SID : 22083095
Name : LIM MIN KYE ANDY

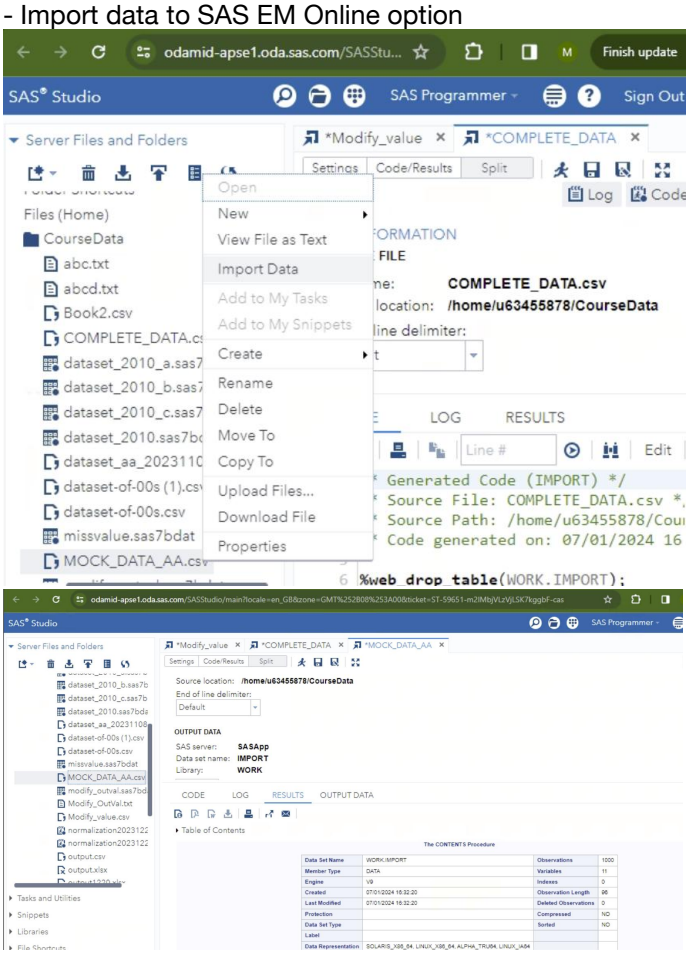
GitHub Link : https://github.com/22083095/wgd7005

Data Import and Preprocessing
Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles.

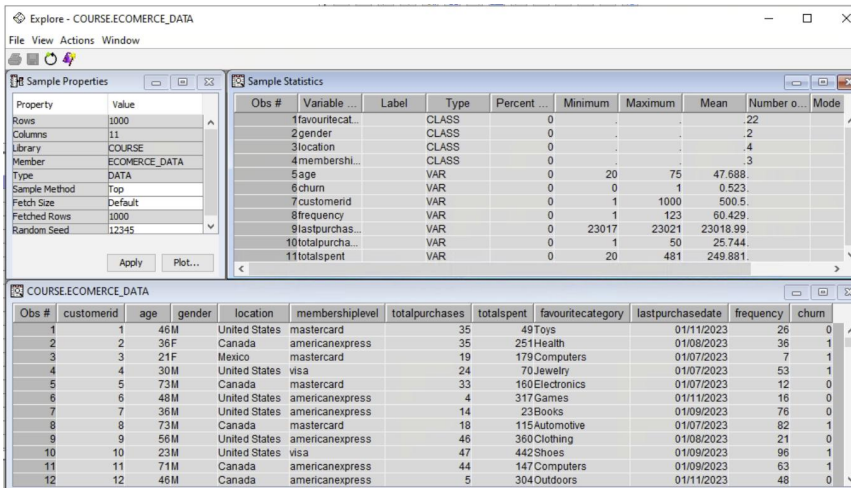
- 1000 response on ecommerce are generated from https://www.mockaroo.com/

```
customerid,age,gender,location,creditcard,totalpurchases,totalspent,favouritecategory,lastpurchasedate,frequency,churn
1,46,M,United States,mastercard,35,49,Toys,01/11/2023,26,0
2,36,F,Canada,americanexpress,35,251,Toys,01/08/2023,36,1
3,21,F,Mexico,mastercard,19,179,Toys,01/07/2023,7,1
4,30,M,United States,visa,24,70,Toys,01/07/2023,53,1
5,73,M,Canada,mastercard,33,160,Toys,01/07/2023,12,0
6,48,M,United States,americanexpress,4,317,Toys,01/11/2023,16,0
7,36,M,United States,americanexpress,14,23,Toys,01/09/2023,76,0
8,73,M,Canada,mastercard,18,115,Toys,01/07/2023,82,1
9,56,M,United States,americanexpress,46,360,Toys,01/08/2023,21,0
10,23,M,United States,visa,47,442,Toys,01/09/2023,96,1
11,71,M,Canada,americanexpress,44,147,Toys,01/09/2023,63,1
12,46,M,Canada,americanexpress,5,304,Toys,01/11/2023,48,0
13,39,M,Canada,visa,48,252,Toys,01/07/2023,27,1
14,29,M,United States,americanexpress,25,456,Toys,01/09/2023,110,0
15,44,F,United States,americanexpress,23,326,Toys,01/09/2023,44,0
16,52,F,United States,visa,30,119,Toys,01/09/2023,66,0
17,33,F,United States,visa,43,473,Toys,01/11/2023,35,1
```

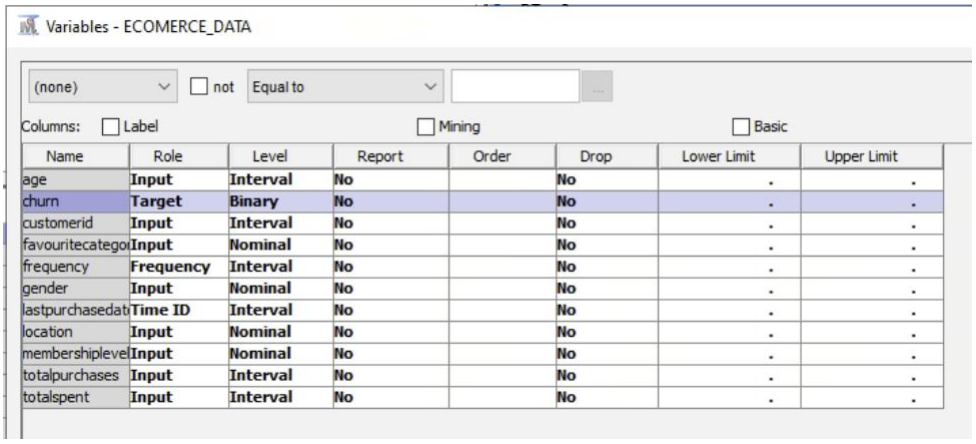
- They are then imported to local file for local SAS EM or SAS EM Online
- Import data to SAS EM Online option



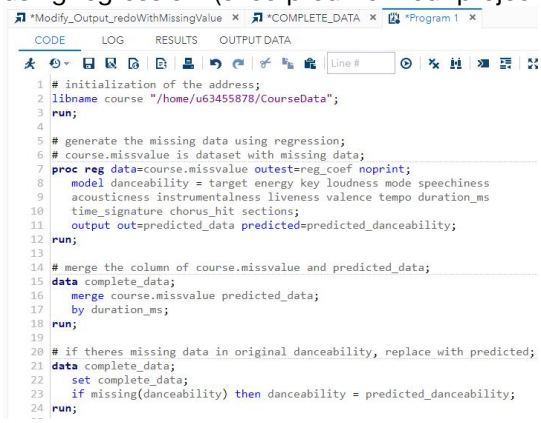
- For local option



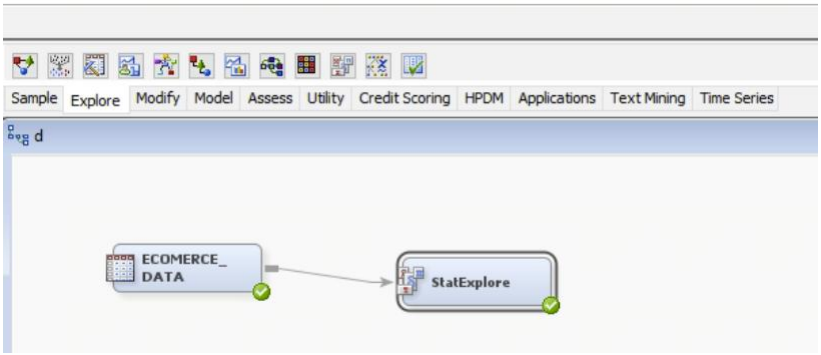
- Variable role edited here



- In our case today, we do not have missing value, if we do have missing value, missing value can be resolved by using regression. (excerpted from our project assignment)



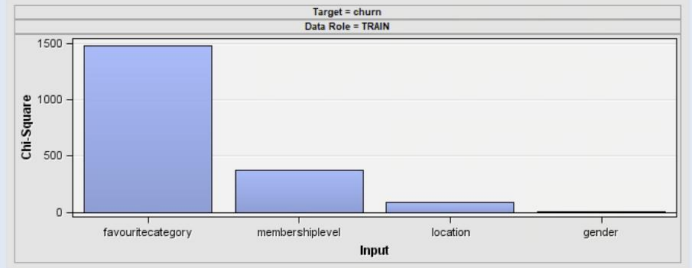
- Furhter insight into customer behaviour and data can be seen below using Explore > Stat Explore



Results - Node: StatExplore Diagram: d

File Edit View Window

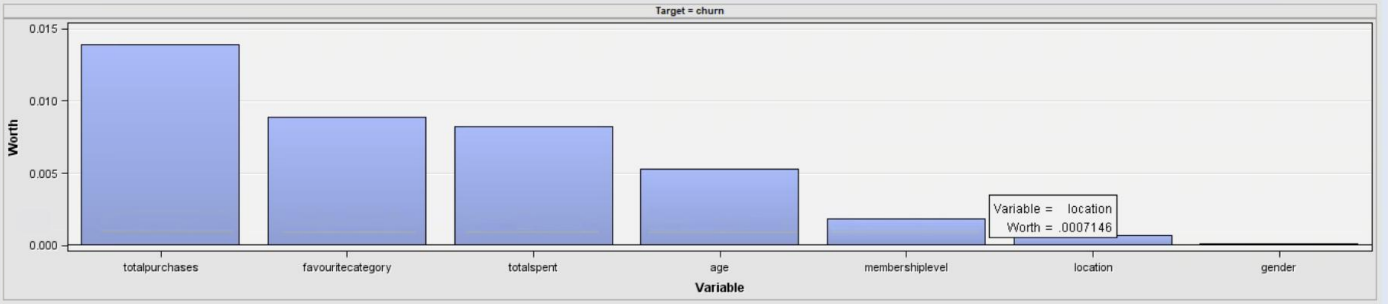
Chi-Square Plot



Output

```
35
36   Class Variable Summary Statistics
37   (maximum 500 observations printed)
38
39   Data Role=TRAIN
40
41                                          Number
42   Data                                     of                                    M
43   Role     Variable Name          Role     Levels    Missing    Mode              Perc
44
45   TRAIN    favouritecategory      INPUT      22          0      Jewelry            5
46   TRAIN    gender                 INPUT       2          0      M                 51
47   TRAIN    location               INPUT       4          0      United States     49
48   TRAIN    membershiplevel        INPUT       3          0      americanexpress   34
49   TRAIN    churn                  TARGET      2          0      1                 50
50
```

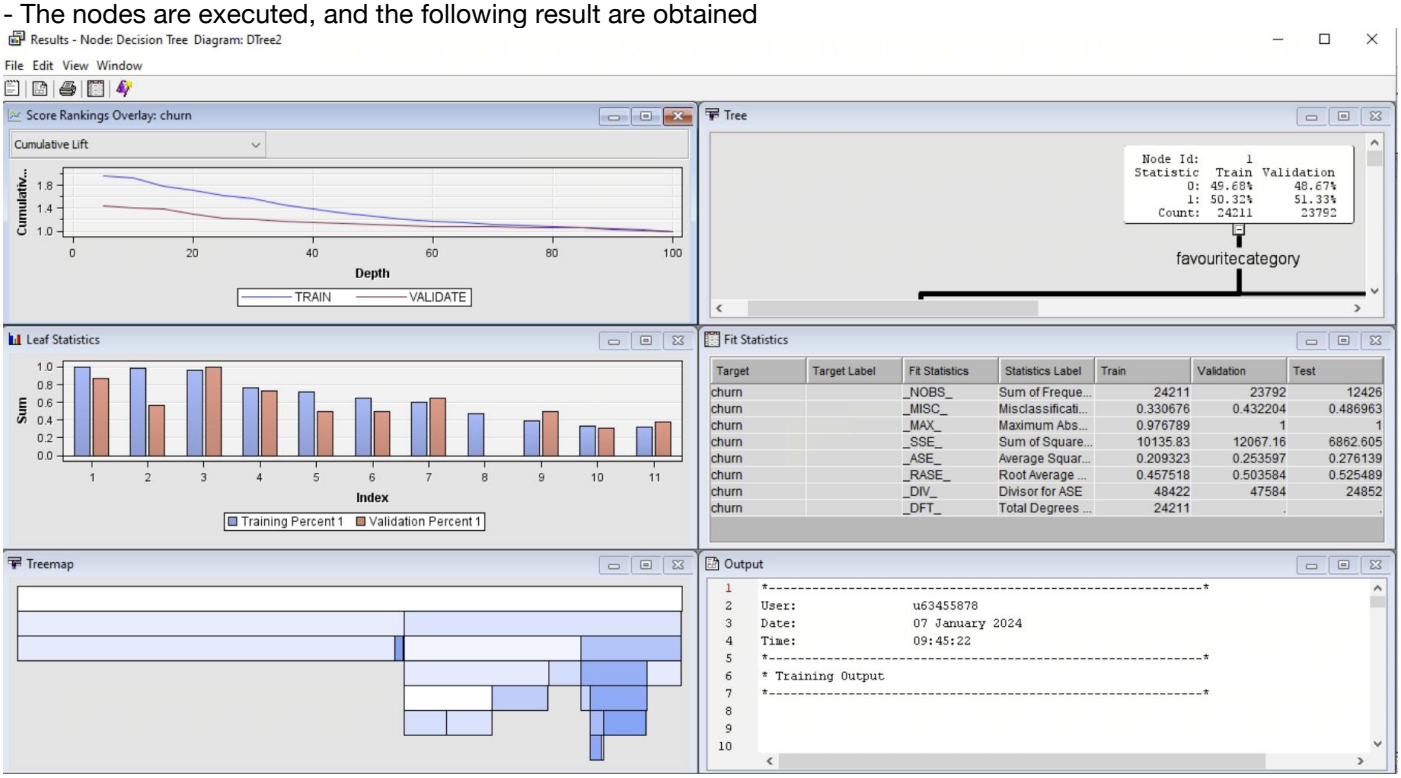
Variable Worth



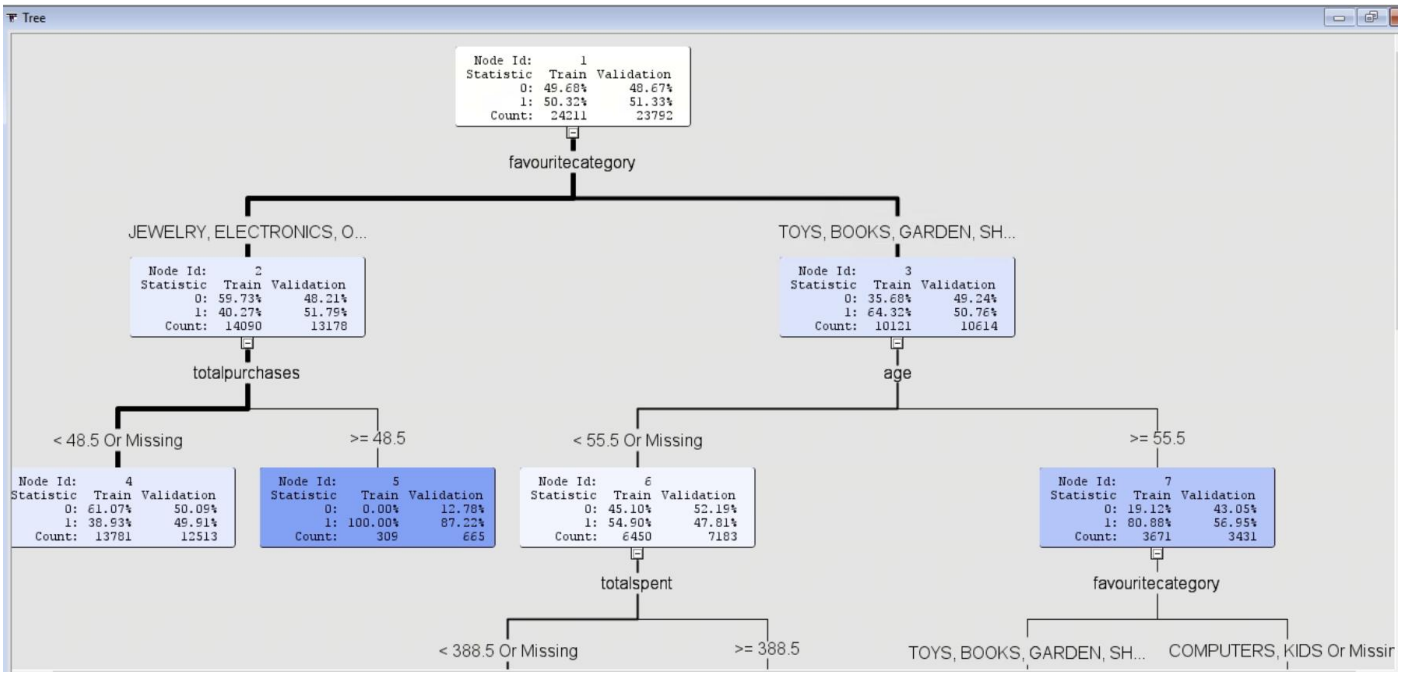
Decision Tree Analysis
Create a decision tree model in SAS Enterprise Miner to analyse customer behaviour.

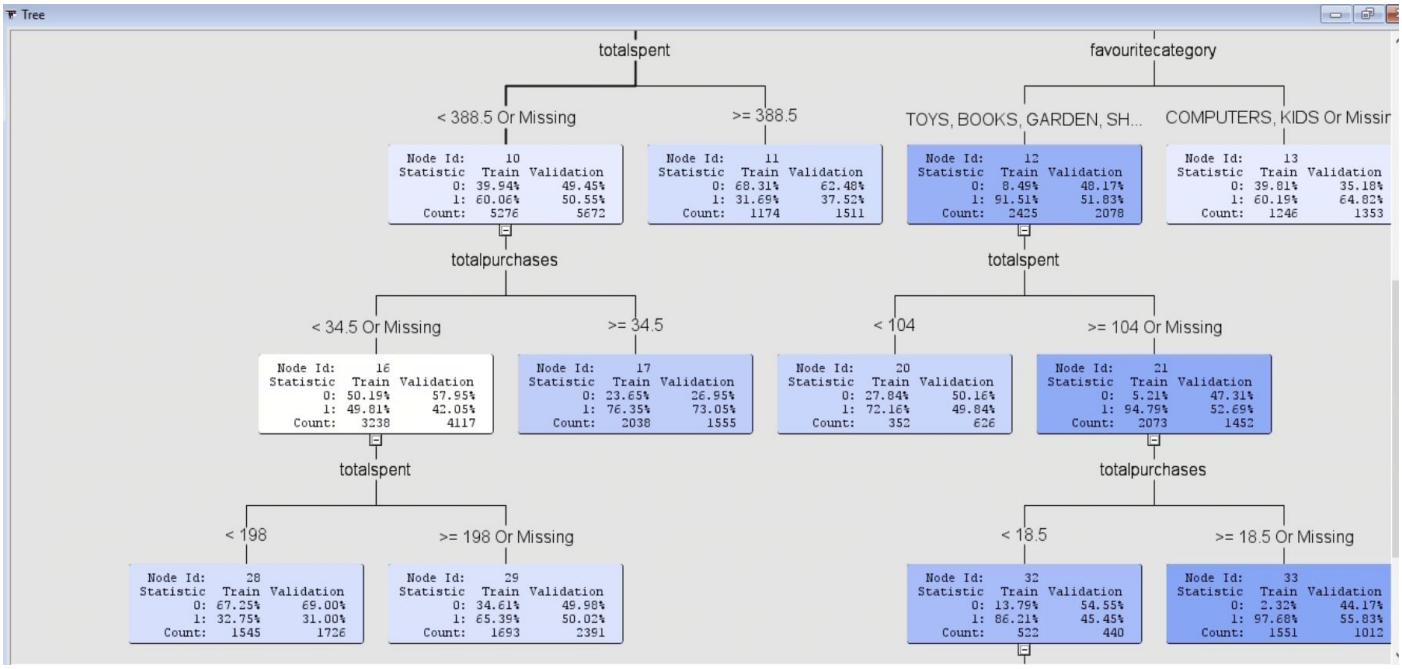
- The following nodes are constructed, and configuration is screenshot if given any.

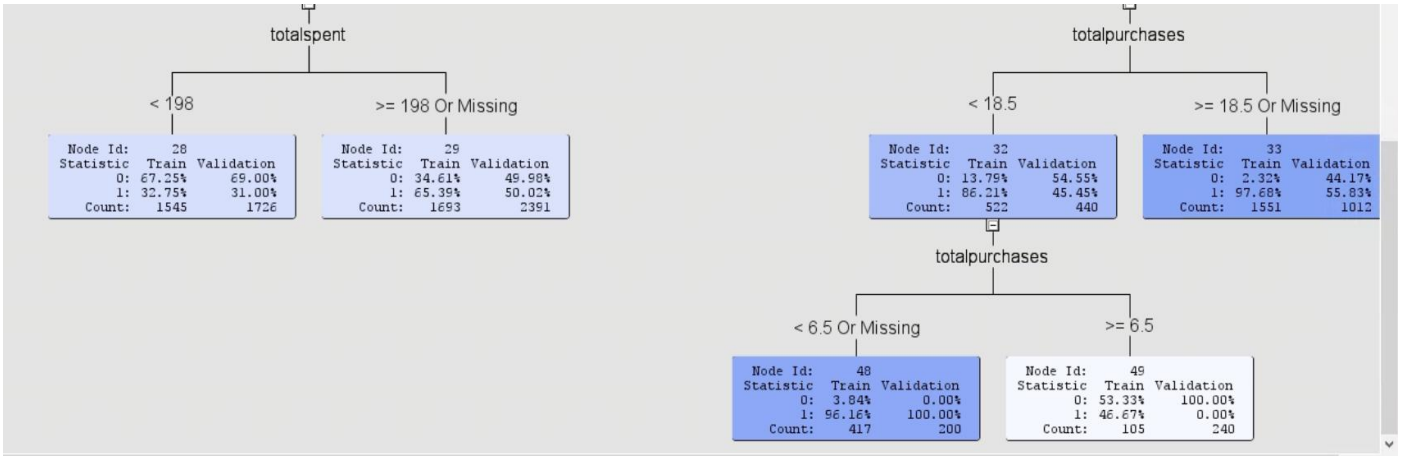
- The nodes are executed, and the following result are obtained



- Tree







--
User: u63455878
Date: 07 January 2024
Time: 09:45:22
--
* Training Output
--

Variable Summary

Role	Measurement Level	Frequency Count
FREQ	INTERVAL	1
ID	INTERVAL	2
INPUT	INTERVAL	3
INPUT	NOMINAL	4
TARGET	BINARY	1
TIMEID	INTERVAL	1

Model Events

Target	Event	Measurement Level	Number of Levels	Order	Label
churn	1	BINARY	2	Descending	

Predicted and decision variables

Type	Variable	Label
TARGET	churn	
PREDICTED	P_churn1	Predicted: churn=1
RESIDUAL	R_churn1	Residual: churn=1
PREDICTED	P_churn0	Predicted: churn=0
RESIDUAL	R_churn0	Residual: churn=0
FROM	F_churn	From: churn
INTO	I_churn	Into: churn
FREQ	frequency	

--
* Score Output
--

--
* Report Output
--

Variable Importance

Variable Name	Label	Number of Splitting Rules	Validation Importance	Ratio of Validation to Training Importance	Importance
favouritecategory		2	1.0000	0.0000	0.0000
totalpurchases		4	0.7331	1.0000	1.3642
totalspent		3	0.6513	0.3390	0.5204
age		1	0.6121	0.0000	0.0000

Tree Leaf Report

Node Id	Depth	Training Observations	Training Percent 1	Validation Observations	Validation Percent 1
4	2	13781	0.39	12513	0.50
17	4	2038	0.76	1555	0.73
29	5	1693	0.65	2391	0.50
33	5	1551	0.98	1012	0.56
28	5	1545	0.33	1726	0.31
13	3	1246	0.60	1353	0.65
11	3	1174	0.32	1511	0.38
48	6	417	0.96	200	1.00
20	4	352	0.72	626	0.50

5	2	309	1.00	665	0.87
49	6	105	0.47	240	0.00

Fit Statistics

Target=churn Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	24211.00	23792.00	12426.00
MISC	Misclassification Rate	0.33	0.43	0.49
MAX	Maximum Absolute Error	0.98	1.00	1.00
SSE	Sum of Squared Errors	10135.83	12067.16	6862.60
ASE	Average Squared Error	0.21	0.25	0.28
RASE	Root Average Squared Error	0.46	0.50	0.53
DIV	Divisor for ASE	48422.00	47584.00	24852.00
DFT	Total Degrees of Freedom	24211.00	.	.

Classification Table

Data Role=TRAIN Target Variable=churn Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	62.1078	85.7487	10313	42.5963
1	0	37.8922	51.6415	6292	25.9882
0	1	22.5348	14.2513	1714	7.0794
1	1	77.4652	48.3585	5892	24.3360

Data Role=VALIDATE Target Variable=churn Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	54.0525	74.6438	8643	36.3273
1	0	45.9475	60.1572	7347	30.8801
0	1	37.6314	25.3562	2936	12.3403
1	1	62.3686	39.8428	4866	20.4523

Event Classification Table

Data Role=TRAIN Target=churn Target Label=' '

False Negative	True Negative	False Positive	True Positive
6292	10313	1714	5892

Data Role=VALIDATE Target=churn Target Label=' '

False	True	False	True

Negative	Negative	Positive	Positive
7347	8643	2936	4866

Assessment Score Rankings

Data Role=TRAIN Target Variable=churn Target Label=' '

Depth	Gain	Cumulative Lift	% Lift	Response	Cumulative % Response	Mean Number of Observations	Posterior Probability
5	95.2760	1.95276	1.95276	98.2712	98.2712	1211	0.98271
10	91.6315	1.87987	1.91632	94.6030	96.4371	1211	0.94603
15	78.3333	1.51715	1.78333	76.3494	89.7449	1210	0.76349
20	70.2805	1.46129	1.70281	73.5382	85.6924	1211	0.73538
25	62.2146	1.29931	1.62215	65.3869	81.6333	1210	0.65387
30	55.5481	1.22226	1.55548	61.5094	78.2784	1211	0.61509
35	46.2792	0.90635	1.46279	45.6114	73.6139	1210	0.45611
40	37.6615	0.77359	1.37661	38.9304	69.2771	1211	0.38930
45	30.9643	0.77359	1.30964	38.9304	65.9068	1210	0.38930
50	25.6020	0.77359	1.25602	38.9304	63.2083	1211	0.38930
55	21.2150	0.77359	1.21215	38.9304	61.0005	1211	0.38930
60	17.5621	0.77359	1.17562	38.9304	59.1622	1210	0.38930
65	14.4686	0.77359	1.14469	38.9304	57.6054	1211	0.38930
70	11.8192	0.77359	1.11819	38.9304	56.2721	1210	0.38930
75	9.5211	0.77359	1.09521	38.9304	55.1156	1211	0.38930
80	7.5119	0.77359	1.07512	38.9304	54.1045	1210	0.38930
85	5.7376	0.77359	1.05738	38.9304	53.2116	1211	0.38930
90	3.9938	0.74335	1.03994	37.4085	52.3341	1210	0.37408
95	1.9450	0.65080	1.01945	32.7508	51.3030	1211	0.32751
100	0.0000	0.63028	1.00000	31.7182	50.3242	1210	0.31718

Data Role=VALIDATE Target Variable=churn Target Label=' '

Depth	Gain	Cumulative Lift	% Lift	Response	Cumulative % Response	Mean Number of Observations	Posterior Probability
5	42.9320	1.42932	1.42932	73.3704	73.3704	1190	0.98976
10	40.1694	1.37407	1.40169	70.5342	71.9523	1190	0.88408
15	39.1489	1.37106	1.39149	70.3798	71.4284	1189	0.75867
20	28.6846	0.97301	1.28685	49.9467	66.0569	1190	0.68170
25	22.4399	0.97445	1.22440	50.0209	62.8513	1189	0.65387
30	21.0555	1.14136	1.21056	58.5888	62.1407	1190	0.62379
35	17.1645	0.93825	1.17165	48.1626	60.1433	1190	0.52355
40	14.6734	0.97225	1.14673	49.9081	58.8646	1189	0.38930
45	12.7342	0.97225	1.12734	49.9081	57.8692	1190	0.38930
50	11.1841	0.97225	1.11184	49.9081	57.0735	1189	0.38930
55	9.9148	0.97225	1.09915	49.9081	56.4219	1190	0.38930
60	8.8570	0.97225	1.08857	49.9081	55.8789	1190	0.38930
65	7.9627	0.97225	1.07963	49.9081	55.4198	1189	0.38930
70	7.1955	0.97225	1.07196	49.9081	55.0260	1190	0.38930
75	6.5312	0.97225	1.06531	49.9081	54.6850	1189	0.38930
80	5.9494	0.97225	1.05949	49.9081	54.3864	1190	0.38930
85	5.4361	0.97225	1.05436	49.9081	54.1228	1190	0.38930
90	3.5040	0.70640	1.03504	36.2612	53.1310	1189	0.34471
95	1.4150	0.63825	1.01415	32.7630	52.0587	1190	0.32463
100	0.0000	0.73102	1.00000	37.5248	51.3324	1189	0.31687

Assessment Score Distribution

Data Role=TRAIN Target Variable=churn Target Label=' '

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	2225	52	0.97716	9.4048
0.75-0.80	1556	482	0.76349	8.4177
0.70-0.75	254	98	0.72159	1.4539
0.65-0.70	1107	586	0.65387	6.9927
0.60-0.65	750	496	0.60193	5.1464
0.45-0.50	49	56	0.46667	0.4337
0.35-0.40	5365	8416	0.38930	56.9204
0.30-0.35	878	1841	0.32291	11.2304

Data Role=VALIDATE Target Variable=churn Target Label=' '

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	1345	532	0.98340	7.8892
0.75-0.80	1136	419	0.76349	6.5358
0.70-0.75	312	314	0.72159	2.6311
0.65-0.70	1196	1195	0.65387	10.0496
0.60-0.65	877	476	0.60193	5.6868
0.45-0.50	0	240	0.46667	1.0087
0.35-0.40	6245	6268	0.38930	52.5933
0.30-0.35	1102	2135	0.32254	13.6054

From the result output, we can deduce that the importance for:
favouritecategory:
Training: Importance ratio is 1.0000
Validation: not provided
totalpurchases:
Training: Importance ratio is 0.7331
Validation: Importance ratio is 1.0000
totalspent:
Training: Importance ratio is 0.6513
Validation: Importance ratio is 0.3390
age:
Training: Importance ratio is 0.6121
Validation: not provided.

Fit Statistics:
Misclassification Rate:
Training: 0.33
Validation: 0.43
Maximum Absolute Error:
Training: 0.98
Validation: 1.00
Sum of Squared Errors (SSE):
Training: 10135.83
Validation: 12067.16

Generally we can observe that the:
- The model seems to perform slightly better on the training dataset than on the validation dataset, as indicated by lower misclassification rates and higher gain values.

- Variable importance suggests that "favouritecategory" is the most important feature, followed by "totalpurchases," "totalspent," and "age."

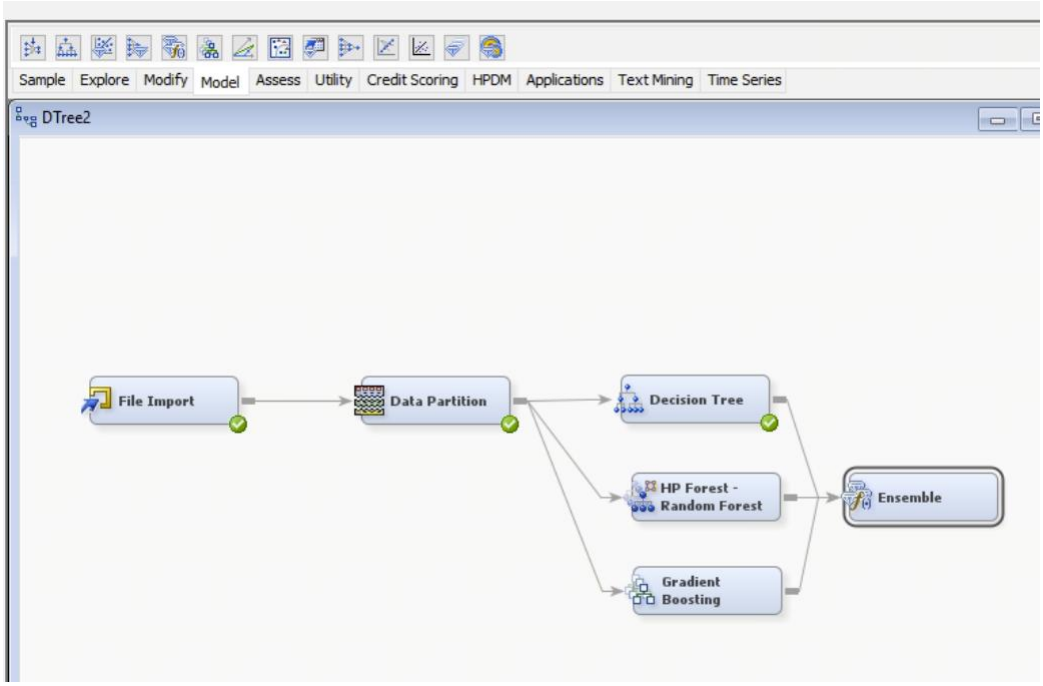
From the report, we know that:
> favouritecategory has been identified as the most important variable.
> Therefore, we can see that customers' preferred categories significantly influence their likelihood of churn.

These are the steps we can do, to improve business and retention, based on the information we know so far:

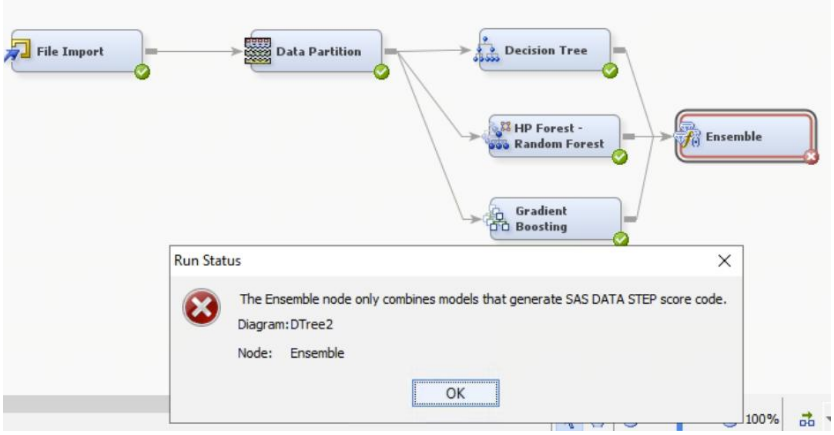
1. Customer Engagement: Leverage insights from "favouritecategory" to tailor marketing strategies and promotions based on customers' preferred product categories. Personalized campaigns can greatly enhance customer engagement.

2. Retention Programs:
Implement targeted retention programs for customers with high totalpurchases and totalspent. We can tailor communication and services to serve this segment's preferences and needs.

Ensemble Methods
Apply Bagging and Boosting, using the Random Forest algorithm as a Bagging example.



- In ensemble, two or more predictive models combined to create a potentially more accurate model
- Works better when model predictions are uncorrelated
- There are 3 ways to ensemble modelling, they are namely Averaging(voting)/ Stacking(Blending)/ Cluster-based selection
- In the diagram below, we can see it is a machine learning technique that involves combining the predictions of multiple models to produce a stronger, more robust model.



Challenge faced:
- Unable to proceed with Ensemble although multiple effort. However, the concept is clear and well known to increase accuracy and robustness to unseen data.
- The model might not be accurate due to the data is randomly generated from a website (does not reflect real customer behaviour)
- Therefore, there might be unrelavent relationship between them.