SCHOOL OF COMPUTER APPLICATION

PREDICTIVE ANALYTICS

BCADS1501

Submitted By:

Name: Siddhant Kumar Patel

Section: BCA DS 36

University Roll No: 1230258423

Semester: 5th

Submitted To:

Mr. Ayushman Bhadauria

Babu Banarasi Das University, Lucknow

BBD City, Ayodhya Road, Lucknow Uttar Pradesh- 226028 Bharat

Case Study

"ACME, a company selling sports products, wants to promote its new product: the XL Original Orange Baseball Cap. To test customer interest, ACME sent a test mailing to 10,000 randomly selected customers and recorded their responses."

Agenda/Definition:

- The agenda is to **build a predictive model** using the specified dataset to estimate an **unknown or future value (the target/output variable)**.
- To apply a suitable machine learning algorithm (e.g., Linear Regression, Decision Tree, Neural Network) to the ACME dataset within SPSS Modeler to establish a relationship between input variables (predictors) and a target variable, and then use that established relationship (the model) to predict the output for new, unseen data.

Outcomes/Learning:

- Proficiency in **data importation** and **pre-processing** (e.g., data auditing, missing value handling, variable type setting) within SPSS Modeler.
- Ability to **select and deploy appropriate modeling nodes** (algorithms) for the specific prediction task (e.g., classification or regression).
- Skills in **evaluating model performance** using appropriate graphs and statistics (e.g., gain charts, accuracy matrices, R^2, RMSE).
- Understanding the end-to-end **CRISP-DM** (Cross-Industry Standard Process for Data Mining) process within a visual environment.

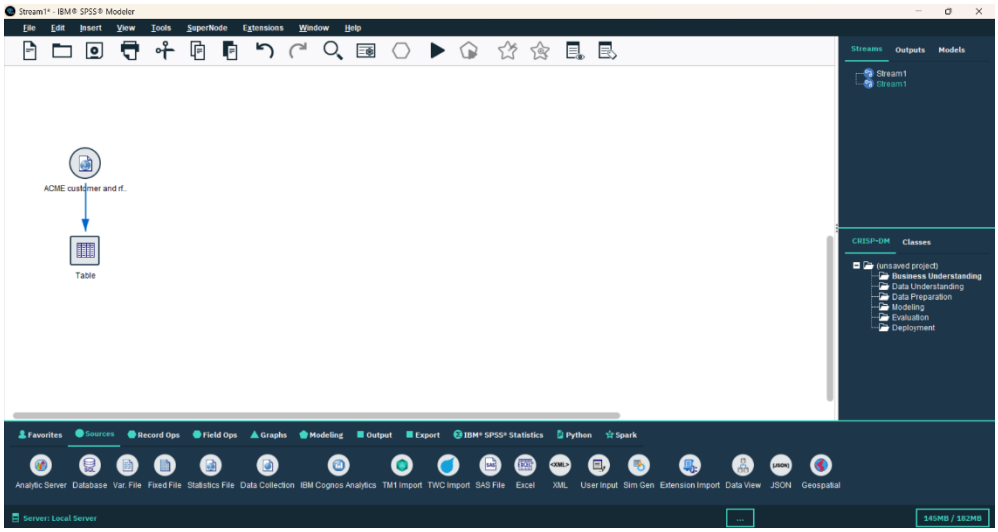
Required Tool:

The specific software required to perform the modeling and prediction is:

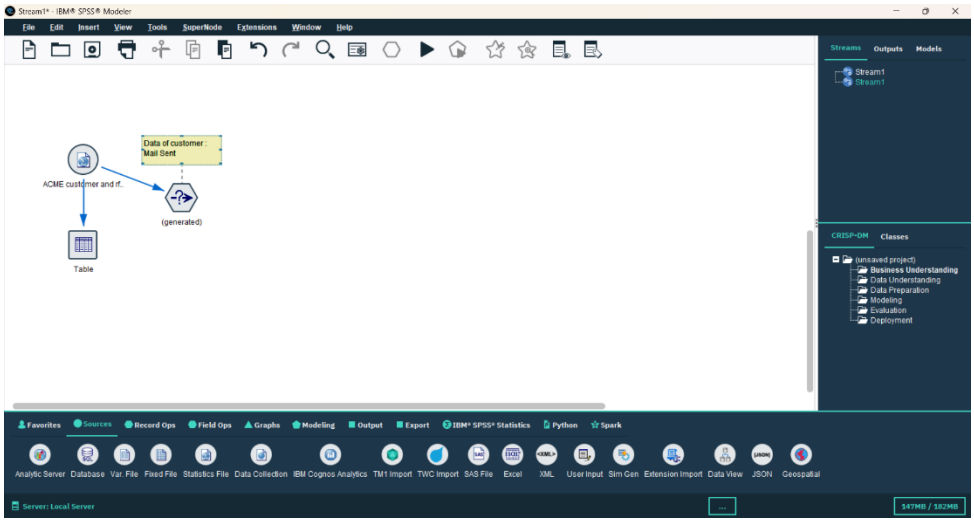
- **IBM SPSS Modeler:** This is the primary tool. It's a graphical workbench for building models without needing to write code.
- **ACME Dataset:** The required data source (in a format compatible with Modeler, like a flat file .csv,.txt, or a database connection).

Steps:

1. Import the dataset, use the table to get all the data (12 columns and 30000 rows)



2. Filter the data for the "YES" value and generate the node, rename as training data.



3. Connect that node with table to get the desired output. (10000 rows)

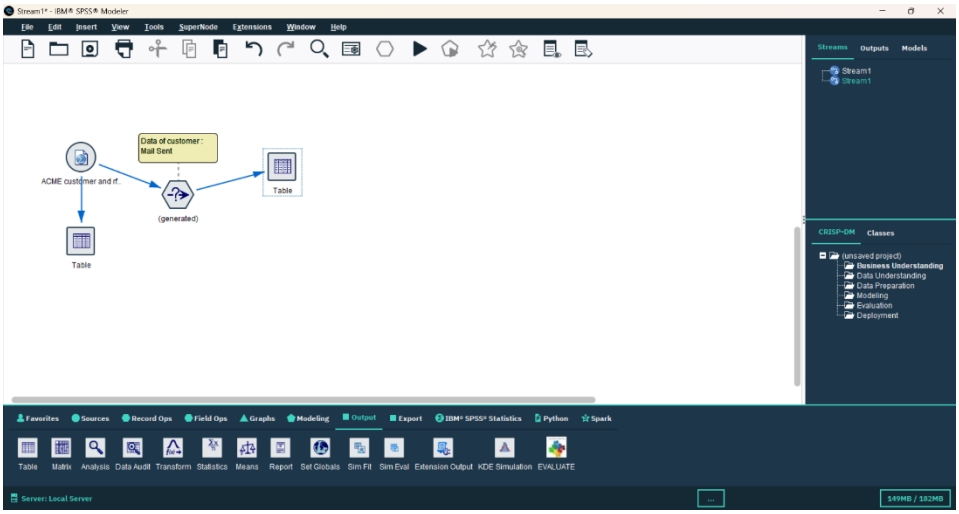
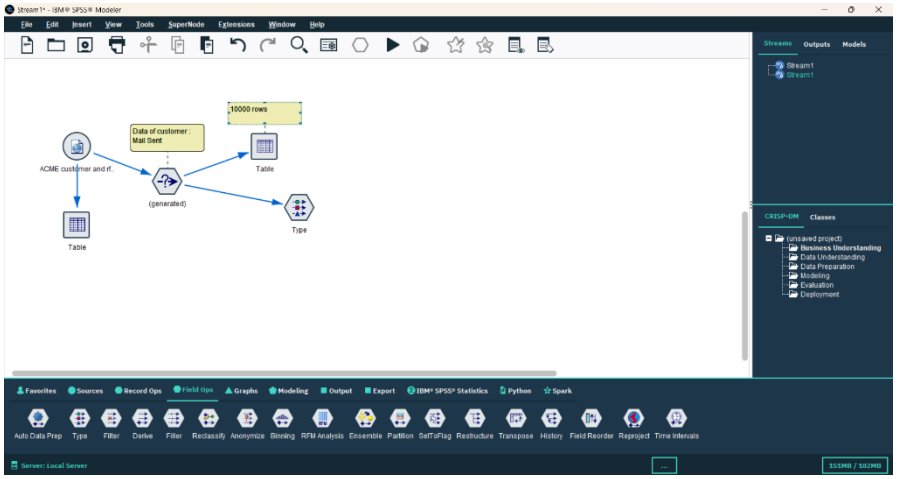


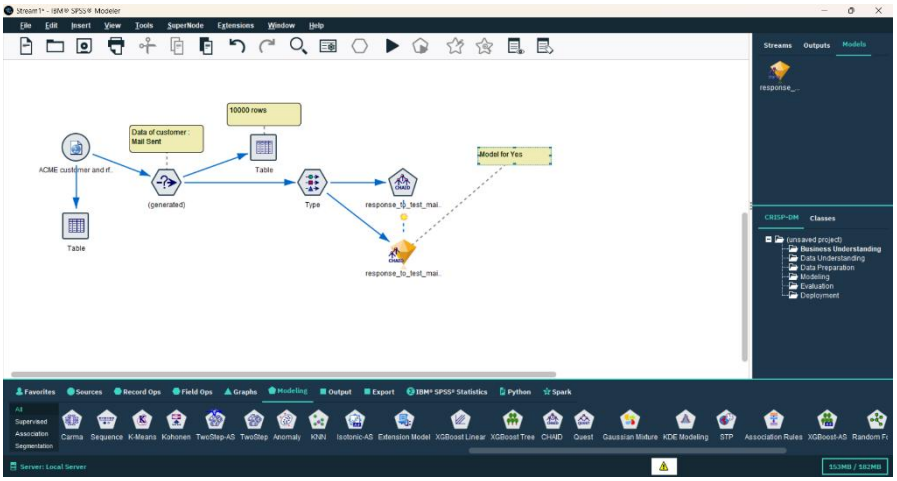
Table Annotations

	customer_id	gender	email_address	postal_code	monetary_value_01_01_2011	frequency_01_01_2011	recency_01_01_2011	has_received_test_mailing
1	723.000	male	name7502@tnet.fr	1818BO	2 medium	3 high	2 medium	yes
2	724.000	female	name25485@wwmail.org	1132DG	1 low	3 high	1 low	yes
3	725.000	male	name15543@wwmail.de	1803YT	3 high	1 low	1 low	yes
4	726.000	male	name28335@zigzag.be	1205WR	3 high	1 low	3 high	yes
5	727.000	female	name5354@tnet.jp	1711ON	1 low	3 high	1 low	yes
6	728.000	female	name20637@wwmail.es	1055FG	2 medium	3 high	1 low	yes
7	729.000	female	name20636@wwmail.es	1254MR	1 low	3 high	1 low	yes
8	730.000	female	name10414@tnet.inc	1723DG	2 medium	3 high	1 low	yes
9	731.000	male	name23372@wwmail.inc	1713AQ	3 high	2 medium	1 low	yes
10	732.000	male	name20635@wwmail.es	1264EC	3 high	2 medium	3 high	yes
11	733.000	female	name5356@tnet.jp	1648BT	3 high	2 medium	1 low	yes
12	734.000	female	name17582@wwmail.de	1285XV	3 high	1 low	3 high	yes
13	735.000	female	name6388@tnet.fr	1282NB	1 low	2 medium	2 medium	yes
14	736.000	male	name10409@tnet.inc	1799IT	3 high	2 medium	1 low	yes
15	737.000	female	name13849@tnet.uk	1802UO	2 medium	3 high	1 low	yes
16	738.000	male	name25473@wwmail.org	1971NK	1 low	3 high	1 low	yes
17	739.000	male	name13848@tnet.uk	1361RL	2 medium	3 high	1 low	yes
18	740.000	female	name23366@wwmail.inc	1164VN	3 high	2 medium	1 low	yes
19	741.000	female	name3188@molbe.cat	1767YN	3 high	1 low	1 low	yes
20	742.000	male	name1606@lomejor.es	1681HP	1 low	3 high	1 low	yes

4. Drag and drop the type node and connect with the training data node.



5. Use the CHAID model and train the model with data.



6. Use the table to get the result after the model is trained.

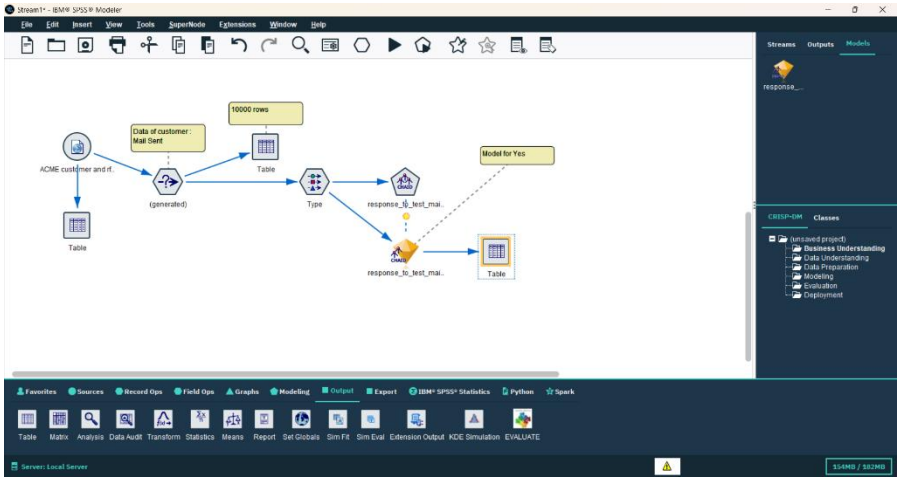
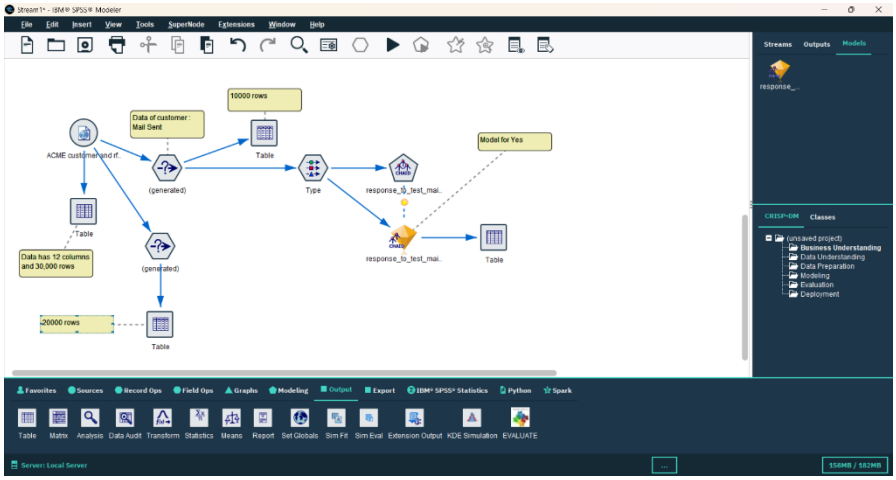


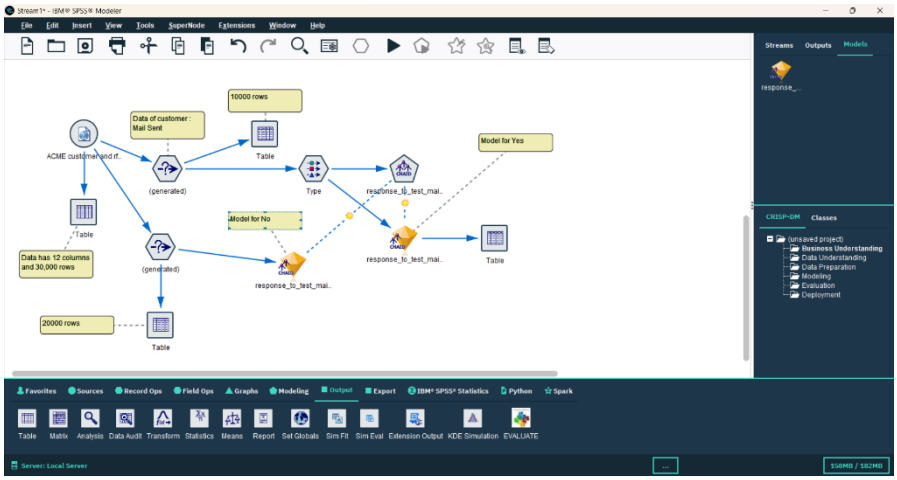
Table Annotations

	number_of_days_between_mailing_and_orderdate	ordered_within_month	$R-response_to_test_mailing_02_01_2011	$RC-response_to_test_mailing_02_01_2011
1	$null$	nap	F	0.953
2	$null$	nap	F	0.993
3	$null$	nap	F	0.953
4	$null$	nap	F	0.953
5	$null$	nap	F	0.993
6	3.000	yes	F	0.953
7	$null$	nap	F	0.993
8	$null$	nap	F	0.953
9	$null$	nap	F	0.911
10	14.000	yes	T	0.625
11	$null$	nap	F	0.911
12	$null$	nap	F	0.953
13	$null$	nap	F	0.998
14	$null$	nap	F	0.911
15	$null$	nap	F	0.953
16	$null$	nap	F	0.993
17	$null$	nap	F	0.953
18	$null$	nap	F	0.911
19	$null$	nap	F	0.953
20	$null$	nap	F	0.993

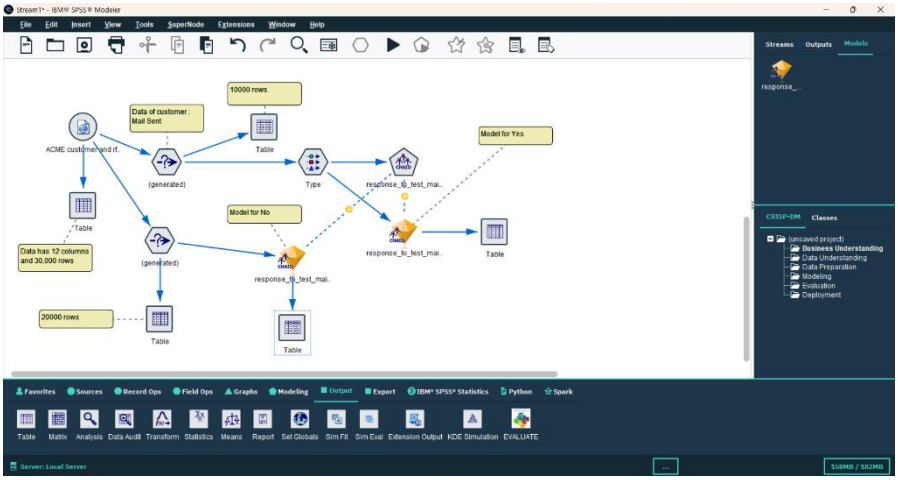
7. Go back and filter the data for the “NO” value and generate the node, rename as test data for testing out model.



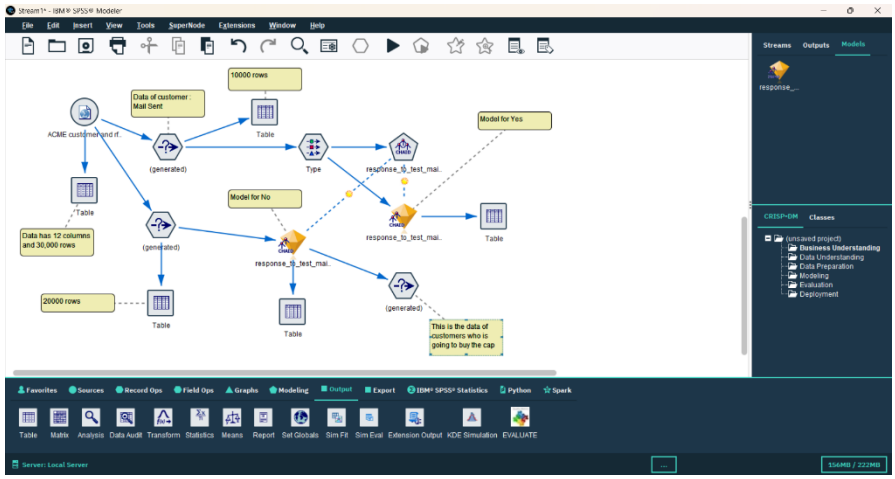
8. Make the copy of the model and connect with the test data.



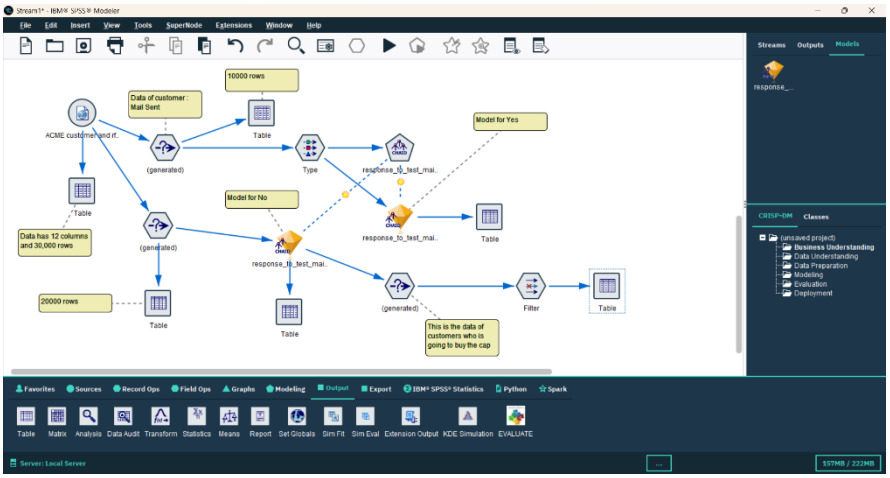
9. Use the table along with all nodes to get the result.



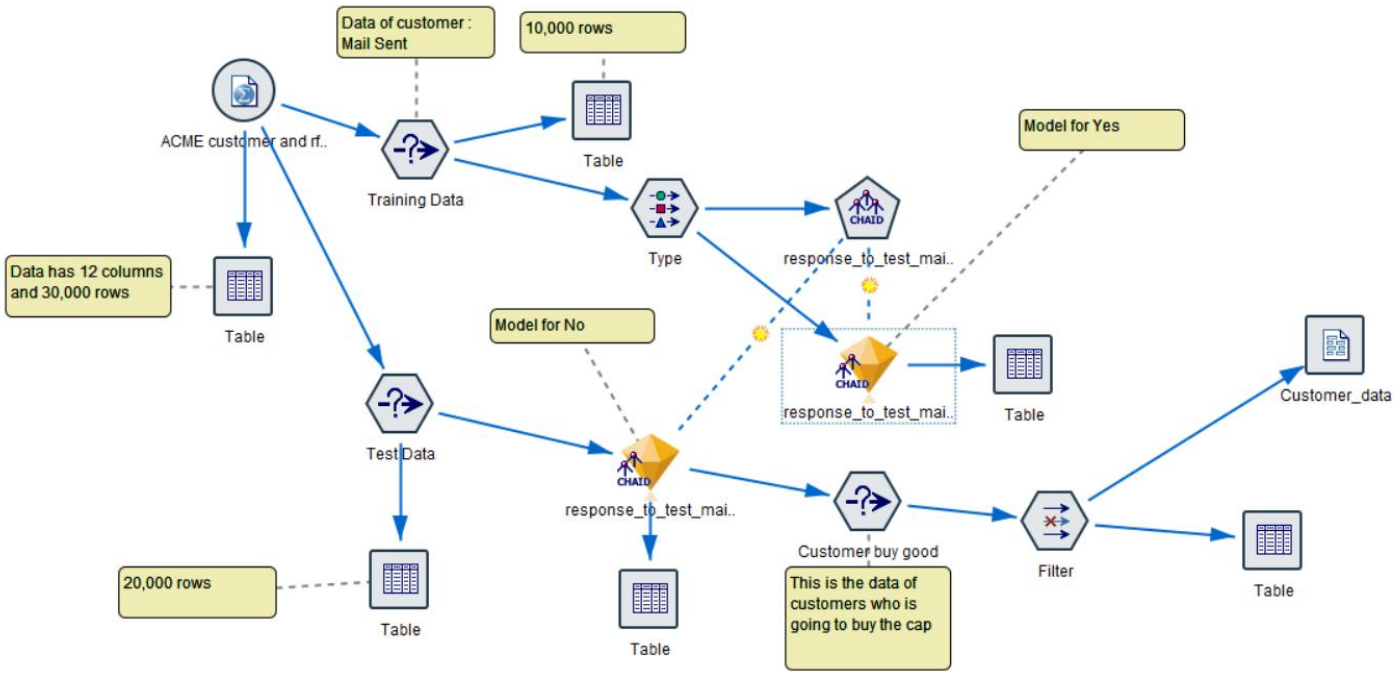
10. Again generate a new node for the “TRUE” data in the model. (Data of customers who are going to buy the cap)



11. Use the filter node to rename the specific columns and filter the data that we needed.



12. After all export the data to your local file with choosing the flat file node and specify the path where you want to save.



Q 1. How many records are in the training dataset?

Ans. There are 30000 Rows in the dataset.

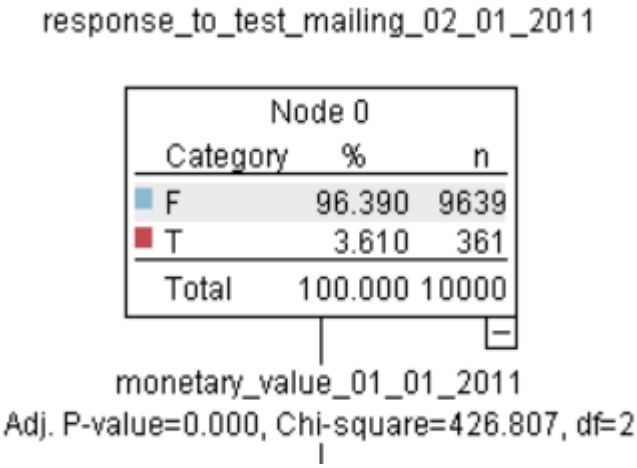
Q 2. How many fields are in the training dataset?

Ans. There are 12 Columns in the dataset.

Q 3. How many customers were included in the test mailing?

Ans. There are 20000 customers are included in the test mailing.

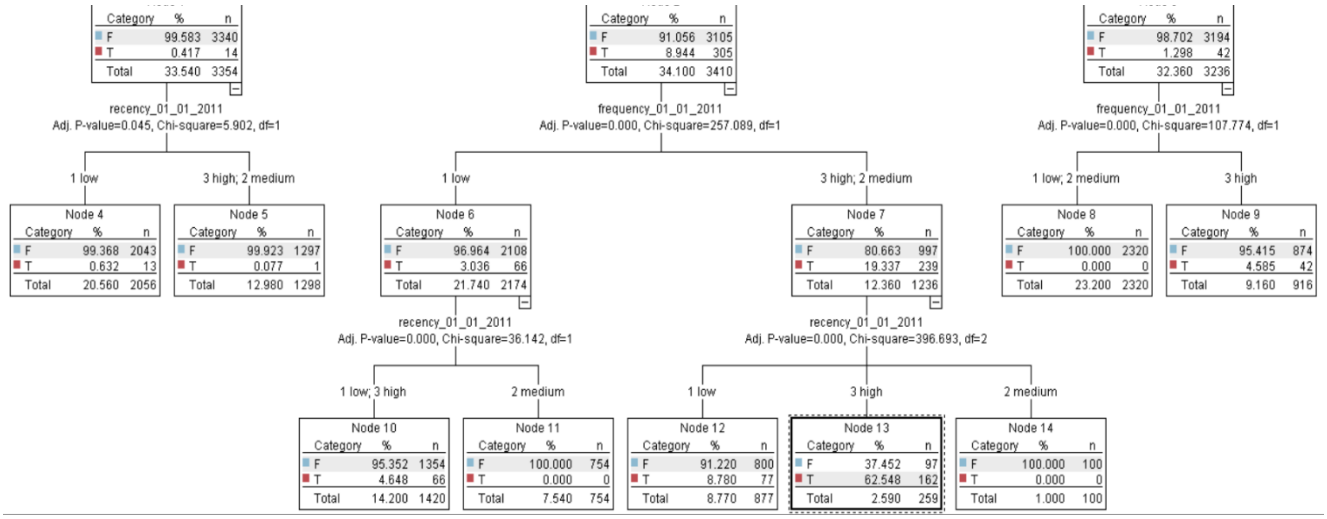
Q 4. Which field is used as the first split?



Ans. **monetary_value_01_01_2011** fields is used as the first split.

Q 5. Which group shows the highest response rate? What is the probability of responding for this group?

Ans. **recency_01_01_2011** shows the highest response rate and the probability of responding for this group is **62.548 %**



Q 6. Identify the two new fields added by the model.

Ans. $R-response_to_test_mailing_02_01_2011 and

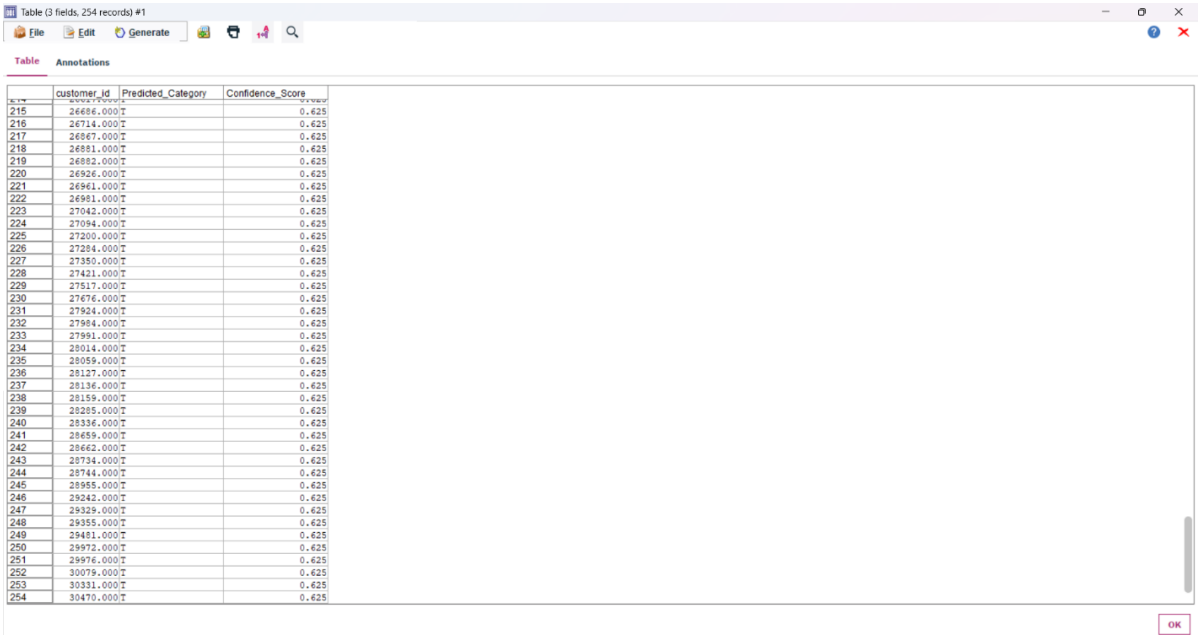
$RC-response_to_test_mailing_02_01_2011 are two new fields added by the model.

Q 7. What do these fields represent?

Ans. The first fields represents the True/False value for the customers who are going to buy the cap or not. The second fields represents the probability of that condition is going to true.

Q 8. How many customers are predicted to respond positively (predicted = T)?

Ans. There are **254** customers are predicted to respond positively (predicted



= T).

After exporting -