ALTERNATIVE ASSESSMENT 1 (50 marks) - WEEK 12

Answer the question below based on the given scenario. Submit your answer within ONE (1) DAY after the question is given in SPECTRUM. Answers should be submitted and saved with the student's name followed by matric number as the file name in the format of .pdf (e.g.Ali_s123456.pdf).

Case Study: E-Commerce Customer Behaviour Analysis

Background:
You will work with a dataset of customer transactions from an e-commerce website, encompassing various customer attributes and purchase history over the last year. The structure provided below is a guideline. Feel free to enhance this dataset by adding relevant attributes that you believe will enrich your analysis. Use the structure as a foundation to create your own sample dataset that reflects realistic customer behaviour.

Dataset Structure:

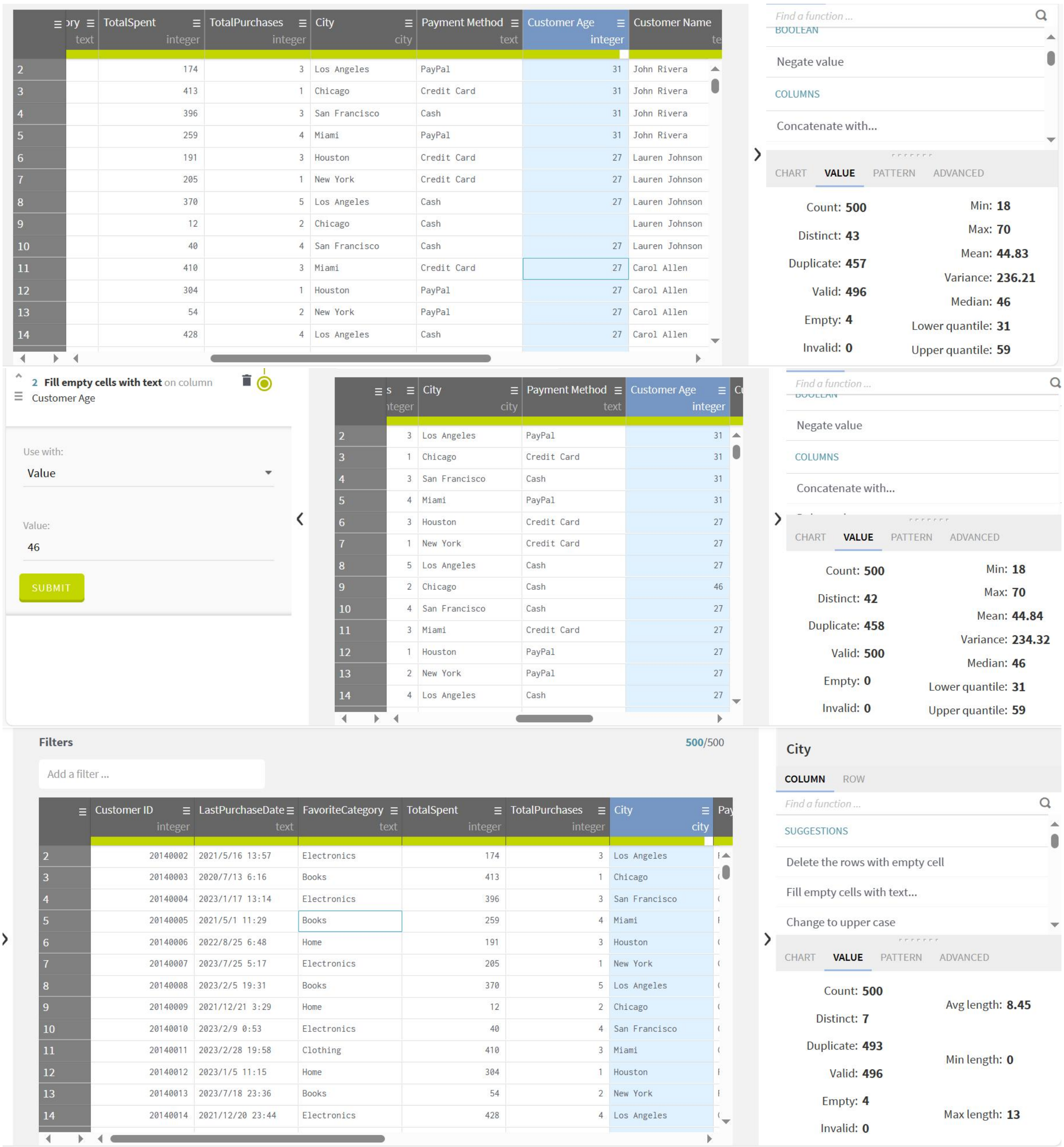
CustomerID: Unique identifier for each customer.
Age: Age of the customer.
Gender: Gender of the customer.
Location: Geographic location of the customer.
MembershipLevel: Indicates the membership level (e.g., Bronze, Silver, Gold, Platinum).
TotalPurchases: Total number of purchases made by the customer.
TotalSpent: Total amount spent by the customer.
FavoriteCategory: The category in which the customer most frequently shops (e.g., Electronics, Clothing, Home Goods).
LastPurchaseDate: The date of the last purchase.
[Additional Attributes]: Consider adding more attributes like customer's occupation, frequency of website visits, etc.
Churn: Indicates whether the customer has stopped purchasing (1 for churned, 0 for active).

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

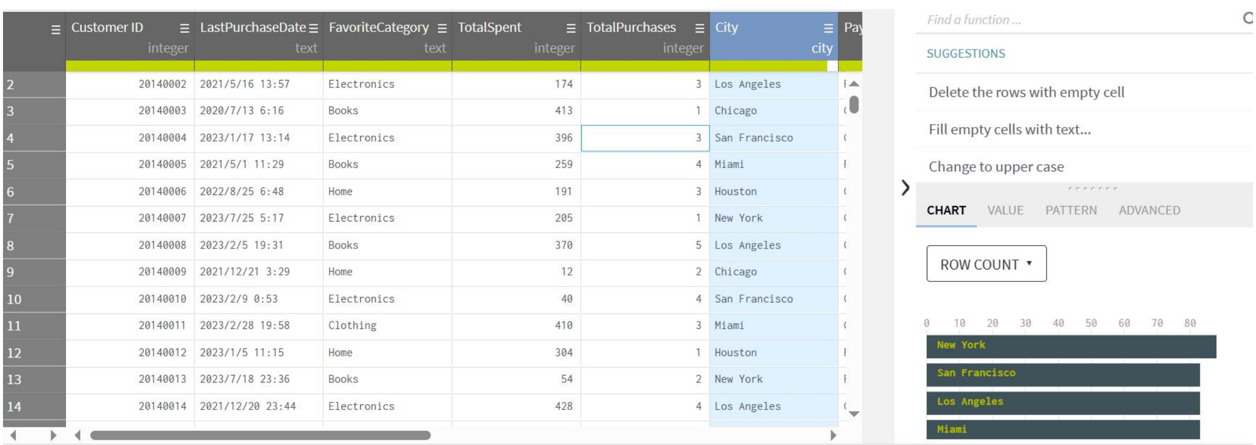
[15 marks]

To see if there are any missing values in each column of the data, here are the columns that are sifted out with missing values:

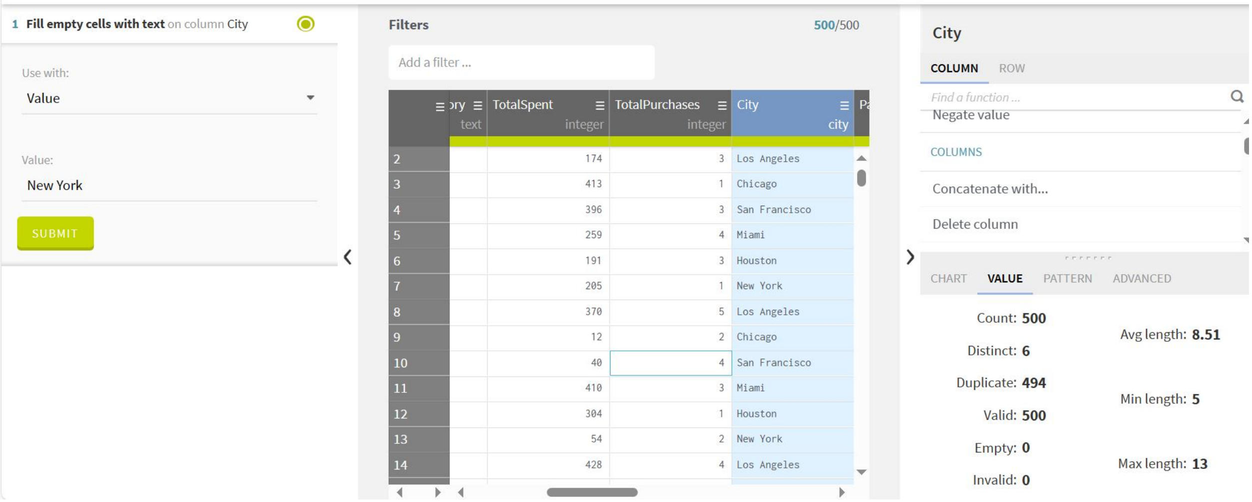
For missing age values, this experiment uses the mean to fill in. Using the mean to fill in missing values helps to maintain the original distribution of the data.



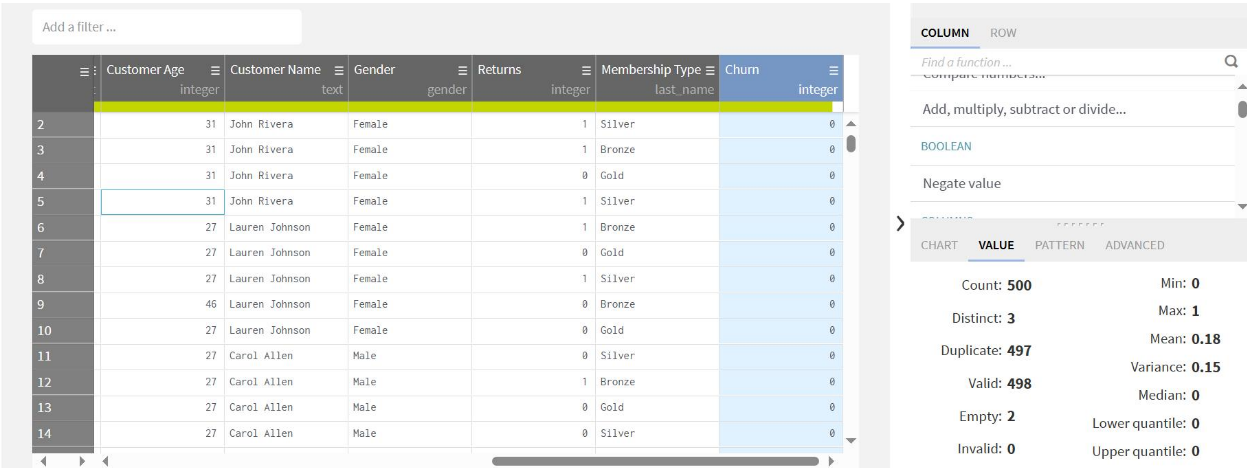
The city names with the highest frequency of occurrence usually represent the major cities in the dataset, and populating such city names helps to maintain the consistency of the overall data distribution. For missing addresses, this experiment uses taking the city name with the highest frequency of occurrence to fill it.

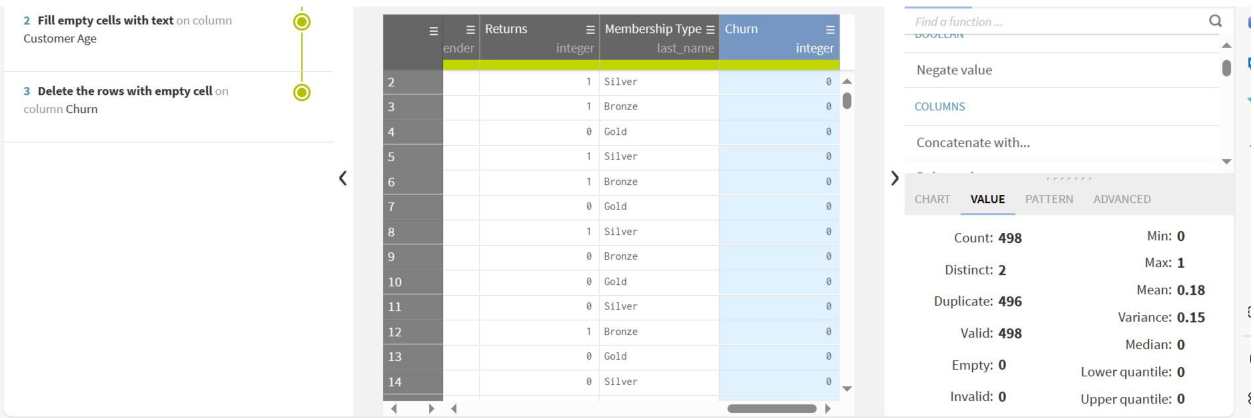


As shown, New York is the most frequent. So fill in the four missing city names as New York.

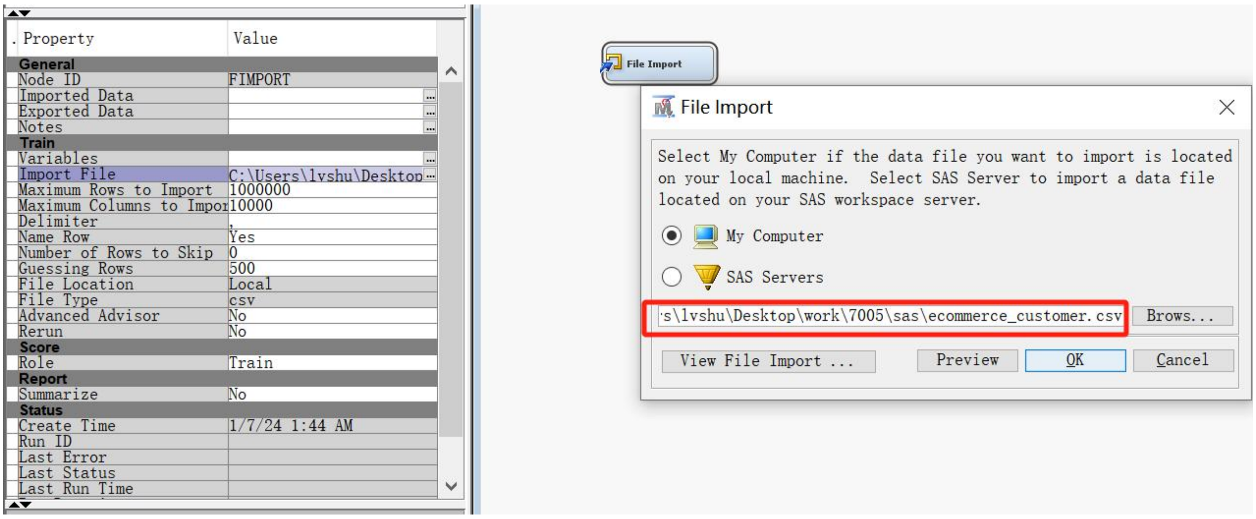


For critical attributes such as missing churn value, where there are few missing values, the deletion of missing values is used to deal with it.Churn value is usually a critical attribute in user churn prediction as it directly reflects whether a user is churned or not. In this case, it is important to ensure the accuracy of this attribute as it is the target variable for model training. Removing missing values avoids introducing uncertainty about the accuracy of user churn prediction during the modeling process.

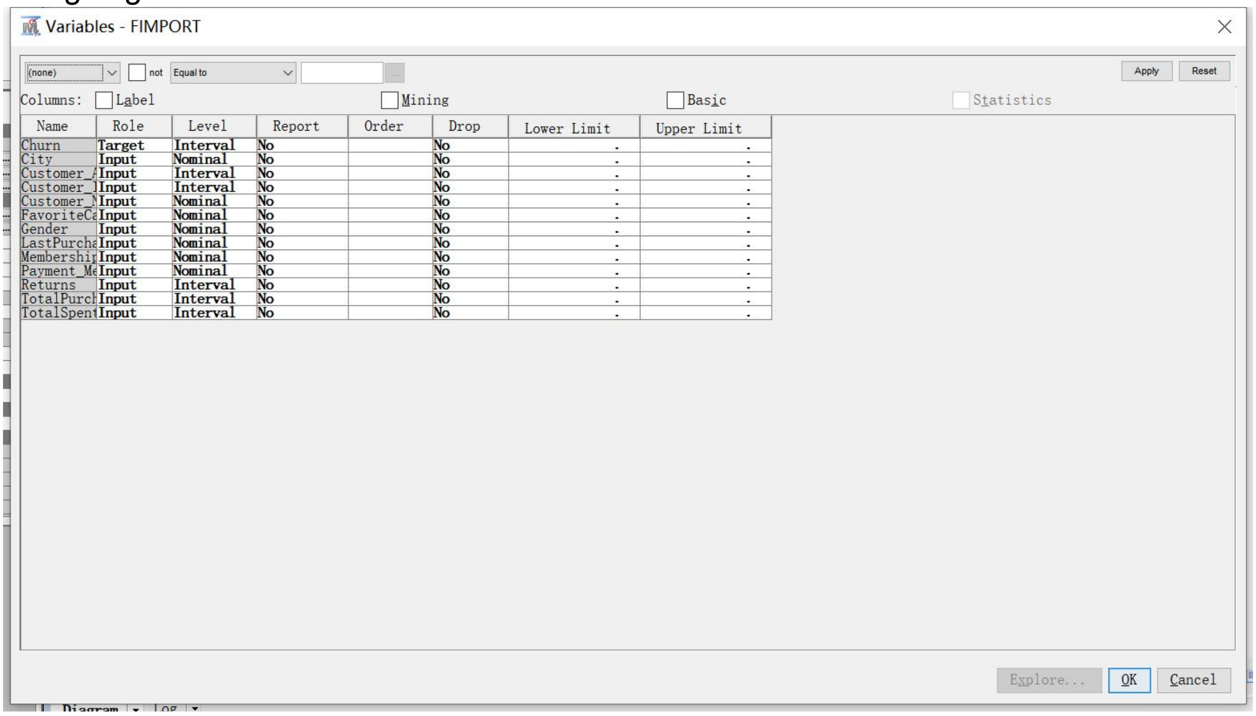




Import data:



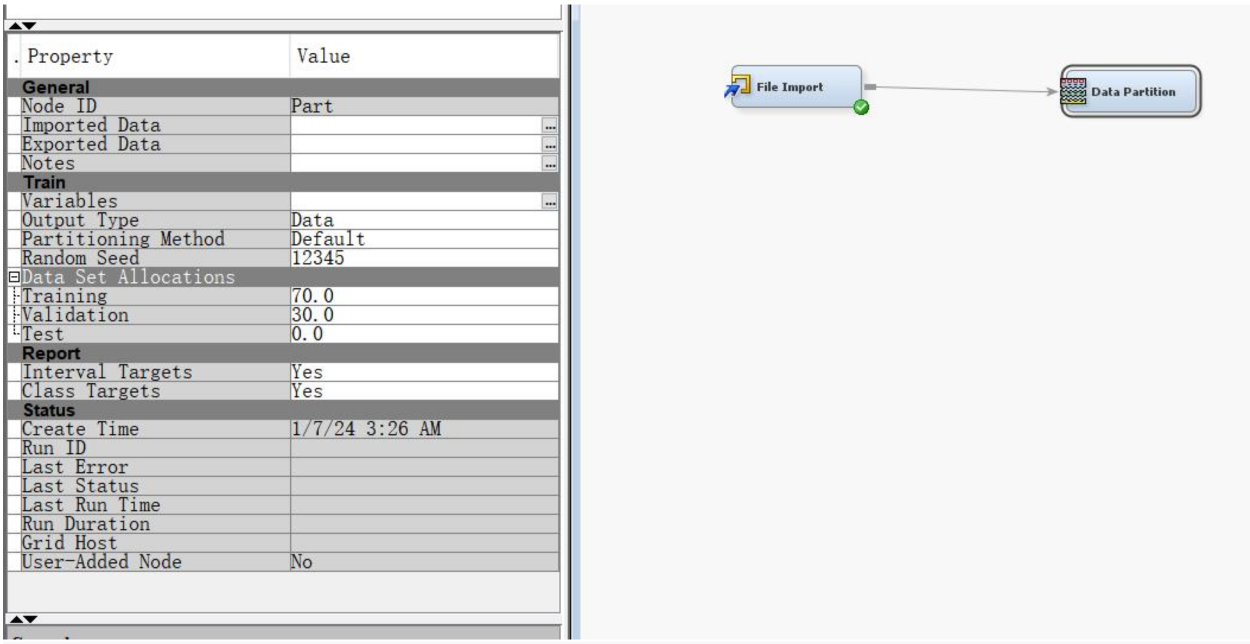
Assigning variable roles:



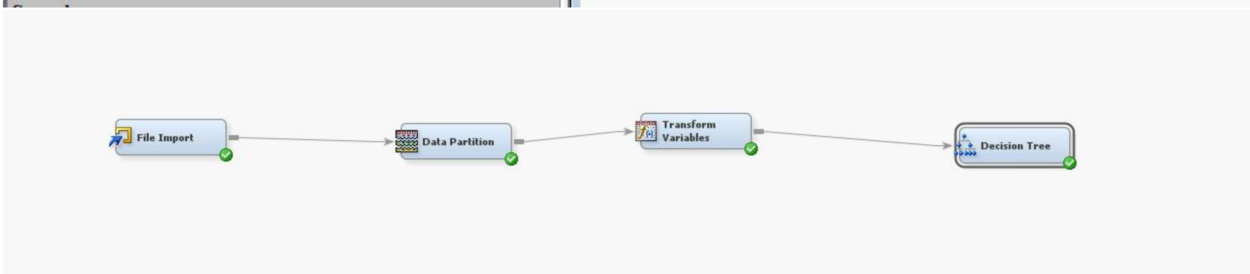
Decision Tree Analysis: Create a decision tree model in SAS Enterprise Miner to analyse customer behaviour.

[20 marks]

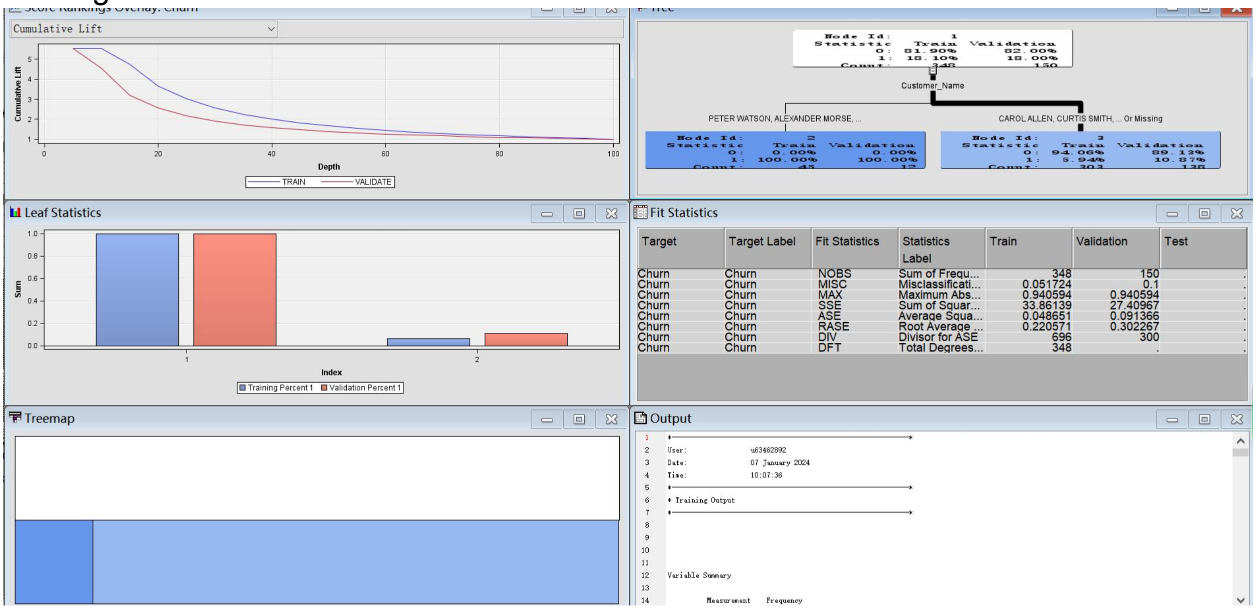
Setting the ratio of training set to testing set:



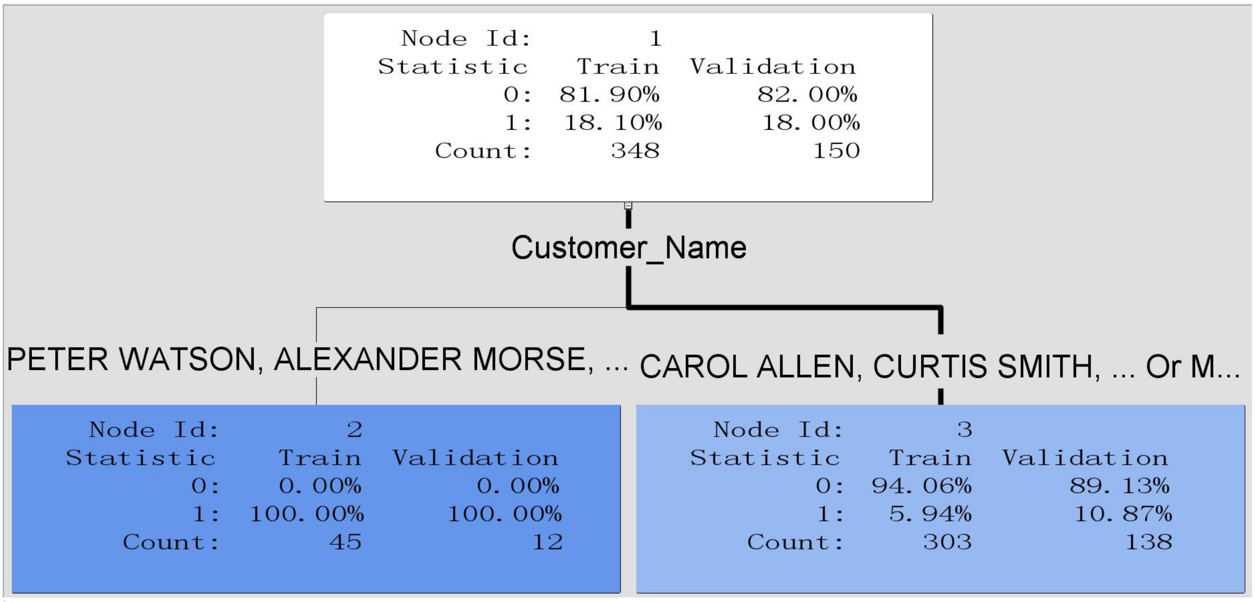
Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	1/7/24 3:26 AM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No



running result:



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	NOBS	Sum of Frequ...	348	150	.
Churn	Churn	MISC	Misclassificati...	0.051724	0.1	.
Churn	Churn	MAX	Maximum Abs...	0.940594	0.940594	.
Churn	Churn	SSE	Sum of Squar...	33.86139	27.40967	.
Churn	Churn	ASE	Average Squa...	0.048651	0.091366	.
Churn	Churn	RASE	Root Average ...	0.220571	0.302267	.
Churn	Churn	DIV	Divisor for ASE	696	300	.
Churn	Churn	DFT	Total Degrees...	348	.	.



Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Customer_Name		1	1.0000	1.0000	1.0000

Tree Leaf Report

Node Id	Depth	Training Observations	Training Percent 1
3	1	303	0.06
2	1	45	1.00

Fit Statistics

Target=Churn Target Label=Churn

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	348.000	150.000
MISC	Misclassification Rate	0.052	0.100
MAX	Maximum Absolute Error	0.941	0.941
SSE	Sum of Squared Errors	33.861	27.410
ASE	Average Squared Error	0.049	0.091
RASE	Root Average Squared Error	0.221	0.302
DIV	Divisor for ASE	696.000	300.000
DFT	Total Degrees of Freedom	348.000	.

Classification Table

Data Role=TRAIN Target Variable=Churn Target Label=Churn

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	94.059	100.000	285	81.8966
1	0	5.941	28.571	18	5.1724
1	1	100.000	71.429	45	12.9310

Data Role=VALIDATE Target Variable=Churn Target Label=Churn

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	89.130	100.000	123	82
1	0	10.870	55.556	15	10
1	1	100.000	44.444	12	8

Event Classification Table

Data Role=TRAIN Target=Churn Target Label=Churn

False Negative	True Negative	False Positive	True Positive
18	285	0	45

Data Role=VALIDATE Target=Churn Target Label=Churn

False Negative	True Negative	False Positive	True Positive
15	123	0	12

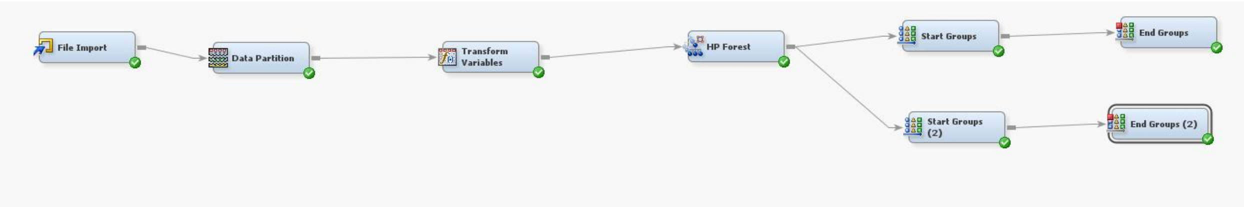
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	NOBS	Sum of Frequencies	348	150	.
Churn	Churn	MISC	Misclassification Rate	0.051724	0.1	.
Churn	Churn	MAX	Maximum Absolute Error	0.940594	0.940594	.
Churn	Churn	SSE	Sum of Squared Errors	33.86139	27.40967	.
Churn	Churn	ASE	Average Squared Error	0.048651	0.091366	.
Churn	Churn	RASE	Root Average Squared Error	0.220571	0.302267	.
Churn	Churn	DIV	Divisor for ASE	696	300	.
Churn	Churn	DFT	Total Degrees of Freedom	348	.	.

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

[10 marks]

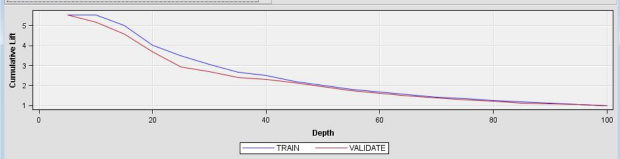
Bagging:

Property	Value
General	
Node ID	Grp
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Rerun	No
General	
Mode	Bagging
Target Group	No
Index Count	10
Minimum Group Size	10
Bagging	
Type	Percentage
Observations	.
Percentage	10.0
Random Seed	12345
Status	
Create Time	1/7/24 8:21 AM
Run ID	837cb136-3825-aa46-baa3-3
Last Error	
Last Status	Complete
Last Run Time	1/7/24 8:23 AM
Run Duration	0 Hr. 0 Min. 2.80 Sec.
Grid Host	
User-Added Node	No



Score Rankings Overlay: Churn

Cumulative Lift



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	ASE	Average Squa...	0.115409	0.126529	.
Churn	Churn	DIV	Divisor for ASE	696	300	.
Churn	Churn	MAX	Maximum Abs...	0.858563	0.958457	.
Churn	Churn	NOBS	Sum of Frequ...	348	150	.
Churn	Churn	RASE	Root Average ...	0.339719	0.355709	.
Churn	Churn	SSE	Sum of Squar...	80.32488	37.95857	.
Churn	Churn	DISF	Frequency of ...	348	150	.
Churn	Churn	MISC	Misclassificati...	0.181034	0.18	.
Churn	Churn	WRONG	Number of Wr...	63	27	.

Summary

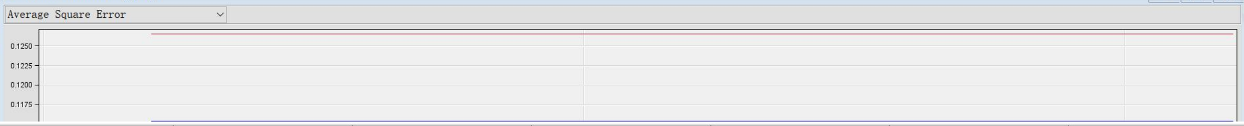
Mode	Group Index	Target
Bagging	10	Churn

Output

```
*------------------------------------------------------------*
User:               u63462892
Date:               07 January 2024
Time:               09:37:33
*------------------------------------------------------------*
* Post Grouping Output
*------------------------------------------------------------*



Model Events

                                   Number
```

Statistics Plot - Bagging

Average Square Error



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	ASE	Average Squared Error	0.115409	0.126529	
Churn	Churn	DIV	Divisor for ASE	696	300	
Churn	Churn	MAX	Maximum Absolute Error	0.858563	0.958457	
Churn	Churn	NOBS	Sum of Frequencies	348	150	
Churn	Churn	RASE	Root Average Squared Error	0.339719	0.355709	
Churn	Churn	SSE	Sum of Squared Errors	80.32488	37.95857	
Churn	Churn	DISF	Frequency of Classified Cases	348	150	
Churn	Churn	MISC	Misclassification Rate	0.181034	0.18	
Churn	Churn	WRONG	Number of Wrong Classifications	63	27	

15	400.270	3.98942	5.00270	72.222	90.566	18	0.19351
20	302.449	0.97479	4.02449	17.647	72.857	17	0.18325
25	249.206	1.29972	3.49206	23.529	63.218	17	0.17096
30	205.125	0.92063	3.05125	16.667	55.238	18	0.15818
35	167.135	0.32493	2.67135	5.882	48.361	17	0.15009
40	148.571	1.22751	2.48571	22.222	45.000	18	0.14376
45	121.656	0.00000	2.21656	0.000	40.127	17	0.13616
50	100.000	0.00000	2.00000	0.000	36.207	17	0.13175
55	81.250	0.00000	1.81250	0.000	32.813	18	0.12794
60	66.507	0.00000	1.66507	0.000	30.144	17	0.12191
65	53.304	0.00000	1.53304	0.000	27.753	18	0.10697
70	42.623	0.00000	1.42623	0.000	25.820	17	0.09689
75	33.333	0.00000	1.33333	0.000	24.138	17	0.07776
80	24.731	0.00000	1.24731	0.000	22.581	18	0.05924
85	17.568	0.00000	1.17568	0.000	21.284	17	0.04946
90	10.828	0.00000	1.10828	0.000	20.064	18	0.04453
95	5.136	0.00000	1.05136	0.000	19.033	17	0.03963
100	0.000	0.00000	1.00000	0.000	18.103	17	0.03486

Data Role=VALIDATE Target Variable=Churn Target Label=Churn

Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	455.556	5.55556	5.55556	100.000	100.000	8	0.28649
10	418.519	4.76190	5.18519	85.714	93.333	7	0.20207
15	358.937	3.47222	4.58937	62.500	82.609	8	0.18704
20	270.370	0.79365	3.70370	14.286	66.667	7	0.18158
25	192.398	0.00000	2.92398	0.000	52.632	8	0.17433
30	171.605	1.58730	2.71605	28.571	48.889	7	0.16877
35	141.090	0.69444	2.41090	12.500	43.396	8	0.15511
40	131.481	1.58730	2.31481	28.571	41.667	7	0.15056
45	112.418	0.69444	2.12418	12.500	38.235	8	0.14311
50	92.593	0.00000	1.92593	0.000	34.667	7	0.13316
55	74.029	0.00000	1.74029	0.000	31.325	8	0.12938
60	60.494	0.00000	1.60494	0.000	28.889	7	0.12568
65	47.392	0.00000	1.47392	0.000	26.531	8	0.11715
70	37.566	0.00000	1.37566	0.000	24.762	7	0.10531
75	27.827	0.00000	1.27827	0.000	23.009	8	0.09628
80	20.370	0.00000	1.20370	0.000	21.667	7	0.07481
85	12.847	0.00000	1.12847	0.000	20.313	8	0.05775
90	6.996	0.00000	1.06996	0.000	19.259	7	0.05032
95	4.895	0.69444	1.04895	12.500	18.881	8	0.04257
100	0.000	0.00000	1.00000	0.000	18.000	7	0.03540

Assessment Score Distribution

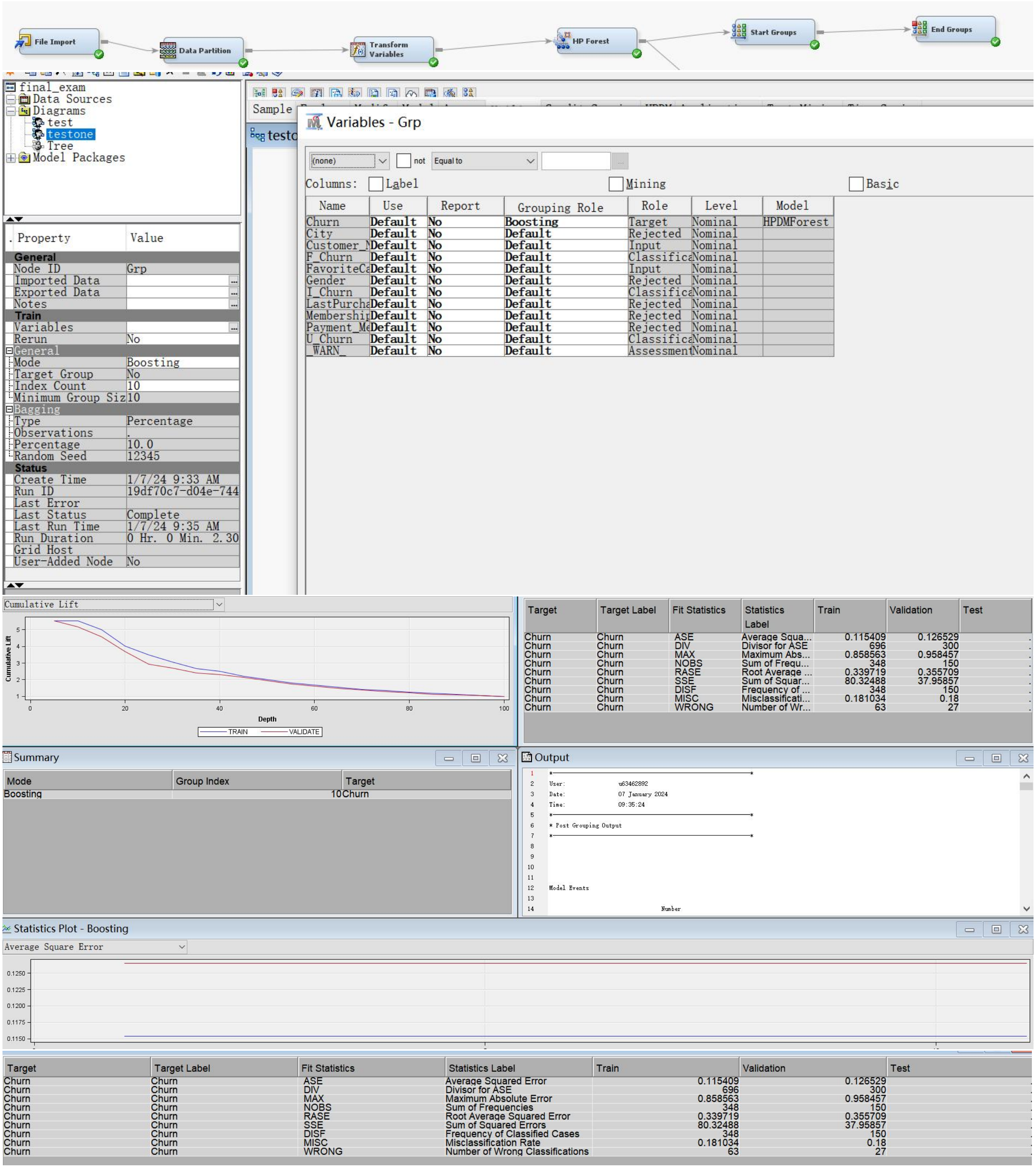
Data Role=TRAIN Target Variable=Churn Target Label=Churn

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.45-0.50	7	0	0.45549	2.0115
0.30-0.35	6	0	0.32066	1.7241
0.25-0.30	11	0	0.26303	3.1609
0.20-0.25	16	0	0.22085	4.5977
0.15-0.20	19	53	0.17189	20.6897
0.10-0.15	4	113	0.12937	33.6207
0.05-0.10	0	57	0.07366	16.3793
0.00-0.05	0	62	0.04102	17.8161

Data Role=VALIDATE Target Variable=Churn Target Label=Churn

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.45-0.50	1	0	0.45150	0.6667
0.30-0.35	2	0	0.31814	1.3333
0.25-0.30	2	0	0.26220	1.3333
0.20-0.25	6	0	0.21933	4.0000
0.15-0.20	13	34	0.17276	31.3333
0.10-0.15	2	47	0.12581	32.6667
0.05-0.10	0	26	0.06969	17.3333
0.00-0.05	1	16	0.04014	11.3333

Boosting:


File Import
Data Partition
Transform Variables
HP Forest
Start Groups
End Groups
final_exam
Data Sources
Diagrams
test
testone
Tree
Model Packages
Property
Value
General
Node ID
Grp
Imported Data
Exported Data
Notes
Train
Variables
Rerun
No
General
Mode
Boosting
Target Group
No
Index Count
10
Minimum Group Siz
10
Bagging
Type
Percentage
Observations
Percentage
10.0
Random Seed
12345
Status
Create Time
1/7/24 9:33 AM
Run ID
19df70c7-d04e-744
Last Error
Last Status
Complete
Last Run Time
1/7/24 9:35 AM
Run Duration
0 Hr. 0 Min. 2.30
Grid Host
User-Added Node
No
Variables - Grp
(none)
not
Equal to
Columns:
Label
Mining
Basic
Name
Use
Report
Grouping Role
Role
Level
Model
Churn Default No Boosting Target Nominal HPDMForest
City Default No Default Rejected Nominal
Customer_N Default No Default Input Nominal
F_Churn Default No Default Classifica Nominal
FavoriteCa Default No Default Input Nominal
Gender Default No Default Rejected Nominal
I_Churn Default No Default Classifica Nominal
LastPurcha Default No Default Rejected Nominal
Membershi Default No Default Rejected Nominal
Payment_M Default No Default Rejected Nominal
U_Churn Default No Default Classifica Nominal
WARN Default No Default Assessment Nominal
Cumulative Lift
Depth
TRAIN
VALIDATE
Summary
Mode
Group Index
Target
Boosting
10
Churn
Output
User:
Date:
07 January 2024
Time:
09:35:24
* Post Grouping Output
Model Events
Number
Statistics Plot - Boosting
Average Square Error
Target
Target Label
Fit Statistics
Statistics Label
Train
Validation
Test
Churn Churn ASE Average Squared Error 0.115409 0.126529
Churn Churn DIV Divisor for ASE 696 300
Churn Churn MAX Maximum Absolute Error 0.858563 0.958457
Churn Churn NOBS Sum of Frequencies 348 150
Churn Churn RASE Root Average Squared Error 0.339719 0.355709
Churn Churn SSE Sum of Squared Errors 80.32488 37.95857
Churn Churn DISF Frequency of Classified Cases 348 150
Churn Churn MISC Misclassification Rate 0.181034 0.18
Churn Churn WRONG Number of Wrong Classifications 63 27

Predicted and decision variables

Type	Variable	Label
TARGET	Churn	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

Group Summary

Mode	Group Index	Target
Boosting	10	Churn

Fit Statistics

Target=Churn Target Label=Churn

Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.115	0.127
DIV	Divisor for ASE	696.000	300.000
MAX	Maximum Absolute Error	0.859	0.958
NOBS	Sum of Frequencies	348.000	150.000
RASE	Root Average Squared Error	0.340	0.356
SSE	Sum of Squared Errors	80.325	37.959
DISF	Frequency of Classified Cases	348.000	150.000
MISC	Misclassification Rate	0.181	0.180
WRONG	Number of Wrong Classifications	63.000	27.000

Classification Table

Data Role=TRAIN Target Variable=Churn Target Label=Churn

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	81.8966	100	285	81.8966
1	0	18.1034	100	63	18.1034

Data Role=VALIDATE Target Variable=Churn Target Label=Churn

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	82	100	123	82
1	0	18	100	27	18

Event Classification Table

Data Role=TRAIN Target=Churn Target Label=Churn

False Negative	True Negative	False Positive	True Positive
63	285	0	0

Data Role=VALIDATE Target=Churn Target Label=Churn

False Negative	True Negative	False Positive	True Positive
27	123	0	0

Data Role=TRAIN Target Variable=Churn Target Label=Churn

Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	452.381	5.52381	5.52381	100.000	100.000	18	0.35769
10	452.381	5.52381	5.52381	100.000	100.000	17	0.23902
15	400.270	3.98942	5.00270	72.222	90.566	18	0.19351
20	302.449	0.97479	4.02449	17.647	72.857	17	0.18325
25	249.206	1.29972	3.49206	23.529	63.218	17	0.17096
30	205.125	0.92063	3.05125	16.667	55.238	18	0.15818
35	167.135	0.32493	2.67135	5.882	48.361	17	0.15009
40	148.571	1.22751	2.48571	22.222	45.000	18	0.14376
45	121.656	0.00000	2.21656	0.000	40.127	17	0.13616
50	100.000	0.00000	2.00000	0.000	36.207	17	0.13175
55	81.250	0.00000	1.81250	0.000	32.813	18	0.12794
60	66.507	0.00000	1.66507	0.000	30.144	17	0.12191
65	53.304	0.00000	1.53304	0.000	27.753	18	0.10697
70	42.623	0.00000	1.42623	0.000	25.820	17	0.09689
75	33.333	0.00000	1.33333	0.000	24.138	17	0.07776
80	24.731	0.00000	1.24731	0.000	22.581	18	0.05924
85	17.568	0.00000	1.17568	0.000	21.284	17	0.04946
90	10.828	0.00000	1.10828	0.000	20.064	18	0.04453
95	5.136	0.00000	1.05136	0.000	19.033	17	0.03963
100	0.000	0.00000	1.00000	0.000	18.103	17	0.03486

Data Role=VALIDATE Target Variable=Churn Target Label=Churn

Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	455.556	5.55556	5.55556	100.000	100.000	8	0.28649
10	418.519	4.76190	5.18519	85.714	93.333	7	0.20207
15	358.937	3.47222	4.58937	62.500	82.609	8	0.18704
20	270.370	0.79365	3.70370	14.286	66.667	7	0.18158
25	192.398	0.00000	2.92398	0.000	52.632	8	0.17433
30	171.605	1.58730	2.71605	28.571	48.889	7	0.16877
35	141.090	0.69444	2.41090	12.500	43.396	8	0.15511
40	131.481	1.58730	2.31481	28.571	41.667	7	0.15056
45	112.418	0.69444	2.12418	12.500	38.235	8	0.14311
50	92.593	0.00000	1.92593	0.000	34.667	7	0.13316
55	74.029	0.00000	1.74029	0.000	31.325	8	0.12938
60	60.494	0.00000	1.60494	0.000	28.889	7	0.12568
65	47.392	0.00000	1.47392	0.000	26.531	8	0.11715

Deliverables:

A report detailing each step of the process, including the rationale behind your choices and any challenges faced. An analysis of the decision tree and ensemble methods, with insights into customer behavior and suggestions for business strategy.

[5 marks]

For this experiment, I first had to find a dataset that met the requirements. After finding the dataset, I imported the data into SAS. Viewed the data in SAS. It was found that the three columns city, age & churn had missing data. For the missing city value, I used the method of taking the plurality to fill in the missing values. Plurality is particularly useful for filling in missing values for categorical variables. This is because for categorical variables, the mean and median may not make sense and the plurality is an intuitive and easy-to-understand alternative. Using plurals to fill in missing values can also help preserve the distributional properties of the original variable because plurals are the most common values in the original data and have less impact on the overall distribution. For the missing value of AGE, I chose to use the median to fill in the missing value. Because age is a numeric variable and the missing values are relatively evenly distributed, using the median is a simple and common method. This helps to maintain the central tendency of the overall dataset. For missing churn values, I chose to deal with them by deleting the missing values. Because the percentage of missing values is relatively small, the impact of missing values is minimal for analysis or modeling tasks. And churn as a target attribute, its value has a large impact on the result. So it is processed by deletion. After processing the data, I started setting up the specified variable roles to use churn as the target attribute.

Then I partitioned the data and adjusted the ratio of the training set to the test set to 7:3, followed by variable transformation. The model of random numbers was chosen to analyze the data. The results obtained are shown in the second question. The misclassification Rate is 0.052 on the training set and 0.100 on the validation set. this means that the model misclassifies at a rate of 5.2% on the training set and 10% on the validation set. A lower misclassification rate is usually good, but further consideration of the model's generalization performance may be needed. Maximum Absolute ErrorThe maximum absolute error was 0.941 on both the training and validation sets. In addition, the performance of the validation set can be used to assess the model's generalization ability. Further optimization or tuning of the model parameters may be required to avoid overfitting.

The biggest challenge I faced was the third question. At first, I was at a loss as to how to set up bagging and boosting in SAS. In my analysis, I urgently needed to optimize the model and therefore needed to utilize these two integrated learning techniques. Looking at the SAS website, I finally found the solution, which is to use the start group statement to adjust the model's mode of operation. Specifically, by setting the mode parameter in the start group statement, I was able to easily choose whether to use bagging or boosting, and I learned that the index count is 10 by default, which provides a default setting for my model that I can adapt to my specific needs. I also encountered some confusion in determining which model to use. I was hesitant to decide whether a random forest -based model or a decision tree-based model would be more appropriate for my analysis. Eventually, I decided to go with the decision tree-based model because it seemed to fit my research question better.

The analysis shows that the number of customer purchases, spending and churn are positively correlated, which may indicate that customers have higher expectations of quality and experience when purchasing a product or service and are more willing to spend more if these expectations are met. Therefore, specialized customer retention strategies, such as exclusive member services and regular promotions, can be designed for high-frequency purchasing, high-spending, and high-satisfaction customers to further cement their loyalty. Utilize purchase history data to implement targeted marketing campaigns to encourage low-frequency purchasing, low-spending, but high-satisfaction customers to increase the number of purchases and spending.

Dataset Links：https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis

Github：

Project link:https://odamid-apse1.oda.sas.com/SASStudio/main?locale=zh_CN&zone=GMT%252B08%253A00&ticket=ST-65233-JFdqOjXjMPh7HIlUJITX-cas

User:22064827@siswaum.edu.my password:Wcl39602

Objective:

The case study aims to assess students' ability to apply decision tree and ensemble methods in a practical context, demonstrating their understanding of the concepts and their ability to derive meaningful business insights from data analysis.

- End -