WQD7005 Data Mining

Group 2

Alternative Assessment 1

Name	Low Boon Kiat
Matric Number	17138399
Submission Date	7th January 2024

Table of Contents

1.0 Introduction

1.1 Project Overview

The project involves working with a dataset of customer transactions from an e-commerce website, encompassing various customer attributes and purchase history. The project 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. Three tools used in this project include Talend Data Preparation, Talend Data Integration and SAS Enterprise Miner. The role of Talend Data Preparation is to perform basic data preprocessing. The role of Talend Data Integration is to combine several datasets into one. The role of SAS Enterprise Miner is to perform data mining, data preprocessing and predictive modelling.

1.2 Dataset Description

The dataset is downloaded from Kaggle, containing information about customers of an e-commerce company. It consists of 20 columns stored in 2 sheets with a total of 5630 records. Table below summarizes the variables names and description in the dataset (Sheet 01).

Variables	Description
CustomerID	Unique customer ID
Churn	Churn Flag (1 for churned, 0 for active)
Tenure	Tenure of customer in organization
PreferredLoginDevice	Preferred login device of customer
CityTier	City tier
WarehouseToHome	Distance in between warehouse to home of customer
PreferredPaymentMode	Preferred payment method of customer
Gender	Gender of customer
HourSpendOnApp	Number of hours spend on mobile application or website
NumberOfDeviceRegistered	Total number of deceives is registered on customer
PreferedOrderCat	Preferred order category of customer in last month
SatisfactionScore	Satisfactory score of customers on service
MaritalStatus	Marital status of customer
NumberOfAddress	Total number of addresses added on customer
Complain	Any complaint has been raised in last month
OrderAmountHikeFromlastYear	Percentage increases in order from last year
CouponUsed	Total number of coupons has been used in last month
OrderCount	Total number of orders has been places in last month
DaySinceLastOrder	Day Since last order by customer
CashbackAmount	Average cashback in last month

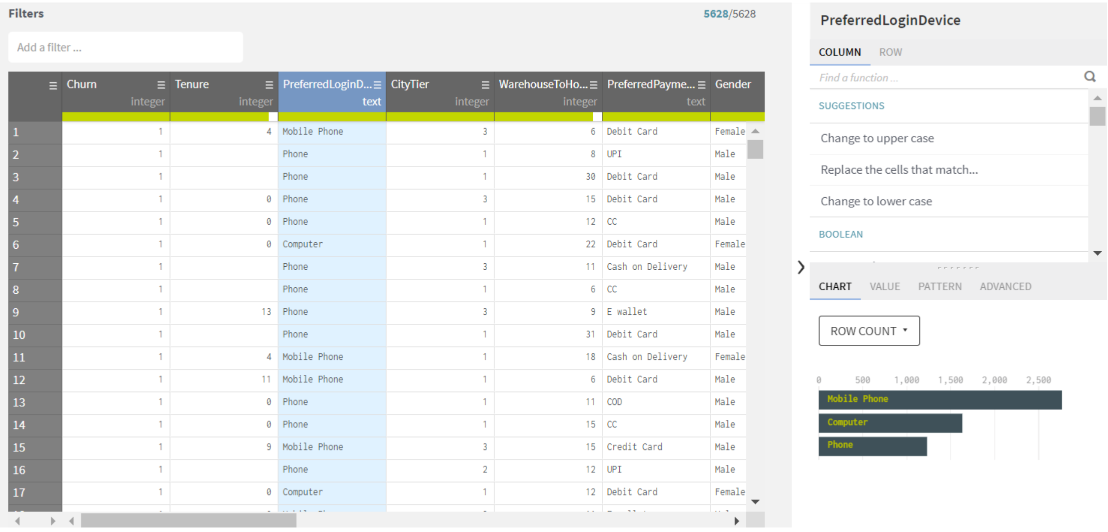
Table below summarizes the variables names and description in the dataset (Sheet 02). Marital status information was collected from customers at a later time.

Variables	Description
CustomerID	Unique customer ID
MaritalStatus	Marital status of customer

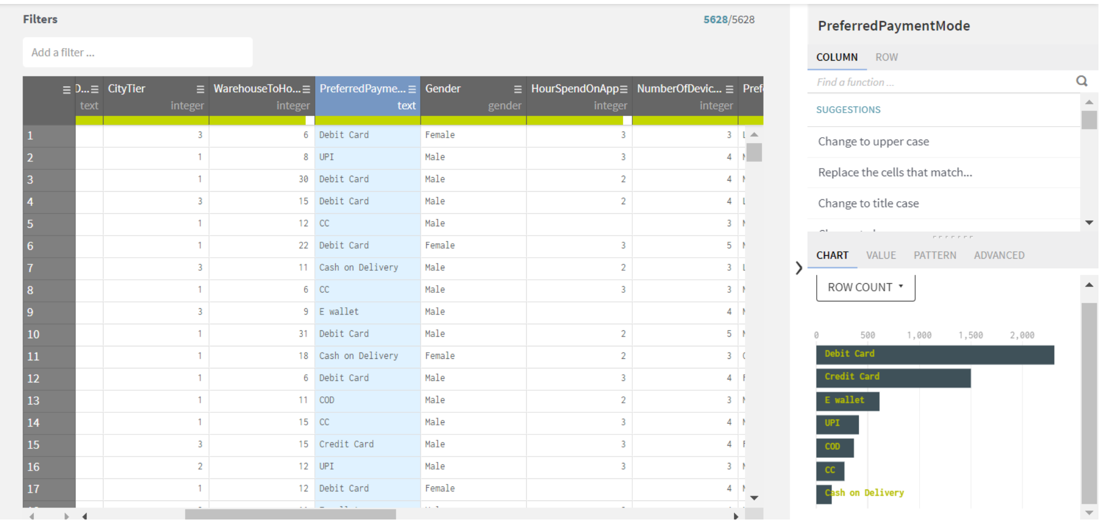
2.0 Data Pre-processing using Talend Data Preparation

2.1 Handling Inconsistent Data

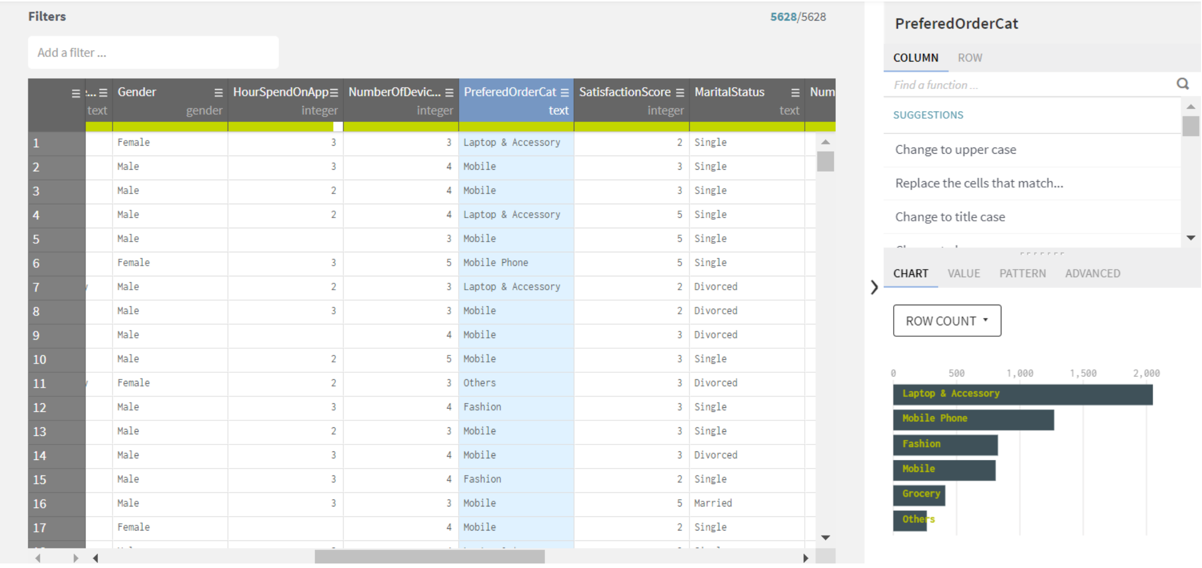
In the “PreferredLoginDevice” column, “Phone” and “Mobile Phone” referred to the same type of device. To ensure data consistency, “Phone” was replaced with “Mobile Phone”.



In the “PreferredPaymentMode” column, “Credit Card” and “CC” referred to the same type of payment mode whereas “Cash on Delivery” and “COD” referred to the same type of payment mode. To ensure data consistency, “CC” was replaced with “Credit Card” whereas “COD” was replaced with “Cash on Delivery”. In addition, “UPI” was replaced with “Unified Payments Interface”.



In the “PreferedOrderCat” column, “Mobile Phone” should be a subset of “Mobile”. To ensure data consistency, “Mobile Phone” was replaced with “Mobile”.



2.2 Handling Outliers

In the “WarehouseToHome” column, there were 2 records with values 126 and 127. These two values were far from majority records as shown in the data distribution chart. To handle outliers, 126 and 127 were adjusted to 26 and 27 respectively to align them within the appropriate range for this column.

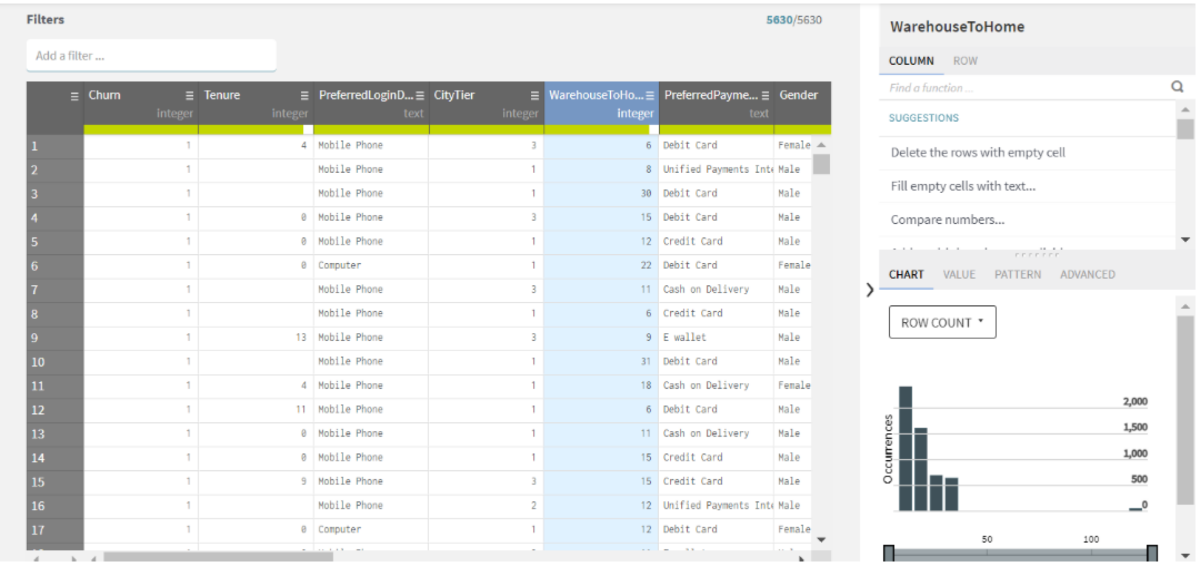
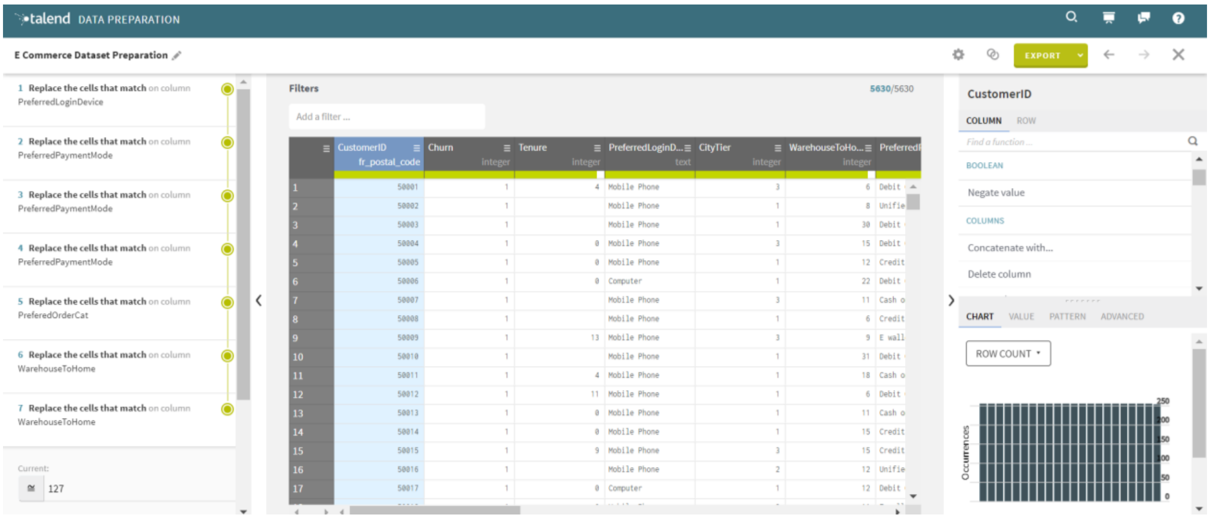
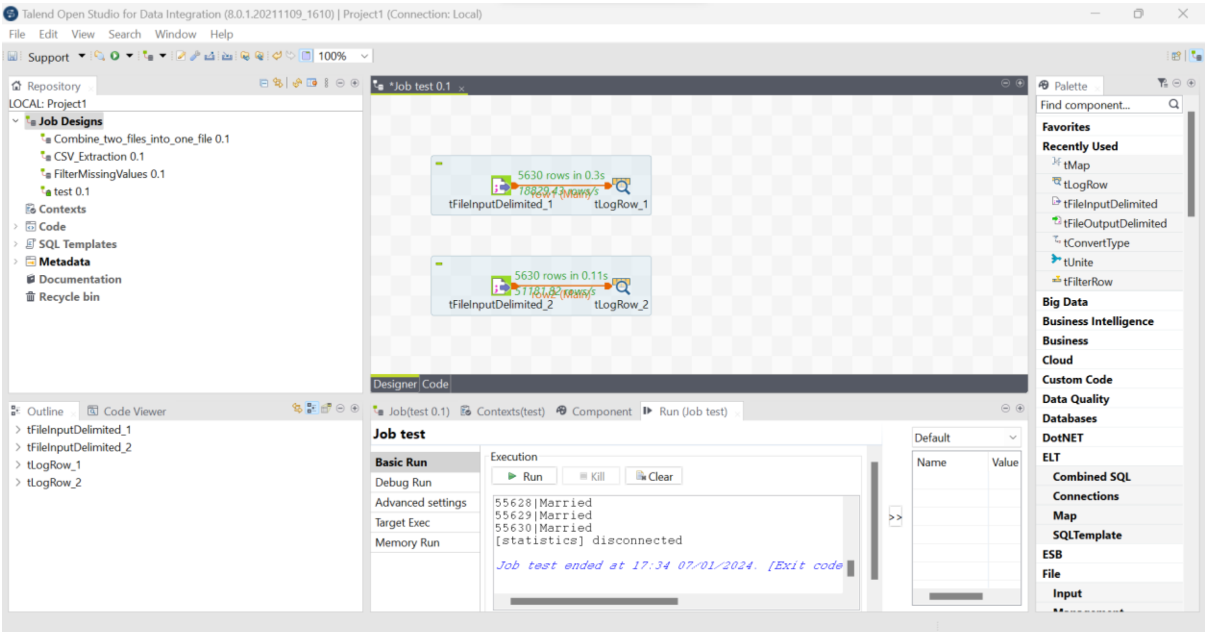


Figure below illustrates all the data pre-processing steps performed in Talend Data Preparation.

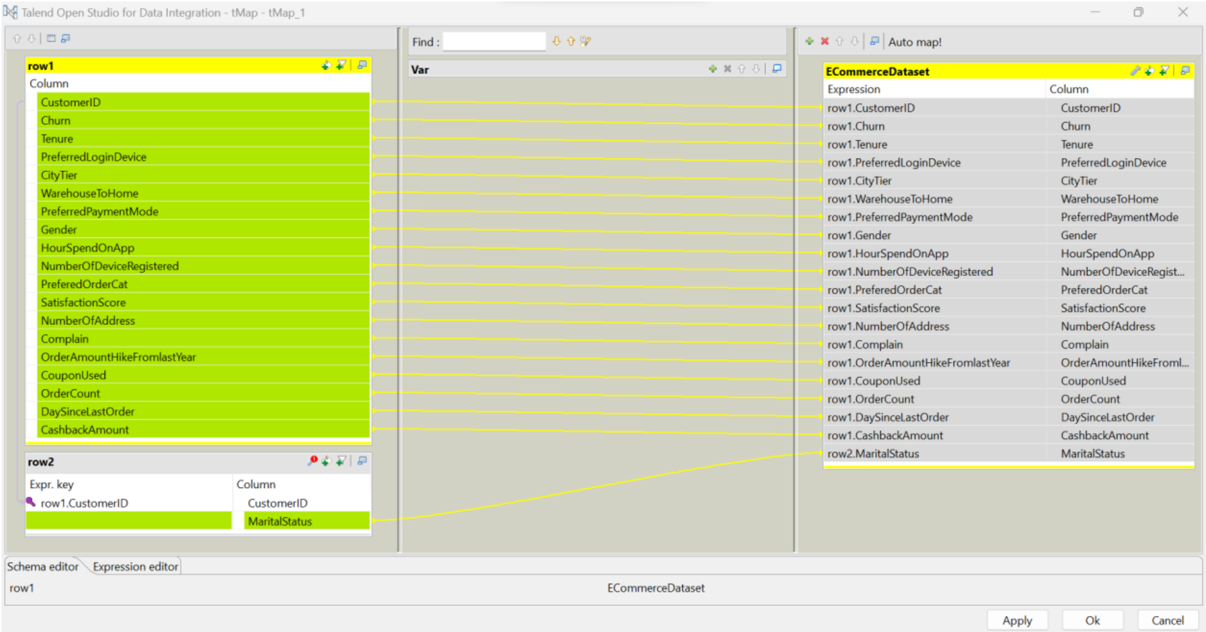


3.0 Data Integration using Talend Data Integration

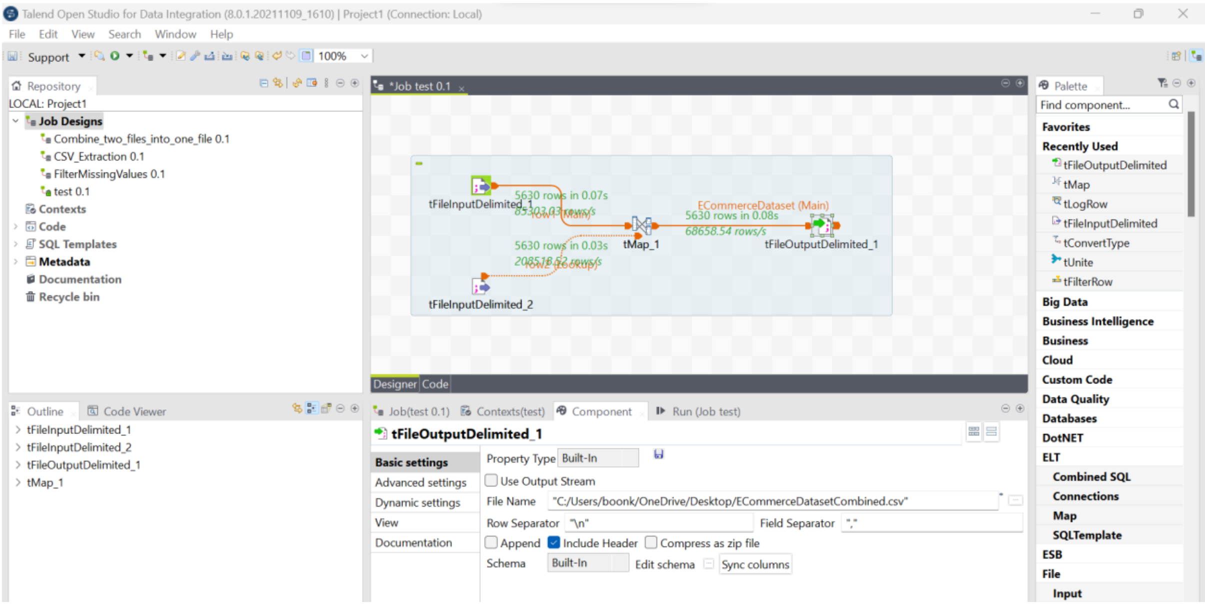
Use “tFileInputDelimited” to import both each dataset.



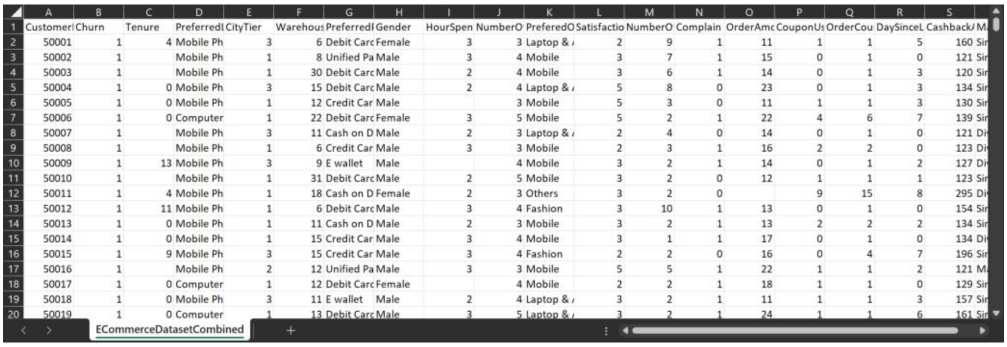
Use “tMap” to join two datasets. Drag the “CustomerID” column in the first dataset to join with the “CustomerID” column in the second dataset.



Use “tFileOutputDelimited” to export the joined dataset.



Display the CSV output from Talend Data Integration.



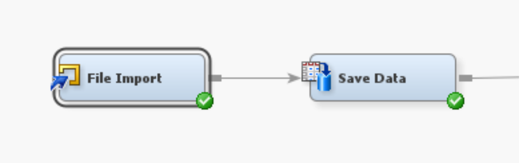
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	Customerl	Churn	Tenure	Preferredl	CityTier	Warehous	Preferredl	Gender	HourSpen	NumberO	PreferedO	Satisfactio	NumberO	Complain	OrderAmc	CouponUs	OrderCou	DaySinceL	Cashback/	M
2	50001	1	4	Mobile Ph	3	6	Debit Carc	Female	3	3	Laptop & /	2	9	1	11	1	1	5	160	Si
3	50002	1		Mobile Ph	1	8	Unified Pa	Male	3	4	Mobile	3	7	1	15	0	1	0	121	Si
4	50003	1		Mobile Ph	1	30	Debit Carc	Male	2	4	Mobile	3	6	1	14	0	1	3	120	Si
5	50004	1	0	Mobile Ph	3	15	Debit Carc	Male	2	4	Laptop & /	5	8	0	23	0	1	3	134	Si
6	50005	1	0	Mobile Ph	1	12	Credit Car	Male		3	Mobile	5	3	0	11	1	1	3	130	Si
7	50006	1	0	Computer	1	22	Debit Carc	Female	3	5	Mobile	5	2	1	22	4	6	7	139	Si
8	50007	1		Mobile Ph	3	11	Cash on D	Male	2	3	Laptop & /	2	4	0	14	0	1	0	121	Di
9	50008	1		Mobile Ph	1	6	Credit Car	Male	3	3	Mobile	2	3	1	16	2	2	0	123	Di
10	50009	1	13	Mobile Ph	3	9	E wallet	Male		4	Mobile	3	2	1	14	0	1	2	127	Di
11	50010	1		Mobile Ph	1	31	Debit Carc	Male	2	5	Mobile	3	2	0	12	1	1	1	123	Si
12	50011	1	4	Mobile Ph	1	18	Cash on D	Female	2	3	Others	3	2	0		9	15	8	295	Di
13	50012	1	11	Mobile Ph	1	6	Debit Carc	Male	3	4	Fashion	3	10	1	13	0	1	0	154	Si
14	50013	1	0	Mobile Ph	1	11	Cash on D	Male	2	3	Mobile	3	2	1	13	2	2	2	134	Si
15	50014	1	0	Mobile Ph	1	15	Credit Car	Male	3	4	Mobile	3	1	1	17	0	1	0	134	Di
16	50015	1	9	Mobile Ph	3	15	Credit Car	Male	3	4	Fashion	2	2	0	16	0	4	7	196	Si
17	50016	1		Mobile Ph	2	12	Unified Pa	Male	3	3	Mobile	5	5	1	22	1	1	2	121	M
18	50017	1	0	Computer	1	12	Debit Carc	Female		4	Mobile	2	2	1	18	1	1	0	129	Si
19	50018	1	0	Mobile Ph	3	11	E wallet	Male	2	4	Laptop & /	3	2	1	11	1	1	3	157	Si
20	50019	1	0	Computer	1	13	Debit Carc	Male	3	5	Laptop & /	3	2	1	24	1	1	6	161	Si

ECommerceDatasetCombined

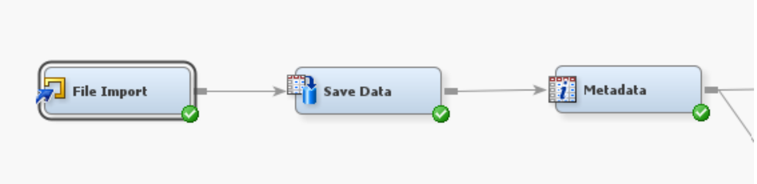
4.0 Data Import and Pre-processing using SAS Enterprise Miner

4.1 Importing Data

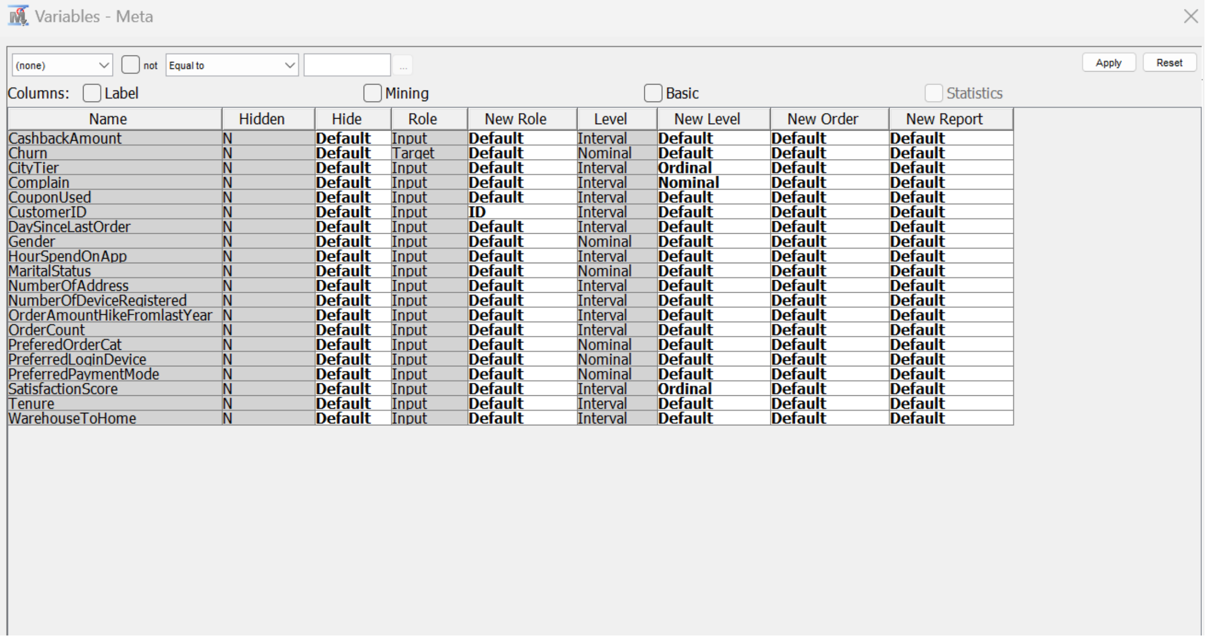
Import the CSV file using "File Import" node. Save it as a SAS file.



Specify the column metadata using "Metadata" node.



The roles and measurement levels of some variables were re-assigned to correctly define the column metadata. For instance, "customerID" should be assigned the role of ID instead of input because it serves as an identifier for individual customers rather than being used as an input feature for modelling. "CityTier" and "SatisfactionScore" should be considered as ordinal variables. "City Tier" ranks cities into different tiers, typically based on their economic development and other similar factors, making it an ordinal variable. For "SatisfactionScore", the numbers represent a respondent's level of satisfaction with a product or service, making it an ordinal variable. "Complain" taking values of 0 or 1 should be considered as nominal variable because it indicates the presence or absence of a complaint without any inherent order or ranking, hence it should be treated as a nominal variable.



Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report
CashbackAmount	N	Default	Input	Default	Interval	Default	Default	Default
Churn	N	Default	Target	Default	Nominal	Default	Default	Default
CityTier	N	Default	Input	Default	Interval	Ordinal	Default	Default
Complain	N	Default	Input	Default	Interval	Nominal	Default	Default
CouponUsed	N	Default	Input	Default	Interval	Default	Default	Default
CustomerID	N	Default	Input	ID	Interval	Default	Default	Default
DaySinceLastOrder	N	Default	Input	Default	Interval	Default	Default	Default
Gender	N	Default	Input	Default	Nominal	Default	Default	Default
HourSpendOnApp	N	Default	Input	Default	Interval	Default	Default	Default
MaritalStatus	N	Default	Input	Default	Nominal	Default	Default	Default
NumberOfAddress	N	Default	Input	Default	Interval	Default	Default	Default
NumberOfDeviceRegistered	N	Default	Input	Default	Interval	Default	Default	Default
OrderAmountHikeFromlastYear	N	Default	Input	Default	Interval	Default	Default	Default
OrderCount	N	Default	Input	Default	Interval	Default	Default	Default
PreferedOrderCat	N	Default	Input	Default	Nominal	Default	Default	Default
PreferredLoginDevice	N	Default	Input	Default	Nominal	Default	Default	Default
PreferredPaymentMode	N	Default	Input	Default	Nominal	Default	Default	Default
SatisfactionScore	N	Default	Input	Default	Interval	Ordinal	Default	Default
Tenure	N	Default	Input	Default	Interval	Default	Default	Default
WarehouseToHome	N	Default	Input	Default	Interval	Default	Default	Default

4.2 Handling Missing Data

Check for missing values using StatExplore Node.

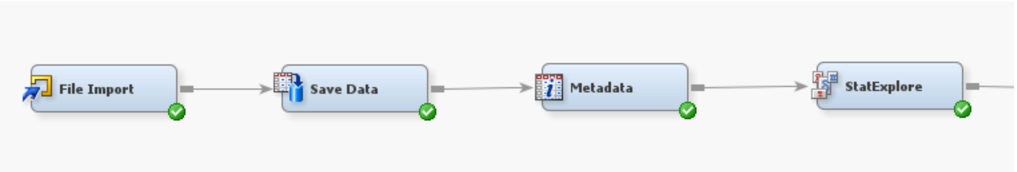


Figure below shows variables summary after specifying the metadata.

Variable Summary

Role	Measurement Level	Frequency Count
ID	INTERVAL	1
INPUT	INTERVAL	10
INPUT	NOMINAL	6
INPUT	ORDINAL	2
TARGET	NOMINAL	1

Figure below shows class variables summary. None of the class variables have missing values.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	CityTier	INPUT	3	0	1	65.12	3	30.59
TRAIN	Complain	INPUT	2	0	0	71.51	1	28.49
TRAIN	Gender	INPUT	2	0	Male	60.11	Female	39.89
TRAIN	MaritalStatus	INPUT	3	0	Married	53.04	Single	31.90
TRAIN	PreferedOrderCat	INPUT	5	0	Mobile	36.94	Laptop & Accessory	36.41
TRAIN	PreferredLoginDevice	INPUT	2	0	Mobile Phone	70.98	Computer	29.02
TRAIN	PreferredPaymentMode	INPUT	5	0	Debit Card	41.10	Credit Card	31.51
TRAIN	SatisfactionScore	INPUT	5	0	3	30.16	1	20.67
TRAIN	Churn	TARGET	2	0	0	83.16	1	16.84

Figure below shows interval variables summary. Seven variables namely “CouponUsed”, “DaySinceaLastOrder”, “HourSpendOnApp”, “OrderAmountHikeFromLastYear”, “OrderCount”, “Tenure” and “WarehouseToHome” have missing values.

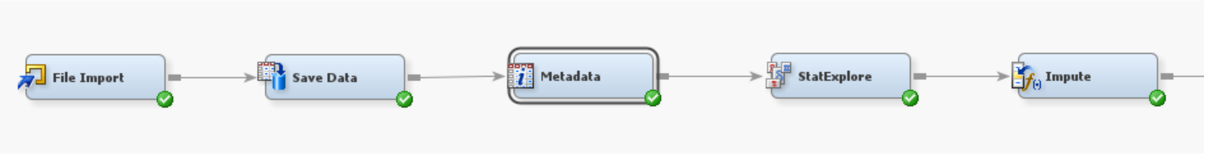
Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
CashbackAmount	INPUT	177.2215	49.19387	5630	0	0	163	325	1.149595	0.973546
CouponUsed	INPUT	1.751023	1.894621	5374	256	0	1	16	2.545653	9.132281
DaySinceLastOrder	INPUT	4.543491	3.654433	5323	307	0	3	46	1.191	4.023964
HourSpendOnApp	INPUT	2.931535	0.721926	5375	255	0	3	5	-0.02721	-0.66708
NumberOfAddress	INPUT	4.214032	2.583586	5630	0	1	3	22	1.088639	0.959229
NumberOfDeviceRegistered	INPUT	3.688988	1.023999	5630	0	1	4	6	-0.39697	0.582849
OrderAmountHikeFromlastYear	INPUT	15.70792	3.675485	5365	265	11	15	26	0.790785	-0.28038
OrderCount	INPUT	3.008004	2.93968	5372	258	1	2	16	2.196414	4.718466
Tenure	INPUT	10.1899	8.557241	5366	264	0	9	61	0.736513	-0.00737
WarehouseToHome	INPUT	15.60271	8.261845	5379	251	5	14	36	0.898406	-0.28639

4.3 Imputing Missing Data

Impute missing values using “Impute” node



The missing values of interval variables were imputed using the mean values. Imputing missing values with the mean assumes that the missing data is missing completely at random (MCAR) or missing at random (MAR) and does not introduce bias.

Property	Value
General	
Node ID	Impt
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	...
Tree Imputation	...
Score	
Hide Original Variables	Yes
Indicator Variables	
Type	None
Source	Imputed Variables
Role	Rejected
Report	
Validation and Test Data	No
Distribution of Missing	No
Status	
Create Time	1/6/24 4:06 PM
Run ID	4eea0a9d-1763-f148-8f23-
Last Error	
Last Status	Complete
Last Run Time	1/7/24 2:46 AM
Run Duration	0 Hr. 0 Min. 3.20 Sec.
Grid Host	
User-Added Node	No

4.4 Assessing Impact of Imputation

Assess the impact of imputation using “StatExplore” node.

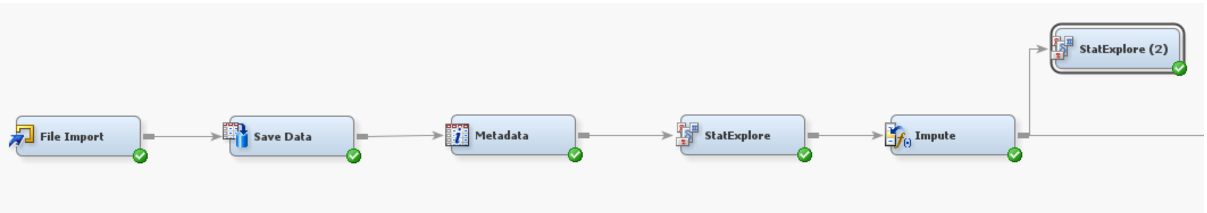


Figure below shows interval variables summary before imputation.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
CashbackAmount	INPUT	177.2215	49.19387	5630	0	0	163	325	1.149595	0.973546
CouponUsed	INPUT	1.751023	1.894621	5374	256	0	1	16	2.545653	9.132281
DaySinceLastOrder	INPUT	4.543491	3.654433	5323	307	0	3	46	1.191	4.023964
HourSpendOnApp	INPUT	2.931535	0.721926	5375	255	0	3	5	-0.02721	-0.66708
NumberOfAddress	INPUT	4.214032	2.583586	5630	0	1	3	22	1.088639	0.959229
NumberOfDeviceRegistered	INPUT	3.688988	1.023999	5630	0	1	4	6	-0.39697	0.582849
OrderAmountHikeFromlastYear	INPUT	15.70792	3.675485	5365	265	11	15	26	0.790785	-0.28038
OrderCount	INPUT	3.008004	2.93968	5372	258	1	2	16	2.196414	4.718466
Tenure	INPUT	10.1899	8.557241	5366	264	0	9	61	0.736513	-0.00737
WarehouseToHome	INPUT	15.60271	8.261845	5379	251	5	14	36	0.898406	-0.28639

Figure below shows interval variables summary after imputation. The imputation did not significantly alter the distribution or central tendencies of the variables.

Interval Variable Summary Statistics
(maximum 500 observations printed)

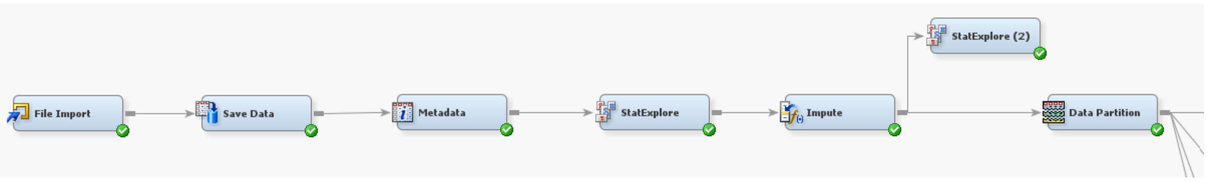
Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
CashbackAmount	INPUT	177.2215	49.19387	5630	0	0	163	325	1.149595	0.973546
IMP_CouponUsed	INPUT	1.751023	1.851038	5630	0	0	1	16	2.605547	9.709842
IMP_DaySinceLastOrder	INPUT	4.543491	3.553382	5630	0	0	4	46	1.224844	4.428875
IMP_HourSpendOnApp	INPUT	2.931535	0.705384	5630	0	0	3	5	-0.02785	-0.55635
IMP_OrderAmountHikeFromlastYear	INPUT	15.70792	3.587926	5630	0	11	15	26	0.810069	-0.14601
IMP_OrderCount	INPUT	3.008004	2.871521	5630	0	1	2	16	2.24851	5.088971
IMP_Tenure	INPUT	10.1899	8.354164	5630	0	0	9	61	0.754404	0.139888
IMP_WarehouseToHome	INPUT	15.60271	8.075545	5630	0	5	14	36	0.919117	-0.15973
NumberOfAddress	INPUT	4.214032	2.583586	5630	0	1	3	22	1.088639	0.959229
NumberOfDeviceRegistered	INPUT	3.688988	1.023999	5630	0	1	4	6	-0.39697	0.582849

5.0 Decision Tree Modelling using SAS Enterprise Miner

5.1 Data Partition

Specify the ratio of training/validation data using "Data Partition" node.

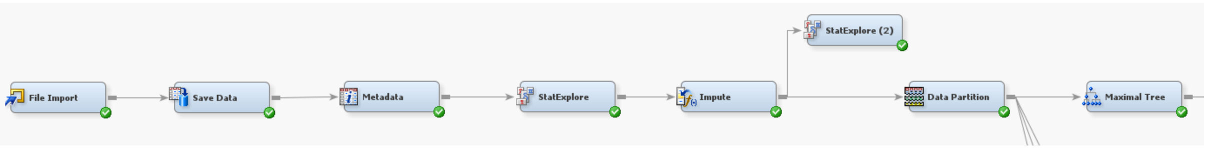


The ratio of training and validation data is 70/30.

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	1/6/24 4:11 PM
Run ID	efb9ef80-6514-9e4d-bffd-(
Last Error	
Last Status	Complete
Last Run Time	1/7/24 5:27 AM
Run Duration	0 Hr. 0 Min. 3.46 Sec.
Grid Host	
User-Added Node	No

5.2 Maximal Decision Tree

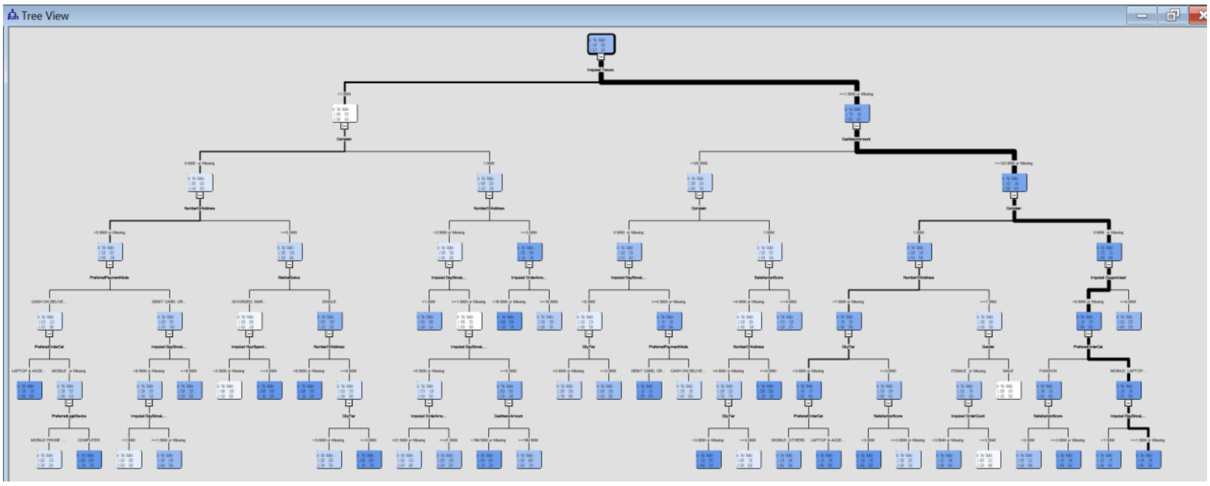
Create the maximal tree using "Decision Tree" node.



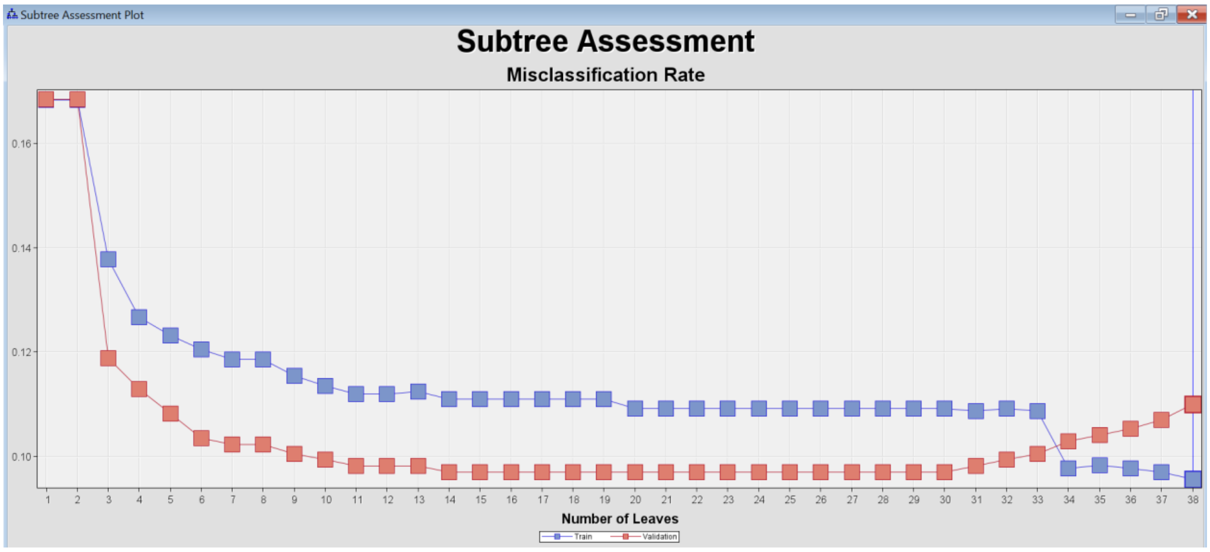
Click on the "…" button at "Interactive" row to open the Interactive Decision Tree tool.

Property	Value
General	
Node ID	Tree
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No

Right click on the root node of the tree and select Train Node. This will grow the tree until stopping rules prohibited further growth. Figure below shows the maximal tree with 38 leaves.



Based on the Subtree Assessment Plot, it appears that the maximal, 38-leaf tree gives a lower misclassification rate than any of its simpler predecessors. However, it is misleading because it applies to training data only. Further optimization is therefore required.



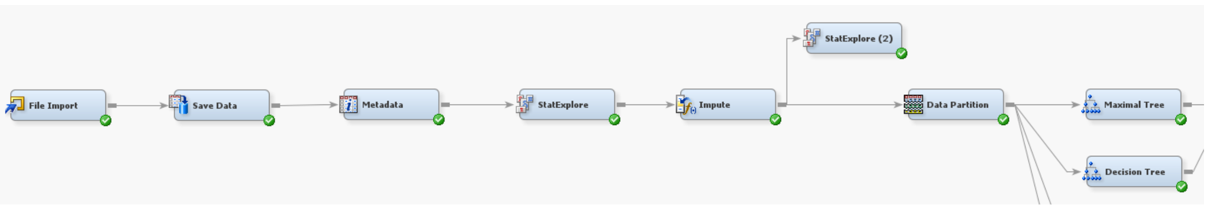
Based on Fit Statistics, misclassification rate is 0.0957 for training dataset and 0.1099 for validation dataset.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		NOBS	Sum of Frequencies	3939	1691	.
Churn		MISC	Misclassification Rate	0.09571	0.109994	.
Churn		MAX	Maximum Absolute Error	0.986111	1	.
Churn		SSE	Sum of Squared Errors	555.5301	267.5079	.
Churn		ASE	Average Squared Error	0.070517	0.079098	.
Churn		RASE	Root Average Squared Error	0.26555	0.281243	.
Churn		DIV	Divisor for ASE	7878	3382	.
Churn		DFT	Total Degrees of Freedom	3939	.	.

5.3 Pruned Decision Tree

Create a decision tree using "Decision Tree" node.

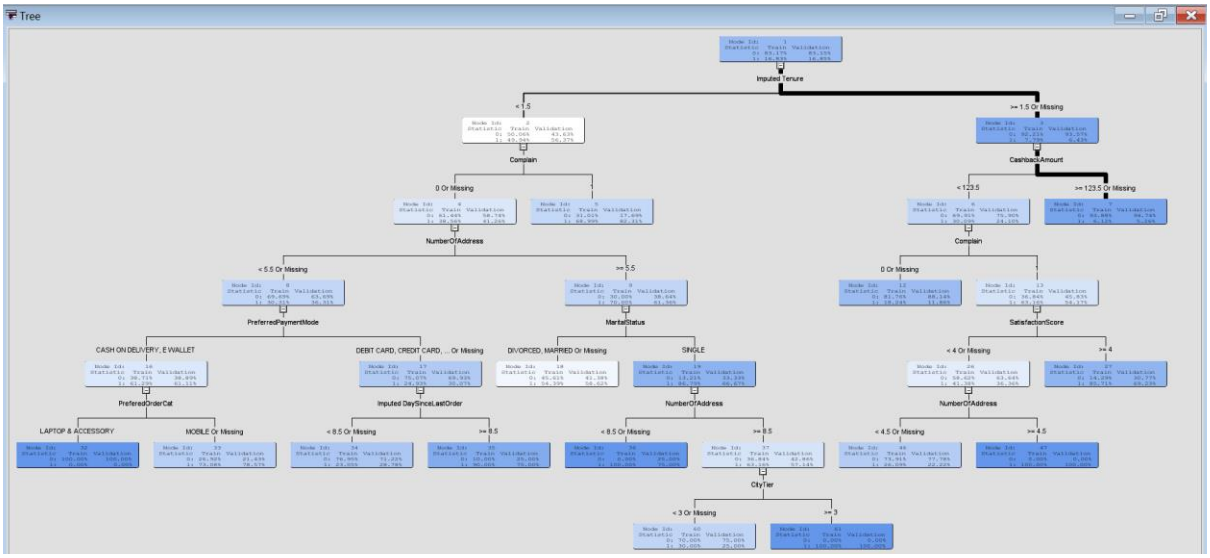


Go to the "Subtree" section of the properties table to specify the tree pruning properties. The method used to prune the maximal tree is Assessment. This means that the algorithms choose the best tree based on the optimality measure specificized by the Assessment Measure. By setting Assessment Measure as Decision, the algorithms will choose a tree that is optimized for making the best decisions (as opposed to best rankings or best probability estimates). Keep other settings as default.

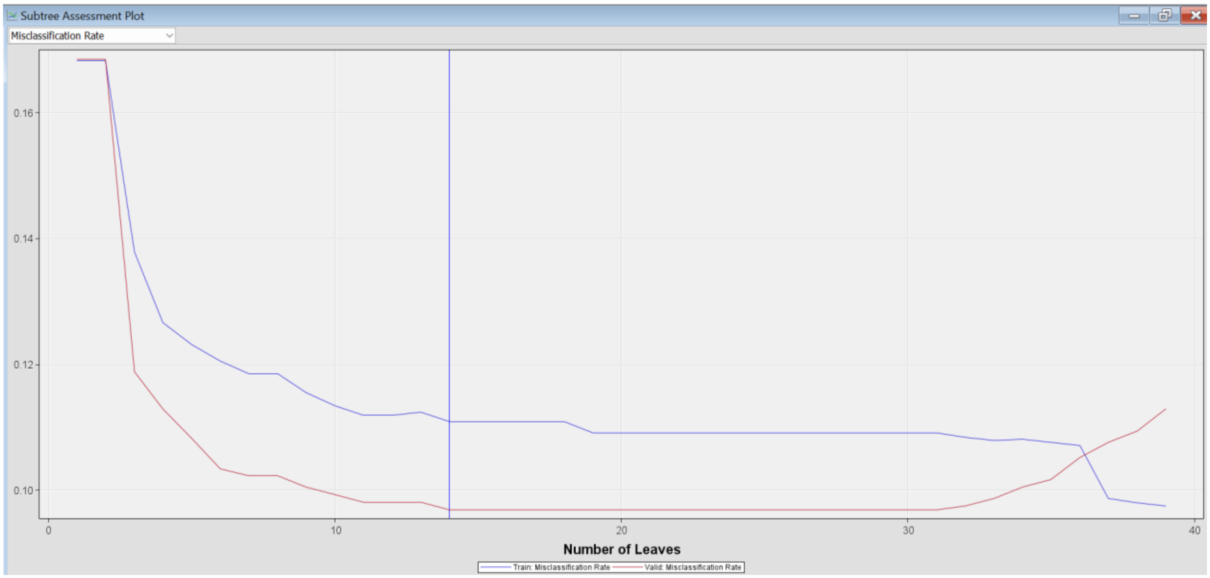
.. Property	Value
General	
Node ID	Tree2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000

Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Interactive Sample	
Create Sample	Default
Sample Method	Random
Sample Size	10000
Sample Seed	12345
Performance	Disk
Score	
Variable Selection	Yes
Leaf Role	Segment
Report	
Precision	4
Tree Precision	4
Class Target Node Color	Percent Correctly Classi
Interval Target Node Color	Average
Node Text	...

Figure below shows the pruned decision tree with 14 leaves.



Based on the Subtree Assessment Plot, it appears that misclassification rate is most optimized when the number of leaves equals 14. The validation misclassification rate plateaued out at 0.097 when number of leaves increased from 15 to 31. Beyond 31, validation misclassification rate increases. Therefore, 14 leaves give the most optimized misclassification rate.



Based on Fit Statistics, misclassification rate is 0.1109 for training dataset and 0.09698 for validation dataset.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		NOBS	Sum of Frequencies	3939	1691	.
Churn		MISC	Misclassification Rate	0.110942	0.096984	.
Churn		MAX	Maximum Absolute Error	0.938846	1	.
Churn		SSE	Sum of Squared Errors	706.6919	278.7777	.
Churn		ASE	Average Squared Error	0.089704	0.08243	.
Churn		RASE	Root Average Squared Error	0.299507	0.287106	.
Churn		DIV	Divisor for ASE	7878	3382	.
Churn		DFT	Total Degrees of Freedom	3939	.	.

The Variable Importance Plot displays the importance of each predictor variable in the model. Only 10 out of 18 input variables are important to the pruned decision tree model.

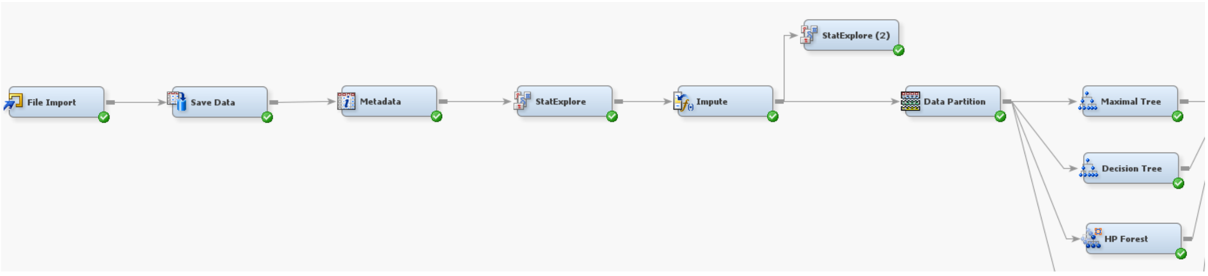
Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
IMP_Tenure	Imputed Tenure	1	1.0000	1.0000	1.0000
Complain		2	0.4765	0.4814	1.0102
NumberOfAddress		3	0.3905	0.1960	0.5019
CashbackAmount		1	0.3130	0.1963	0.6273
PreferredPaymentMode		1	0.2434	0.2114	0.8688
PreferedOrderCat		1	0.1949	0.2351	1.2064
IMP_DaySinceLastOrder	Imputed DaySi...	1	0.1922	0.0988	0.5140
MaritalStatus		1	0.1564	0.0000	0.0000
SatisfactionScore		1	0.1541	0.0820	0.5320
CityTier		1	0.1403	0.1201	0.8557
NumberOfDeviceRegistered		0	0.0000	0.0000	.
PreferredLoginDevice		0	0.0000	0.0000	.
IMP_CouponUsed	Imputed Coupo...	0	0.0000	0.0000	.
IMP_OrderAmountHikeFromlastYear	Imputed Order...	0	0.0000	0.0000	.
IMP_HourSpendOnApp	Imputed HourS...	0	0.0000	0.0000	.
Gender		0	0.0000	0.0000	.
IMP_OrderCount	Imputed Order...	0	0.0000	0.0000	.
IMP_WarehouseToHome	Imputed Wareh...	0	0.0000	0.0000	.

6.0 Ensemble Methods using SAS Enterprise Miner

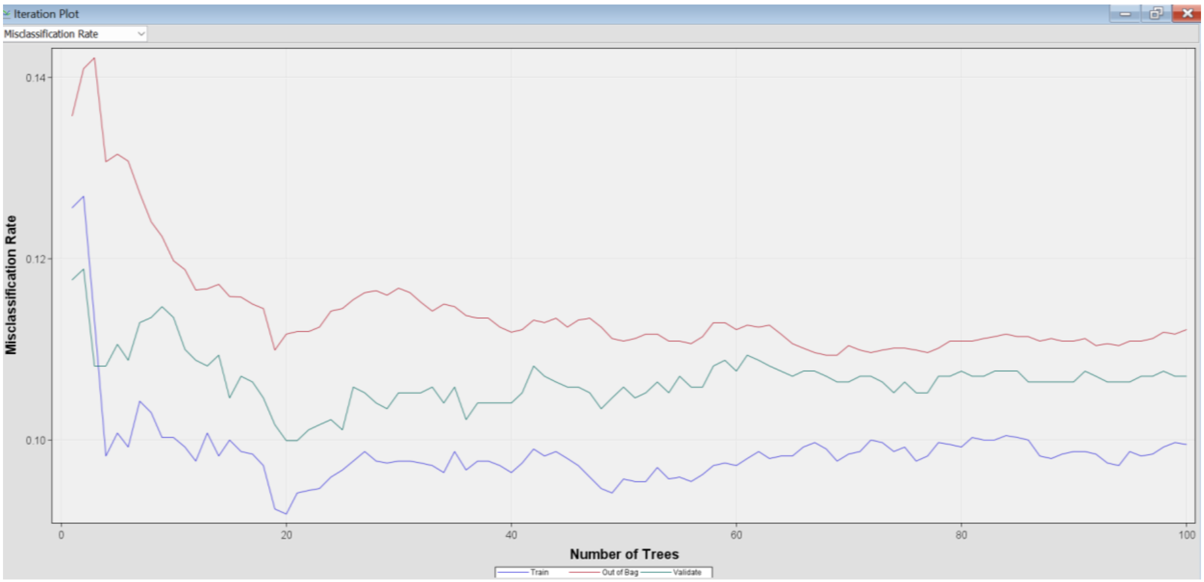
6.1 Bagging

Create a model for Bagging using "HP Forest" node. Keep default settings.



Property	Value
General	
Node ID	HPDMForest
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Tree Options	
Maximum Number of Trees	100
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each Sample	0.6
Number of Obs in Each Sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Consider in S	.
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest Percentage of Obs in Node	1.0E-5
Smallest Number of Obs in Node	1
Split Size	.
Use as Modeling Node	Yes
Score	
Variable Selection	Yes
Variable Importance Method	Loss Reduction
Number of Variables to Consider	25
Cutoff Fraction	0.01

Based on Iteration Plot, misclassification rate plateaued out when number of trees reaches 20.



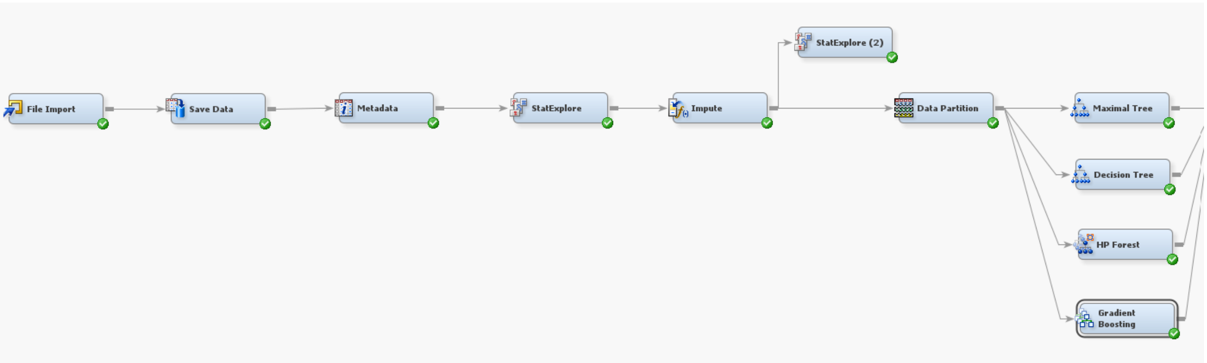
Based on Fit Statistics, misclassification rate is 0.09952 for training dataset and 0.1070 for validation dataset.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		ASE	Average Squared Error	0.073619	0.074995	.
Churn		DIV	Divisor for ASE	7878	3382	.
Churn		MAX	Maximum Absolute Error	0.954333	0.954333	.
Churn		NOBS	Sum of Frequencies	3939	1691	.
Churn		RASE	Root Average Squared Error	0.271329	0.273853	.
Churn		SSE	Sum of Squared Errors	579.9741	253.6341	.
Churn		DISF	Frequency of Classified Cases	3939	1691	.
Churn		MISC	Misclassification Rate	0.099518	0.107037	.
Churn		WRONG	Number of Wrong Classifications	392	181	.

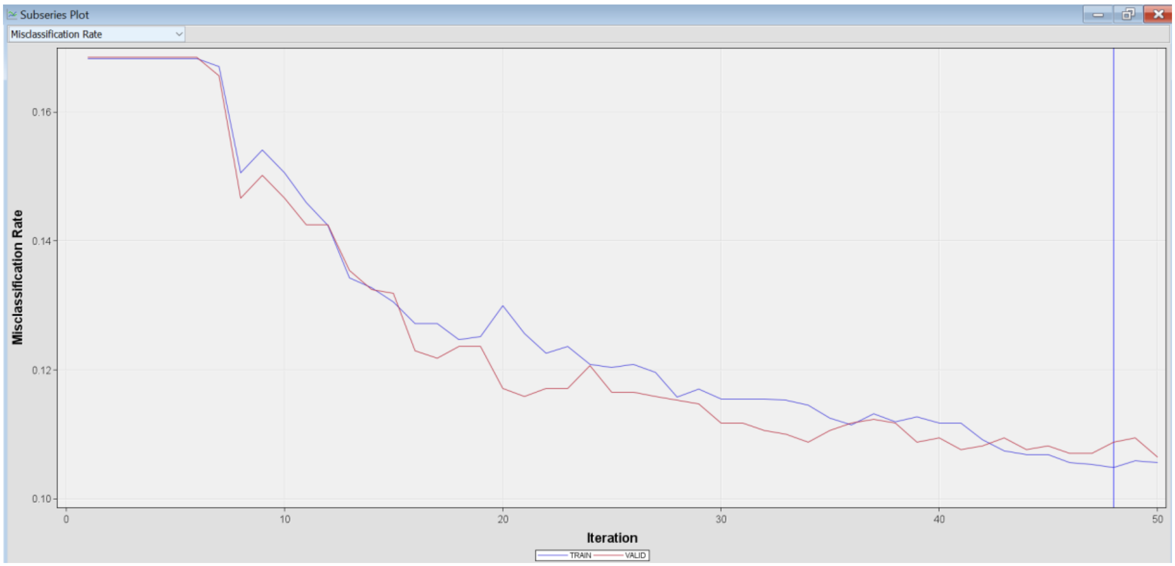
6.2 Boosting

Create a model for Boosting using "Gradient Boosting" node. Keep default settings.



Property	Value
General	
Node ID	Boost
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate Rules	0
Split Size	.
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Decision
Score	
Subseries	Best Assessment Value
Number of Iterations	1
Create H Statistic	No
Variable Selection	Yes
Report	
Observation Based Importance	No
Number Single Var Importance	5

Based on Iteration Plot, misclassification rate plateaued out at 48th iteration.



Based on Fit Statistics, misclassification rate is 0.1056 for training dataset and 0.1064 for validation dataset.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		NOBS	Sum of Frequencies	3939	1691	.
Churn		SUMW	Sum of Case Weights Times Freq	7878	3382	.
Churn		MISC	Misclassification Rate	0.105611	0.106446	.
Churn		MAX	Maximum Absolute Error	0.962012	0.965158	.
Churn		SSE	Sum of Squared Errors	616.1571	252.9251	.
Churn		ASE	Average Squared Error	0.078212	0.074786	.
Churn		RASE	Root Average Squared Error	0.279665	0.27347	.
Churn		DIV	Divisor for ASE	7878	3382	.
Churn		DFT	Total Degrees of Freedom	3939	.	.

6.3 Model Comparison

Compare model performance using "Model Comparison" node.

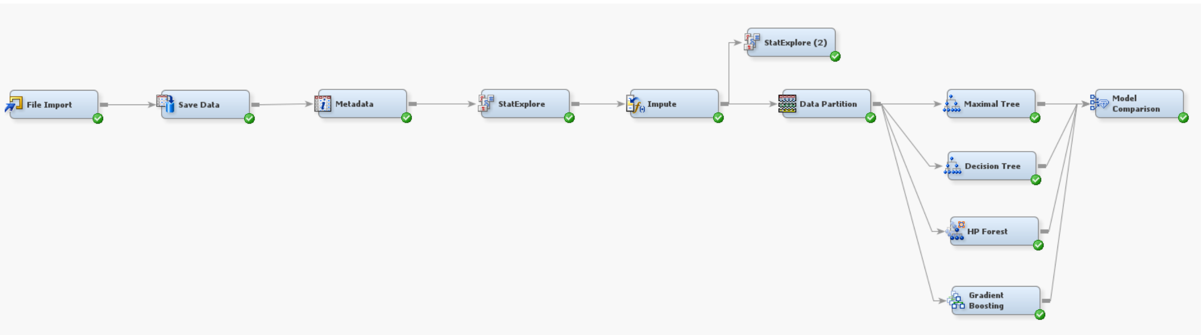


Figure below shows the Fit Statistics for model comparison.

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree2	Decision Tree	0.09698	0.089704	0.11094	0.082430
	Boost	Gradient Boosting	0.10645	0.078212	0.10561	0.074786
	HPDMForest	HP Forest	0.10704	0.073619	0.09952	0.074995
	Tree	Maximal Tree	0.10999	0.070517	0.09571	0.079098

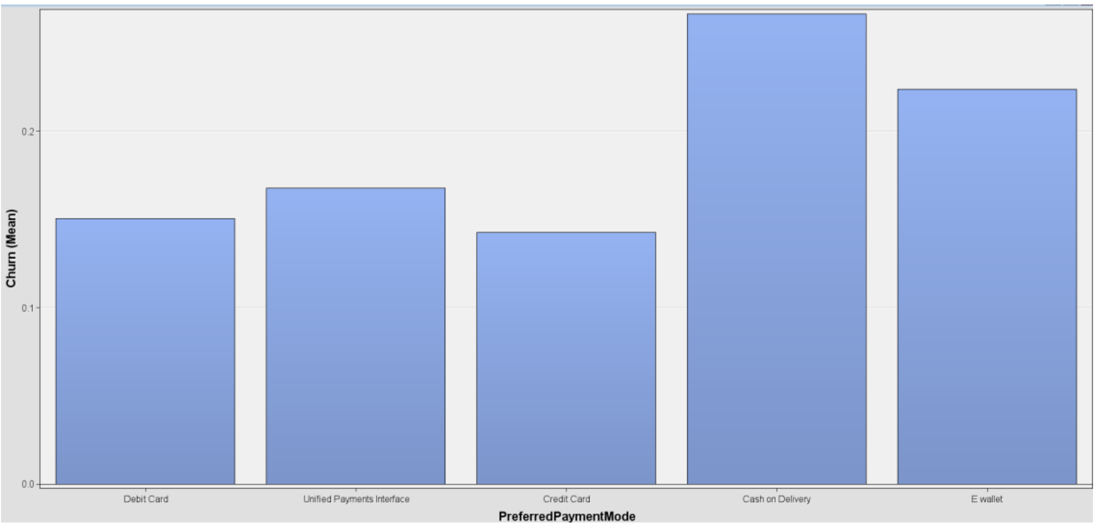
The selected model is pruned Decision Tree (Tree2 in the figure) with a validation misclassification rate of 0.09698 or 9.698%. Although the two ensemble methods helped reduce the training misclassification rate, ensemble methods resulted in higher validation misclassification rate than the pruned Decision Tree. This outcome is not uncommon and can be due to various reasons:

1. **Overfitting in Ensemble Methods:**
    - While ensemble methods (such as Random Forest or Gradient Boosting) aim to reduce overfitting, improper tuning or inadequate control over model complexity might lead to overfitting the training data. This can result in poorer performance on unseen validation data.
2. **Sensitivity to Hyperparameters:**
    - Ensemble methods often have multiple hyperparameters to tune (e.g., number of trees in Random Forest, learning rate in Gradient Boosting). Suboptimal hyperparameters can negatively affect model performance on validation data.

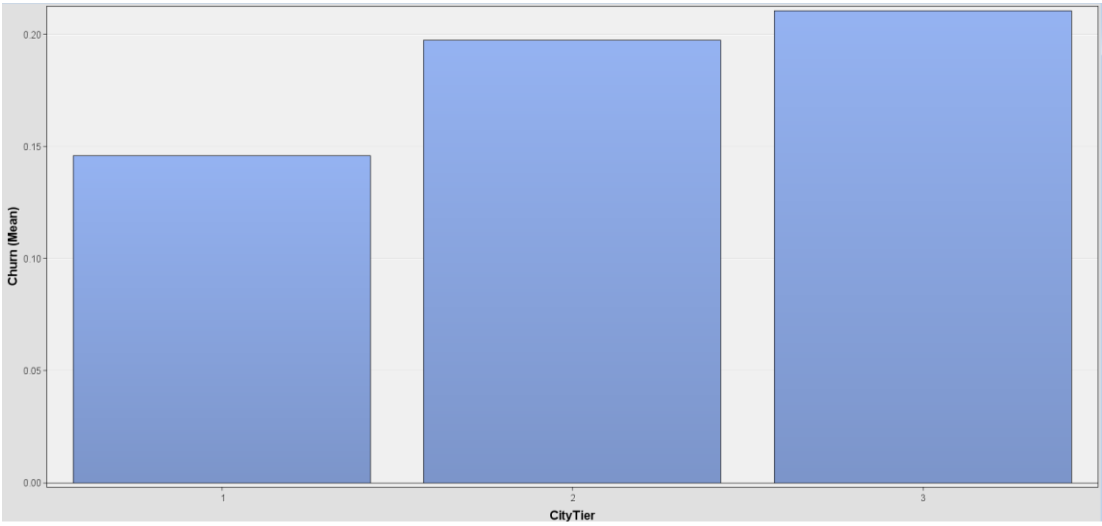
Nevertheless, the resulting difference is not significant (1 - 2%). It is fair to conclude that all the models including pruned Decision Tree, Random Forest and Gradient Boosting managed to deliver good classification accuracy, with small misclassification rate of 9.6 – 10.7%.

7.0 Insights into Customer Behaviour and Suggestions for Business Strategy

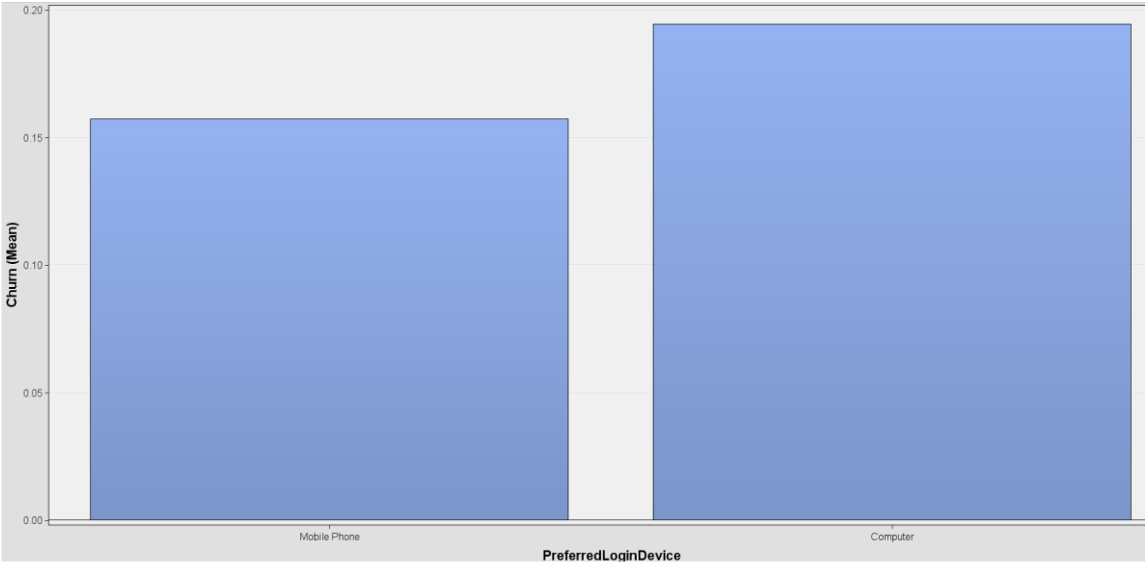
7.1 Insights into Customer Behaviour



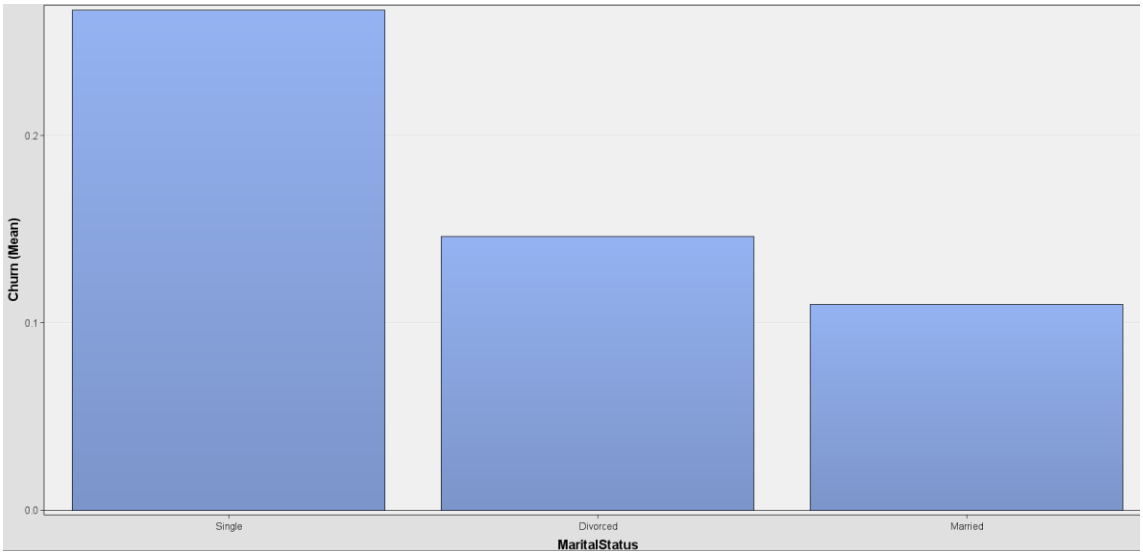
The data shows that customers using cash on delivery and e-wallet have higher-than-average churn rates. This suggests potential issues with these payment methods.



Tier 3 has a higher churn rate than Tier 2 and Tier 1. Also, both Tier 1 and Tier 2 cities still have a significant level of customer churn, which means that targeted strategies will be needed for each tier.



It is observed that customers who prefer to log in on a computer are more likely to churn than those who prefer to use a phone.



Single customers seem to show a higher churn rate than customers who are either married or divorced.

7.2 Suggestions for Business Strategy

Based on the extracted insights, below are some recommendations for business strategy.

- Investigate the reasons why COD, E-wallet and UPI customers are churning, evaluate the payment process and make improvements where needed. Explore new payment options and methods that may be more appealing to customers.

- The city tiers of their customers should be considered when developing strategies to reduce churn. Business should consider the demographics and purchasing power of their customers in different city tier to determine the optimal approach for reducing churn. Finally, it's important to consider the product categories being sold, as some products may be more likely to drive customer loyalty than others.

- Enhance the features and functionality available on the desktop version of the website. Also, improve the general user experience, usability and speed of both mobile phone and computer, to ensure that users have a seamless experience using their website.
- Offer personalized deals, services, or products to single customers based on their preferences and behaviour. Tailoring offerings to their interests can increase engagement and loyalty. Create loyalty programs or exclusive benefits to incentivize single customers to continue using the company's products or services.

8.0 Conclusion

In summary, Talend Data Preparation, Talend Data Integration and SAS Enterprise Miner were used to work with a dataset of customer transactions from an e-commerce website, encompassing various customer attributes and purchase history. Decision tree and ensemble methods were used to model the customer churn. Based on the results, it is concluded that all the models including pruned Decision Tree, Random Forest and Gradient Boosting managed to deliver good classification accuracy, with small misclassification rate of 9.6 – 10.7%. The selected model is pruned Decision Tree with a validation misclassification rate of 0.09698 or 9.698%. The two ensemble methods namely Random Forest and Gradient Boosting helped reduce the training misclassification rate but resulted in higher validation misclassification rate likely due to overfitting the training data. This can result in poorer performance on unseen validation data. Several insights were extracted through data mining, from which the company can tailor their business strategies to reduce customer churn rate.

9.0 GitHub Repository

Link: https://github.com/boon-kiat/customerchurn.git