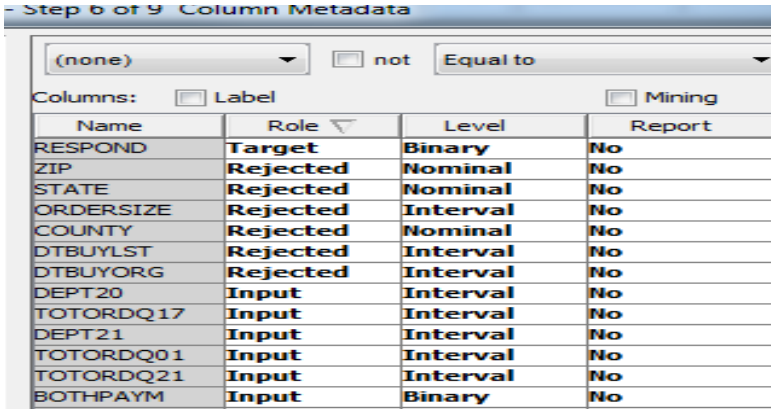
Catalog Logistic Regression

A mail order catalog retailer wants to save money on printing/mailing catalogs and increase revenue by targeting mailed catalogs to past customers who are **most likely** to purchase from the new 2010 catalog. The goal is to increase revenue per mailed catalog.

The Target variable is RESPOND (0,1)
0 = didn't buy in Q1 or Q2 from 2009 catalog 1 = did buy in Q1 or Q2 from 2009 catalog.



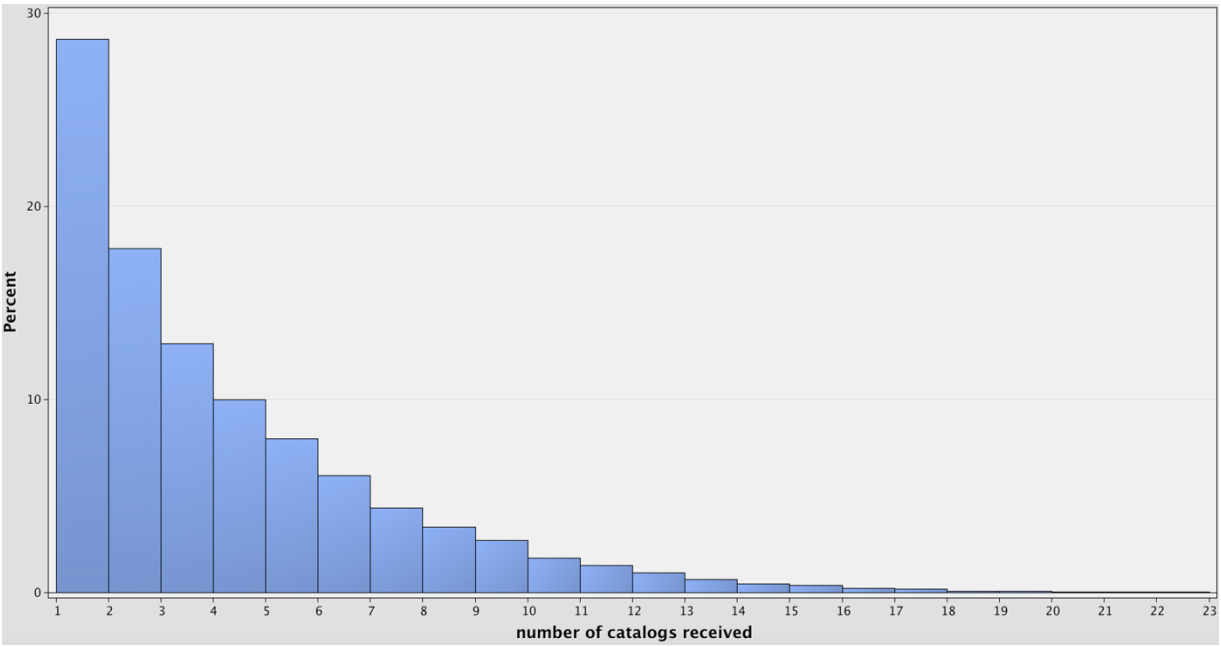
Name	Role	Level	Report
RESPOND	Target	Binary	No
ZIP	Rejected	Nominal	No
STATE	Rejected	Nominal	No
ORDERSIZE	Rejected	Interval	No
COUNTY	Rejected	Nominal	No
DTBUYLST	Rejected	Interval	No
DTBUYORG	Rejected	Interval	No
DEPT20	Input	Interval	No
TOTORDQ17	Input	Interval	No
DEPT21	Input	Interval	No
TOTORDQ01	Input	Interval	No
TOTORDQ21	Input	Interval	No
BOTHPAYM	Input	Binary	No

1. **Create a Diagram – name it Catalog**
   a. Drag the data source onto the diagram and RUN the data source.
   b. From the results, describe the following IVs (use variable labels).
      i. Catalogcnt: number of catalogs received
      ii. DayLast: days since last
      iii. Doll 24: $ last 24 months
      iv. DollarQ variables: tot $ (quarterly)

2. What is the business purpose of this analysis? To target past customers who are most likely to purchase from the new 2010 catalog.
3. How many customers are in the data set? 48,356
4. How many variables are captured for each customer? 98
5. Why shouldn't you just mail a catalog to all customers in the database? Mailing a catalog to all the customers would be ineffective and expensive; the retailer would waste a significant amount of money trying to attain customers who may be projected as 'unlikely to purchase' the new 2010 catalog.
6. Select the data source node on the diagram; select the Options Tab – Preferences
   a. Make sure: Property Sheet Tooltips is **On**, Sample method – **Random**, Fetch Size – **Max**

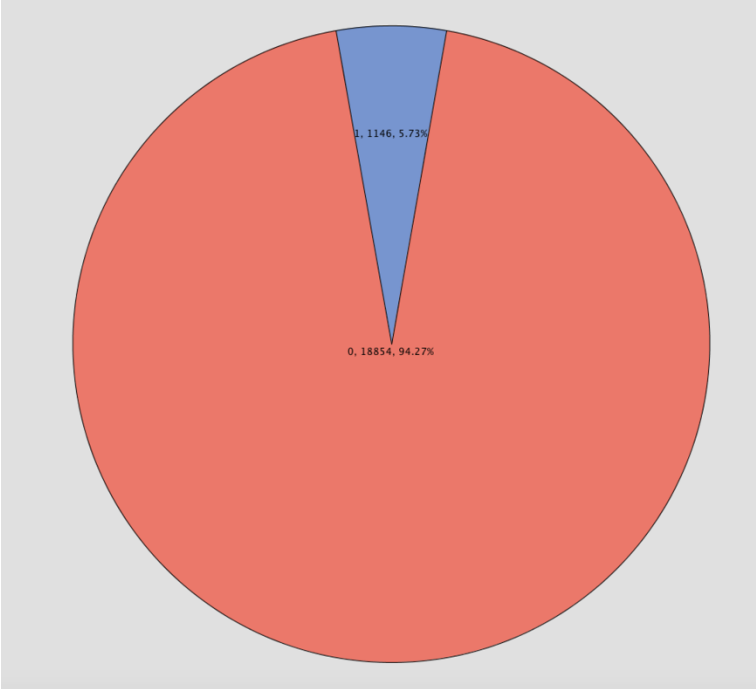
7. **Data Exploration**
   a. Most customers are from what state? CA
   b. What is the min, max and mean number of catalogs that were sent to customers (CatalogCNT)? Min=1; Max=23; Mean=3.76145

   c. **Copy/paste** a histogram of the Catalogcnt with each bar representing 1 catalog. (Actions-Plot)



i. What percentage of customers has only received 1 catalog? 28.66%
ii. What % of customers has received between 9-10 catalogs? 2.70%
iii. Describe the skew of the Catalogcnt distribution. Positive, right skew distribution

d. **Copy/paste** a pie chart of Respond (DV) with both value and % on the Pie chart.



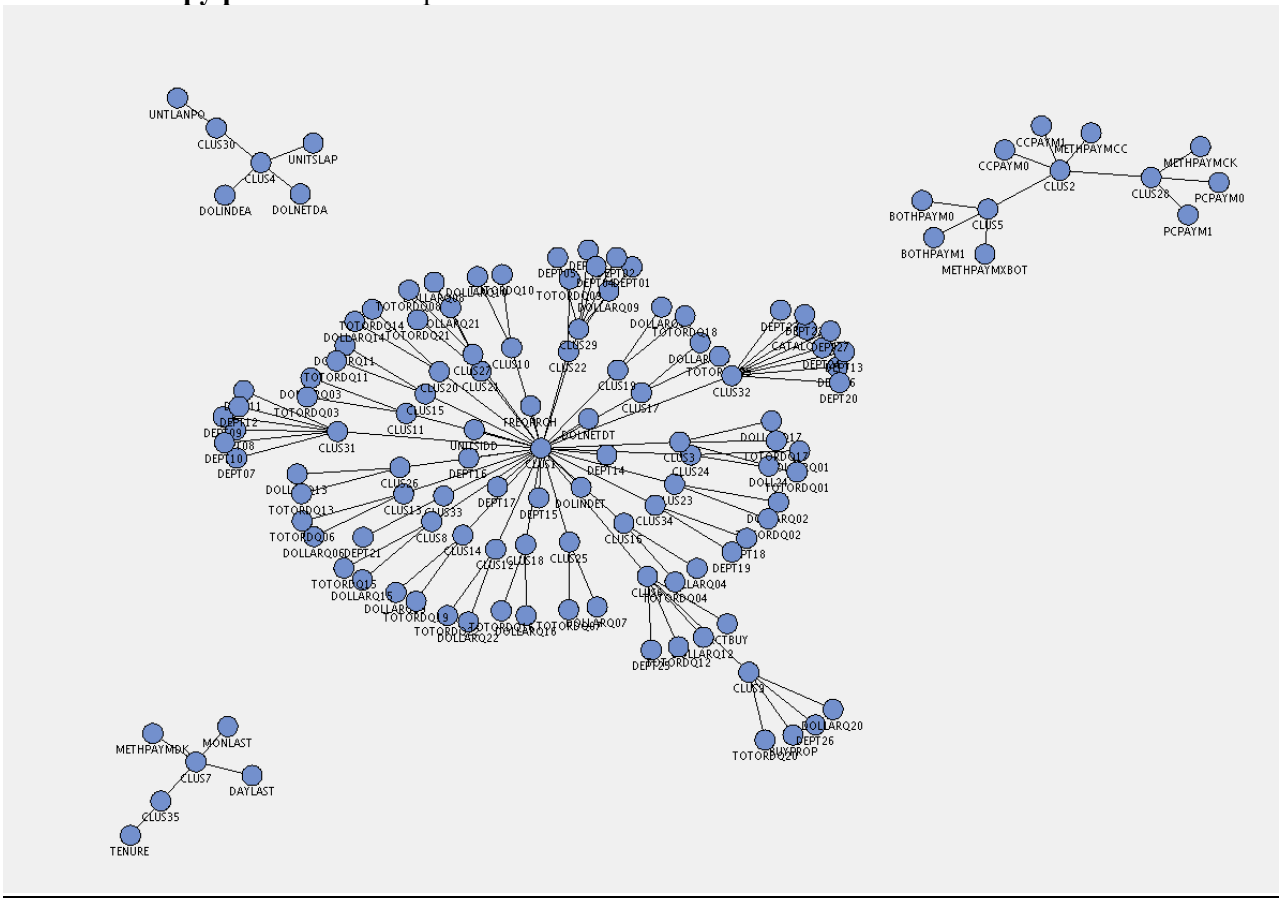
i. What % of customers purchased in Q1 or Q2 from the 2009 catalog? 5.73%

8. **Data Partition Node**
   a. Split the data into 2:1 proportions (put in 2 as training; 1 as validation; 0 as test) - RUN
   b. Summary Statistics window:
      i. What percentage of customers in the data set **DID NOT** purchase in Q1/Q2 2009? 94.3%
   c. Why are the % of 0s and 1s about the same in the training and validation data sets? Both sets were randomly broken up from the same (original) data set

9. **Transform Node** – skew is acceptable for the IVs, do not add transform node

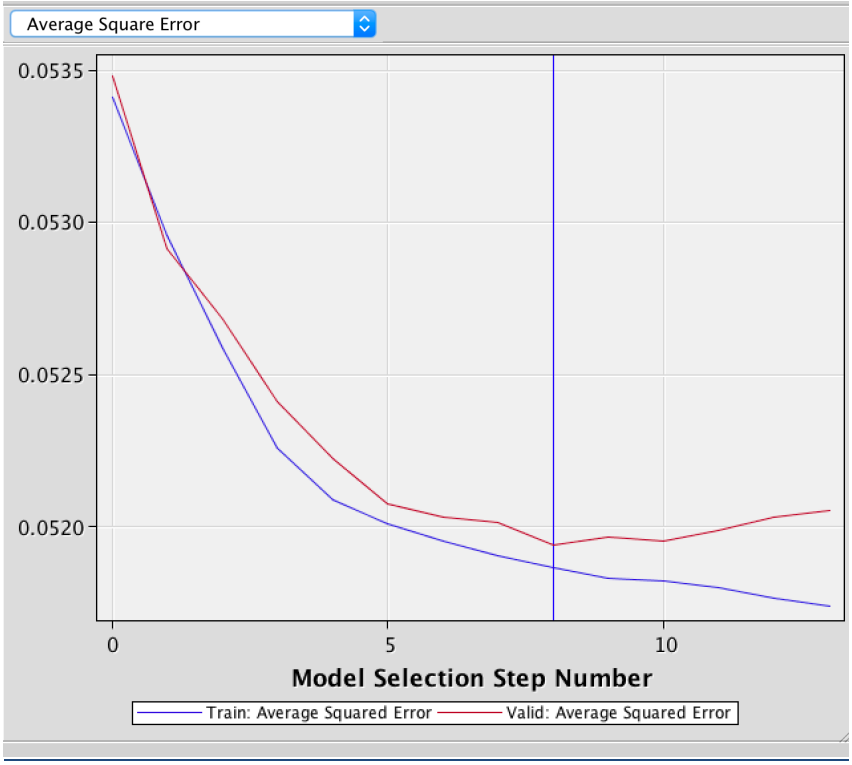
10. **Imputation Node**
    a. No missing data – do not add imputation node

11. **Variable Clustering Node**: (**Note**: Variable Clustering node is different from Clustering node)
    a. Attach Data Partition node to Variable Clustering Node
    b. Properties Panel:
        i. Train: Change Includes Class Variable to YES,
        ii. Score: Change Variable Selection to BEST VARIABLES – RUN
    c. **Copy/paste** the cluster plot.



    d. What is the purpose of the Variable Clustering Node? To reduce a large number of variables to a smaller number of 'best' variables for inclusion in logistic regression models. Reduce redundancy and dimensionality.
    e. What stat determines the best variable to represent a cluster? The lowest 1 – R2 ratio
    f. The 98 IVs are grouped into how many clusters? 35
    g. What variable is selected as the best representative of Cluster #32? CATALOGCNT

12. **Regression Node** - Add a regression node (name it Forward).
    a. Property Panel: Model Selection: Choose **Forward** as the 'selection model', **Validation Error** as the 'selection criterion', **YES** for 'use selection defaults' - RUN.
    b. View – Model – Iteration Plot
        i. **Copy/Paste** the Iteration Plot with Avg Sq Error



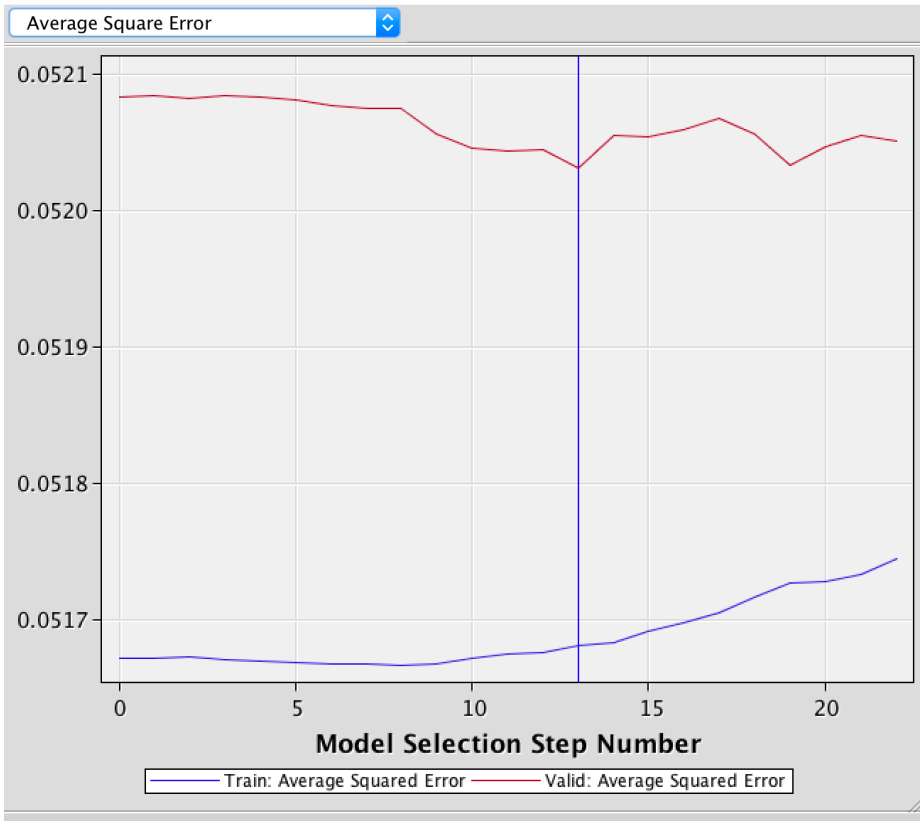
ii. What is happening with the error rate on training and validation sets at Step 8? At step 8, the error rate for the training set decreases which the error rate for the validation set increases

c. Output Window: In the Chart, enter each IV in the Forward logistic regression model

Forward Selection
CATALOGCNT
DOLINDET
MONLAST
TOTORDQ12
TOTORDQ18
TOTORDQ20
TOTORDQ21
TOTORDQ22

d. What two variables are most important in the prediction of a customer purchasing in Q1/Q2 from the 2009 catalog? (Maximum Likelihood Estimates table -use absolute value)
MONLAST; TOTORDQ20

13. Add another regression node (name it Backward).
    a. Choose **Backward** as the selection model with **Validation Error** as the selection criterion and **YES**.
    b. **Copy/Paste** the Iteration Plot with Avg Sq Error.

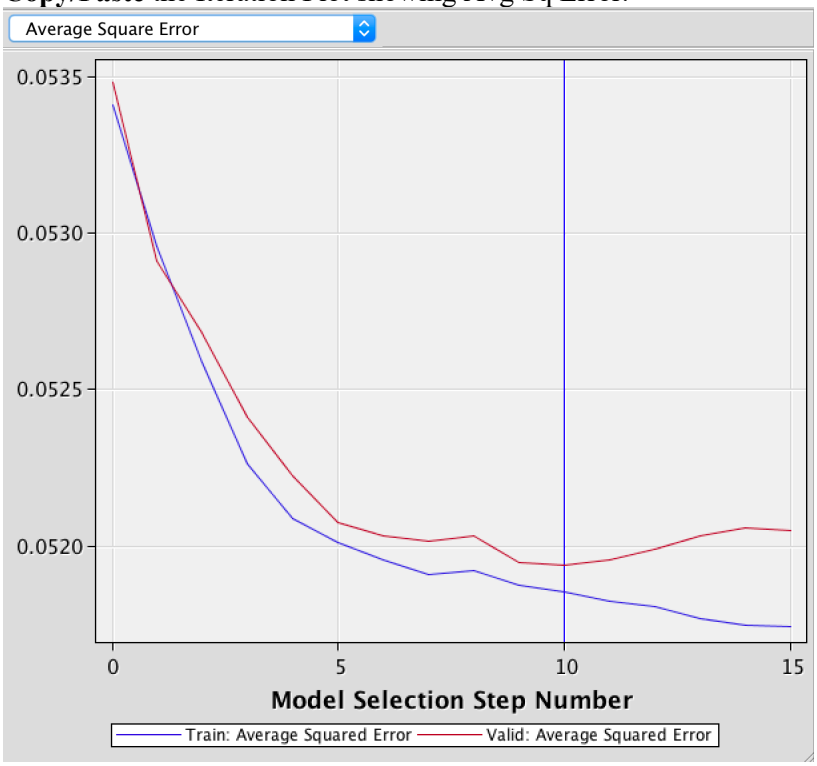


i. At what step does the error rate start to increase for the Validation data set? 13

c. How many IVs are in the final regression model? **The final model is directly above the 4 horizontal lines near the end of the output. 22

d. Which two variables are most important in the prediction of who is likely to purchase from the new catalog? MONLAST; CCPAYM0

14. Add another regression node (name it Stepwise).
    a. Choose **Stepwise** as the selection model with **Validation Error** as the selection criterion and **YES**.
    b. **Copy/Paste** the Iteration Plot showing Avg Sq Error.



i. What does the blue vertical line represent? (DO NOT say optimal tree!) It represents the number of steps (at which the selected model is trained on)

c. In the Chart below, enter each IV in the final logistic regression model.

Stepwise Selection
CATALOGCNT
DEPT03
MONLAST
TOTORDQ12
TOTORDQ18
TOTORDQ20
TOTORDQ21
TOTORDQ22

d. What are the **3 most important variables** in the prediction of those most likely to purchase from the new 2010 catalog? MONLAST; TOTORDQ20; CATALOGCNT

15. **Model Comparison Node**
    a. Connect all regression nodes to it, change the Selection Statistic to **Avg Sq Error,** selection table to **Validation**. RUN.
    b. What does the ROC chart show in the comparison of the three models? All three models are good models
    c. Which model is best based on the lowest **Validation Avg Sq Error**? Stepwise

16. Open the Output Results of the Best Regression Model (from 16 above).
    a. Type out the logistic regression equation using all the variables (you can abbreviate the var names). (0.0886CATALOGCNT + 0.0386DEPT03 -0.1286MONLAST + 0.0365TOTORDQ12 + 0.0424TOTRDQ18 + 0.0972TOTORDQ20 + 0.0415TOTORDQ21 + 0.0628TOTORDQ22)
    b. How will this formula be used? Each record in a new data set can be inputted into the equation
    c. **Copy/Paste** the Odds Ratio Estimates.

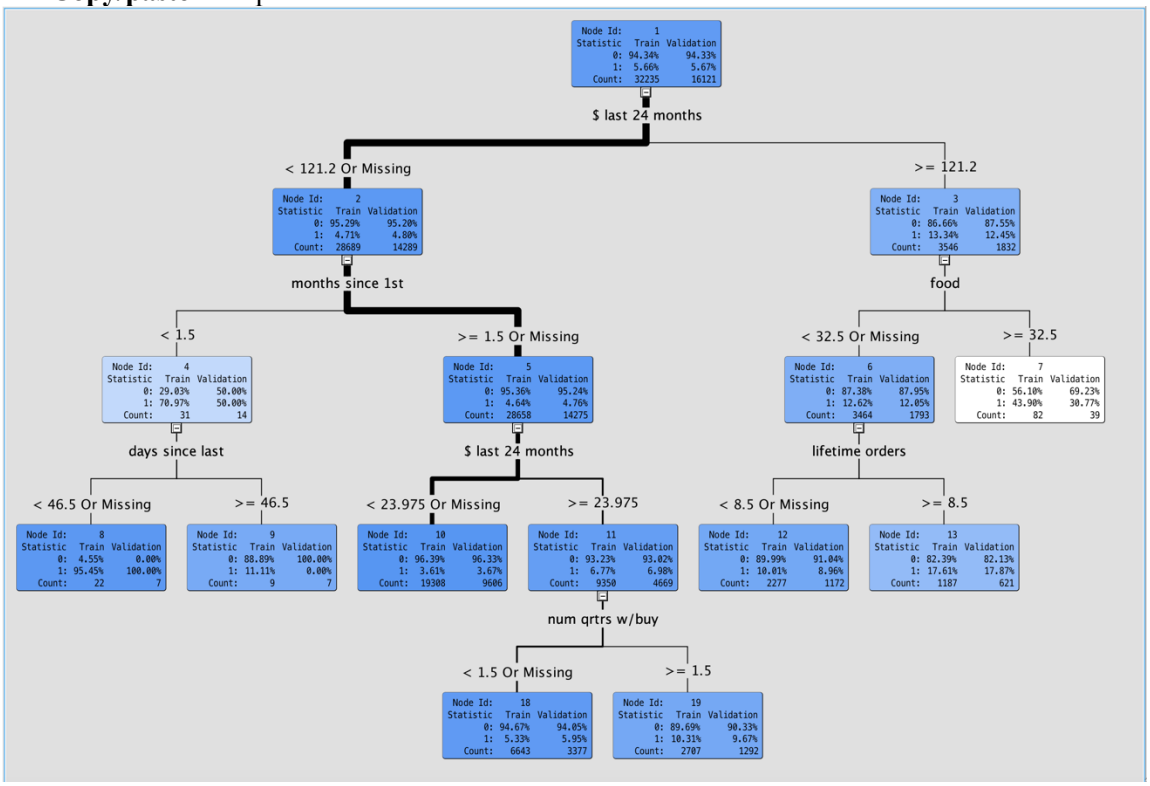
Odds Ratio Estimates

Effect	Point Estimate
CATALOGCNT	1.053
DEPT03	1.025
MONLAST	0.994
TOTORDQ12	1.177
TOTORDQ18	1.265
TOTORDQ20	1.465
TOTORDQ21	1.256
TOTORDQ22	1.437

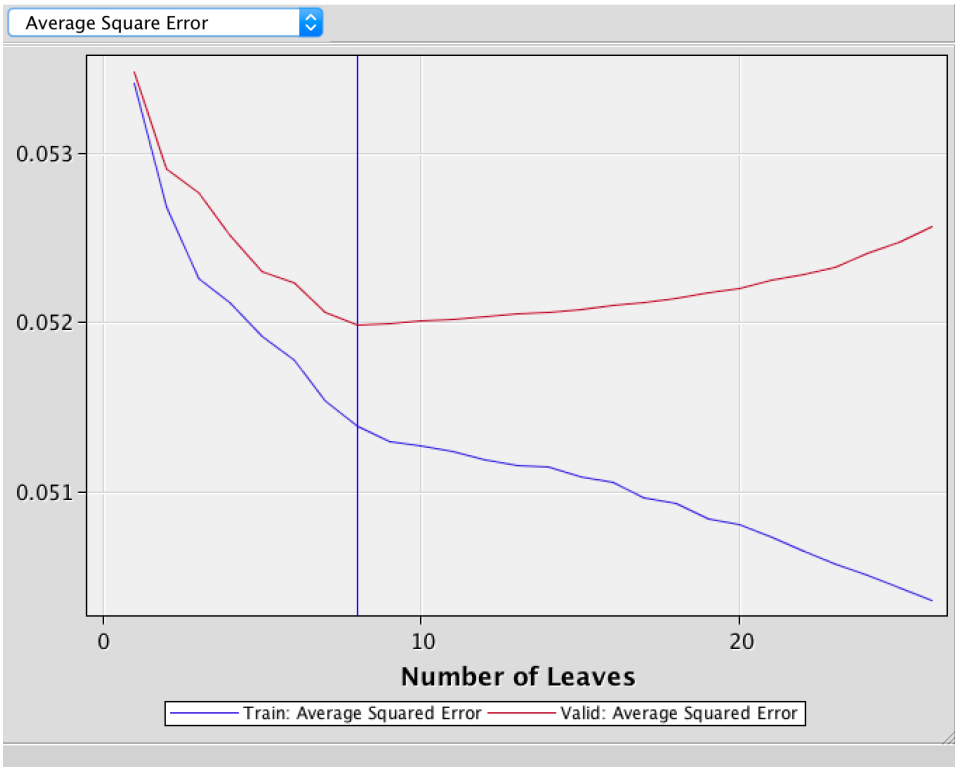
d. Fill in the following chart with the interpretation of the Odds Ratio Estimates:

IV	Point Estimate	Interpretation
Catalogcnt	1.053	For every additional catalog received, there is a 5.3% increase in **odds** of purchasing from the 2010 catalog.
Monlast	0.994	For every additional month since last purchase, there is a 0.6% decrease in **odds** of purchasing from the 2010 catalog.
TotordQ22	1.437	For every additional order in Q22, there is a 43.7% increase in **odds** of purchasing from the 2010 catalog.
Dept03	1.025	For every additional purchase from department 3, there is a 2.5% increase in **odds** of purchasing from the 2010 catalog.

17. Add a Decision Tree node and connect it to the data partition node.
    a. Change the subtree assessment measure to **Average Square Error -** RUN
    b. **Copy/paste** the optimal tree

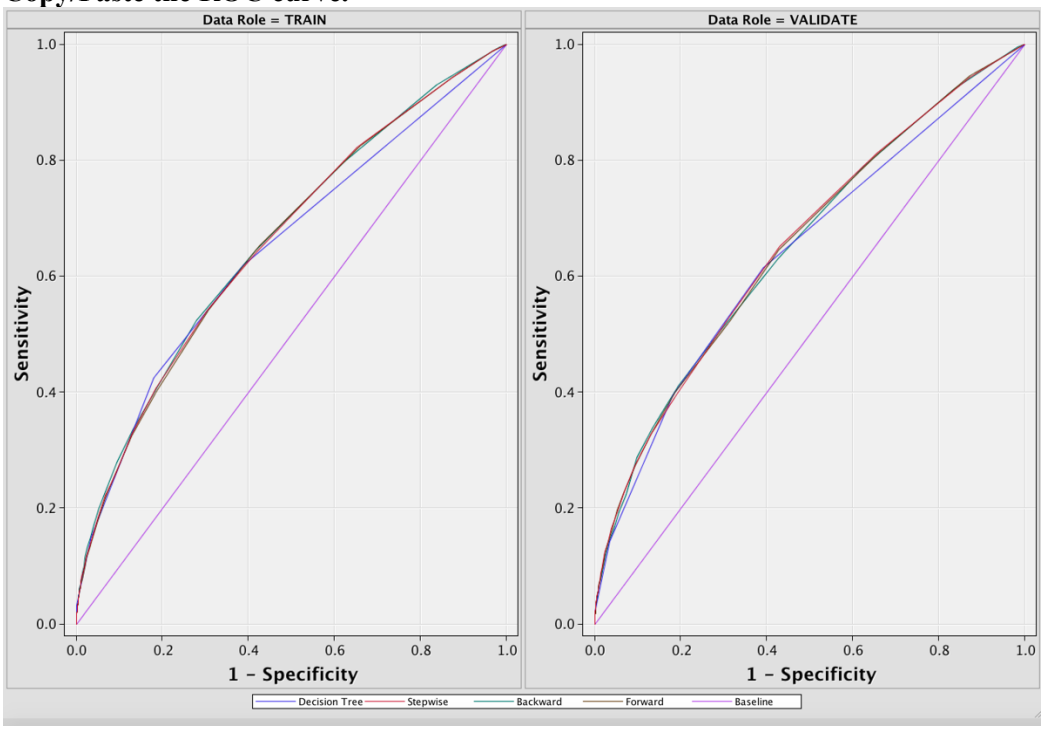


    i. Describe the customers with a 95.45% likelihood of buying from the new 2010 catalog. Those customers spent less than $121.20 (or missing) in the last 24 months, been less than 1.5 months since their first purchase, and has been less than 46.5 days (or missing) since their last purchase.
    ii. What % of customers spending less than $25 are predicted to buy from 2010 catalog? 3.61%
    c. **Copy/paste** the subtree assessment plot.



i. How many leaves are in the optimal tree? 8

18. Connect the Tree node to the Model Comparison node – RUN
    a. **Copy/Paste the ROC curve.**



b. Fill in the chart for the best models based on Validation Avg Sq Error.

	Model Name
Best Model	Stepwise
2nd Best Model	Forward
3rd Best Model	Decision tree
4th Best Model	Backward

19. Using the Odds Ratio Table or Maximum Likelihood Table for the **Best Model** above:
    a. In general, customers don't like getting our catalogs and are less likely to order from the new catalog. TRUE or **FALSE**
    b. Customers who order products from Department 3 are more likely to order again. **TRUE** or FALSE
    c. The more months that pass since a customer's last order, the more likely they will not order from the new catalog. **TRUE** or FALSE
    d. Ordering in Quarter 22 is a better predictor of ordering from next catalog than ordering in Quarter 20. TRUE or **FALSE**