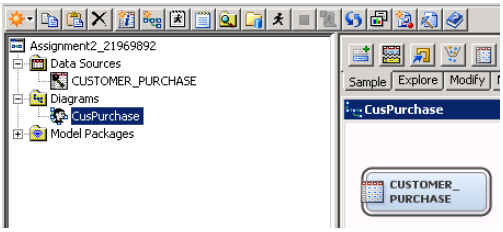
Contents

PART 1: PROJECT SETUP AND EXPLORATORY ANALYSIS

Part 1a: Create project and data source

The project and data source are created as below:



Part 1b: Model input justification

TargetAmt cannot be used as an input for the model predicting TargetBuy because of three main reasons:

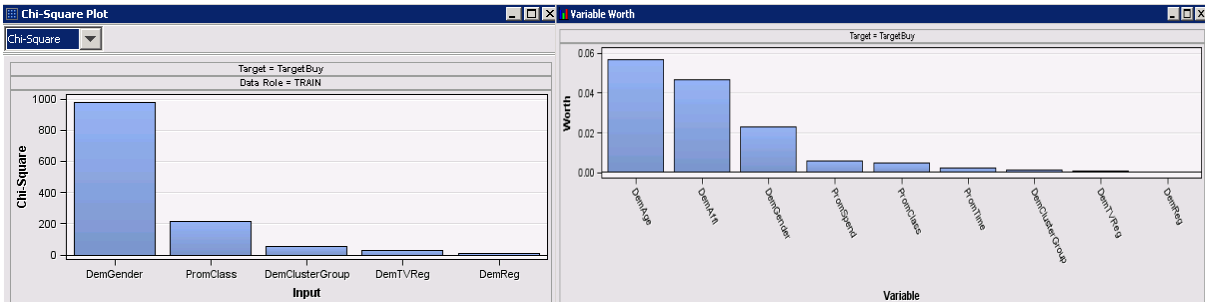
- Variable redundancy: Since TargetBuy (binary: 0 or 1) indicates whether a customer purchased, including TargetAmt (number of items bought) is redundant. If TargetBuy is 0, TargetAmt is 0, and vice versa.
- High correlation: TargetBuy and TargetAmt are highly correlated, which could distort model validity and lead to overfitting, as TargetBuy can be derived from TargetAmt.
- Data leakage: Using TargetAmt to predict TargetBuy introduces data leakage, as TargetAmt depends on TargetBuy, leading to biased results and poor model generalization.

Part 1c: Data exploratory analysis

Adding StatExplore node and connecting it to the data source gives the result as below:

The Chi-square plot below indicates the relationship between predictor variables and the target variable (i.e. TargetBuy) as can be seen, gender (i.e. DemGender) shows the strongest relationship with TargetBuy and hence, is strongly correlated with TargetBuy. This is followed by loyalty status (i.e. PromClass) which shows a moderate relationship with TargetBuy. DemClusterGroup, DemTVReg, and DemReg which have a lower Chi-square values show a weak relationship with the target variable.

Instead of looking into the relationship between explanatory variables and target variable, the Variable Worth plot reflects the importance of a variable in terms of its predictive power in the model and how much it improves the model's performance. Interestingly, DemAge and DemAffl show the highest worth, meaning that they both are the most valuable predictors of customer's likelihood to buy. Despite a strong relationship with TargetBuy, DemGender is ranked 3rd in its predictive power worth, which is followed by PromSpend, PromClass, PromTime, DemClusterGroup, DemTVReg, and DemReg.



The summary output from StatExplore also shows that there are a few missing values existing in both class and interval variables, which might require certain processing if regression model is built during later stages.

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	DemClusterGroup	INPUT	8	478	C	20.67	D	19.62
TRAIN	DemGender	INPUT	4	1691	F	55.08	M	26.04
TRAIN	DemReg	INPUT	6	327	South East	38.89	Midlands	30.51
TRAIN	DemTVReg	INPUT	14	327	London	27.87	Midlands	13.96
TRAIN	PromClass	INPUT	4	0	Silver	38.34	Tin	29.48
TRAIN	TargetBuy	TARGET	2	0	0	75.23	1	24.77

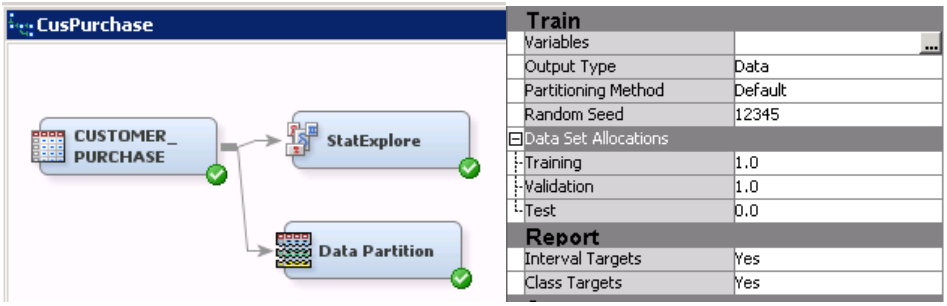
Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
DemAffl	INPUT	8.705986	3.428527	14785	772	0	8	31	0.86677	1.932328
DemAge	INPUT	53.83856	13.21682	14482	1075	18	54	79	-0.08505	-0.84541
PromSpend	INPUT	4338.93	6837.358	15557	0	0.01	2000	110072.4	3.898819	26.74589
PromTime	INPUT	6.556894	4.639008	15362	195	0	5	39	2.295868	8.246241

Part 1d: Data partition

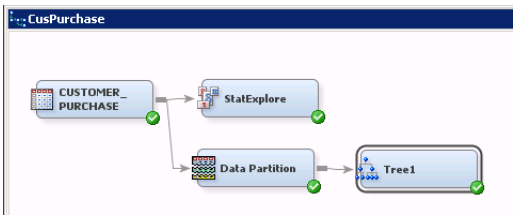
Creating the Partition node with setting Training and Validation under Data Set Allocations to 1.0 gives the split of 50% of the data for training and 50% for validation.

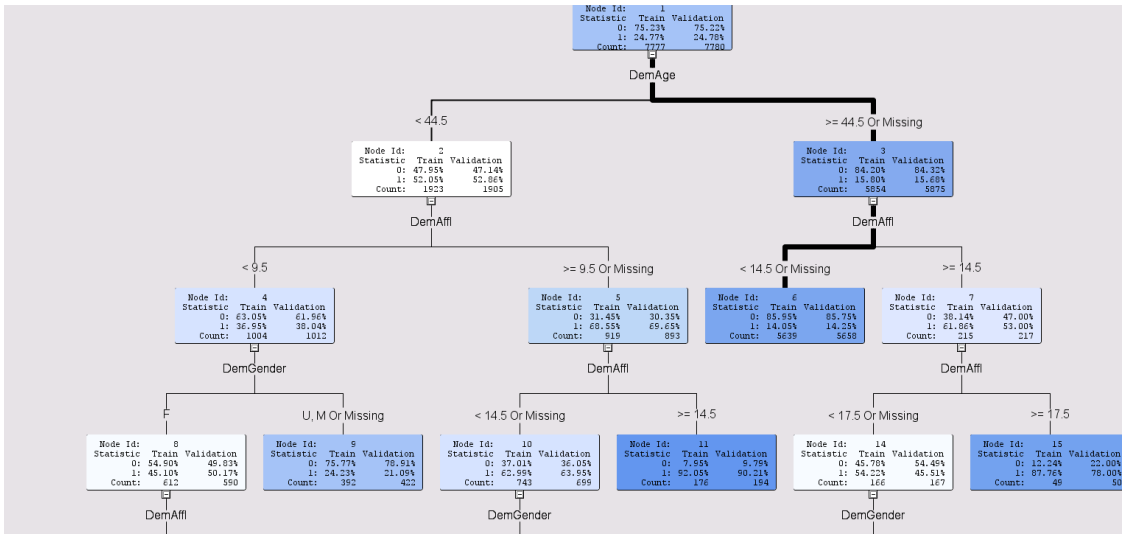


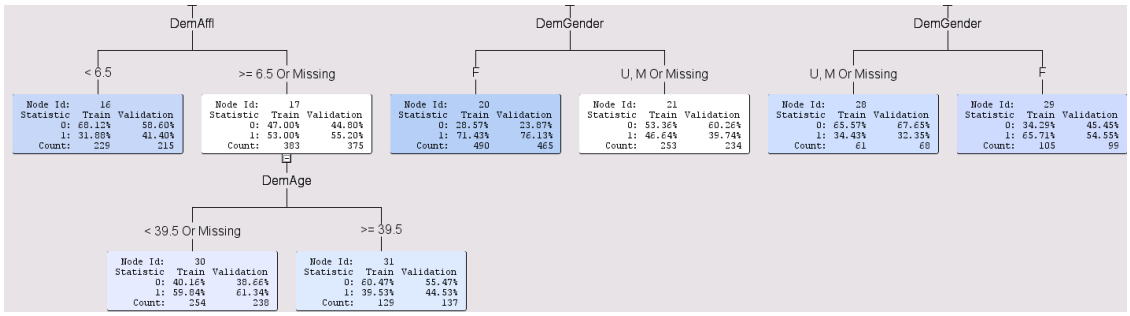
PART 2: DECISION TREE ANALYSIS

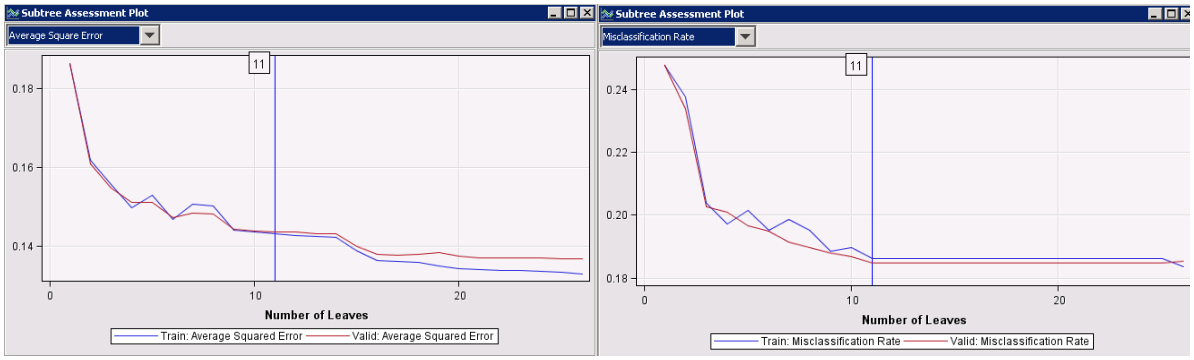
Part 2a: 2 decision trees

Tree 1: Decision tree node is added with keeping 'Decision' as the default option for Assessment Measure. From Tree 1 below, along with the subtree plots using Average Square Error and Misclassification Rate, the optimal number of leaves are 11 because more leaves do not improve the model's performance. DemAge is the variable used for the first split, breaking the tree into 2 branches with DemAge smaller than 44.5, and greater than or equal to 44.5 (or missing). This is supported by the result in part 1c, where DemAge is shown to have the highest variable worth and predictive power. The majority of the records are customers with ages being greater than or equal to 44.5 (or missing) and those with affluence grades smaller than 14.5 (or missing).

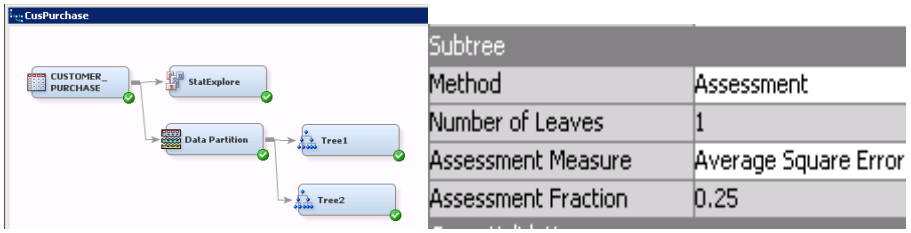




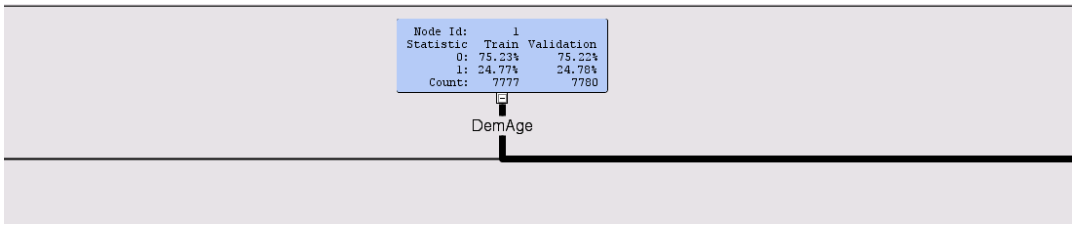


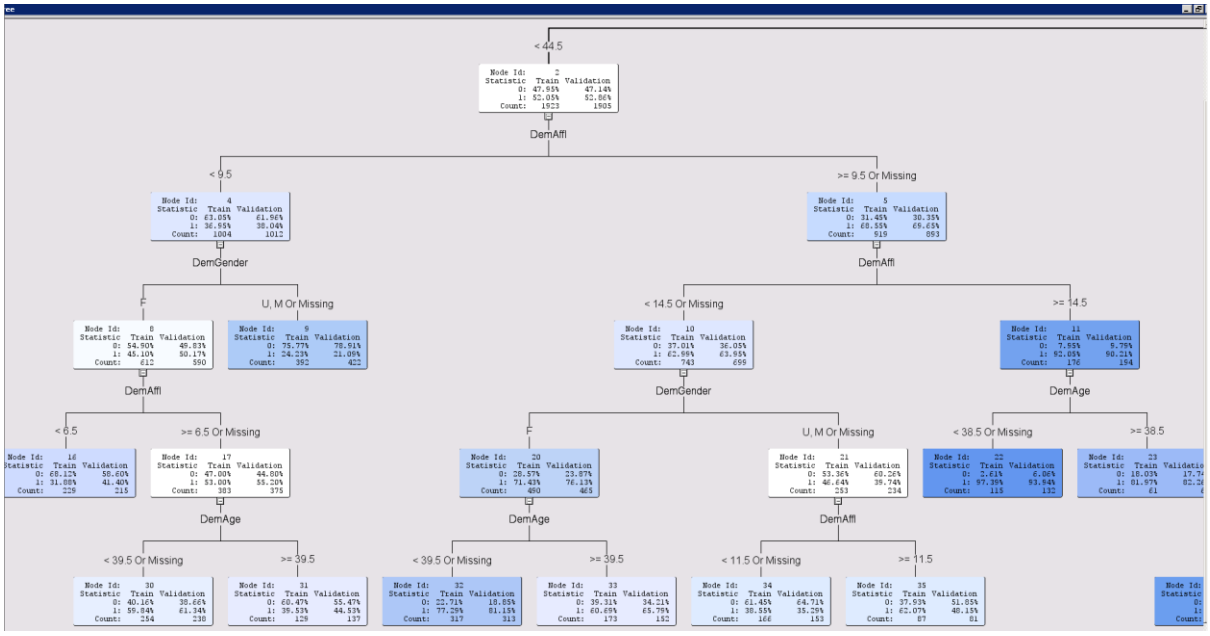


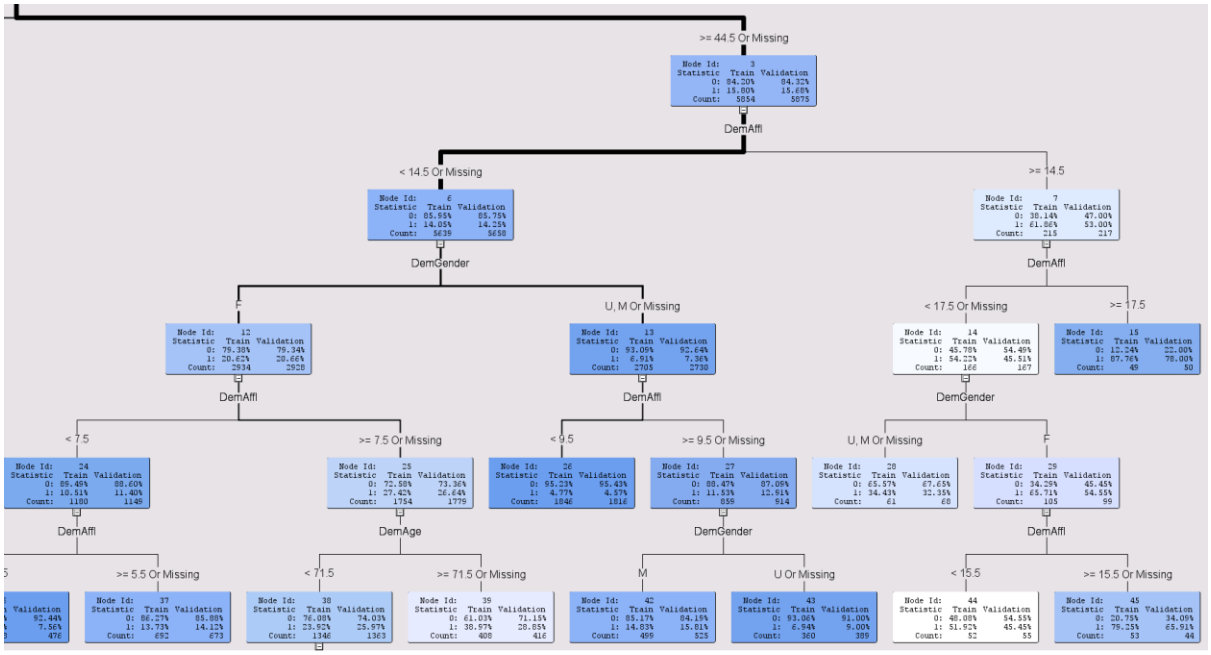
Tree 2: Decision tree node is added with using 'Average Square Error' as the selected option for Assessment Measure to test the performance of decision tree with different measurement technique. From Tree 2 as shown below, along with the subtree plots using Average Square Error and Misclassification Rate, the optimal number of leaves are 22 because more leaves do not improve the model's performance. DemAge is the variable being used for the first split, which is the same as Tree 1, breaking the tree into 2 branches with DemAge being smaller than 44.5, and greater than or equal to 44.5 (or missing). Similar to Tree 1, most of the records are customers with ages being greater than or equal to 44.5 (or missing) and those with affluence grades smaller than 14.5 (or missing).

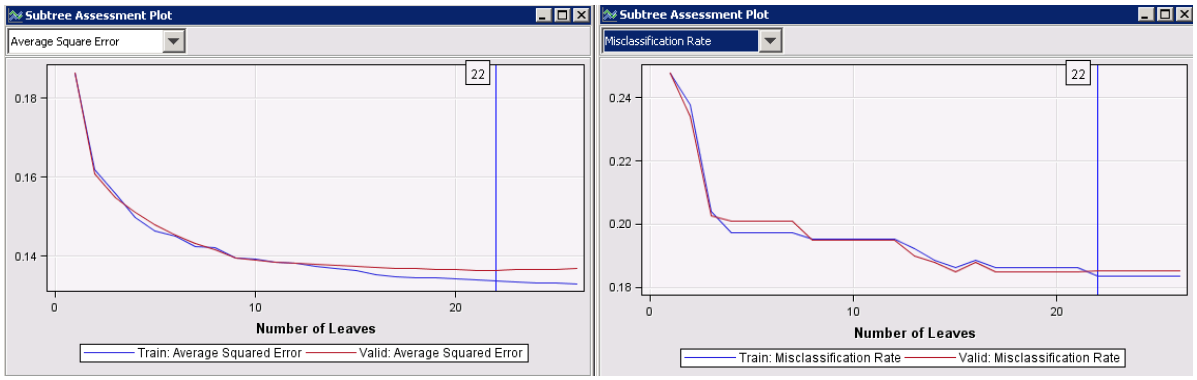


Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25

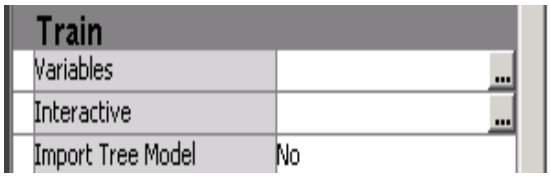


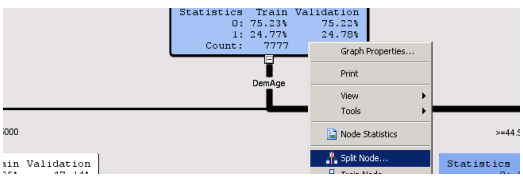






Tree 1 and 2: Competing splits are the alternative splits that could have been compared with the first split but not chosen because the first split shows a better performance. For both Tree 1 and Tree 2, the competing split is found by selecting Tree 1 and the ellipsis next to 'Interactive' in the Property panel. Then, right-clicking on the first node and selecting 'Split Node' option would give the Log value of the competing nodes. The competing splits are the same for both trees and as can be seen, DemAge is selected as the first split because it has the highest -Log(p), which indicates that this split is better at distinguishing the classes of TargetBuy in this case. The competing splits being considered but not chosen are: DemAffl, DemGender, PromSpend, PromClass. This result is consistent with the discussion of variable worth in part 1c.





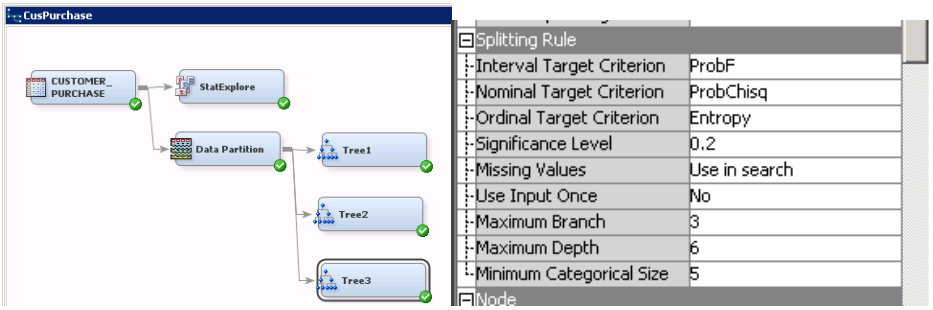
Split Node 1

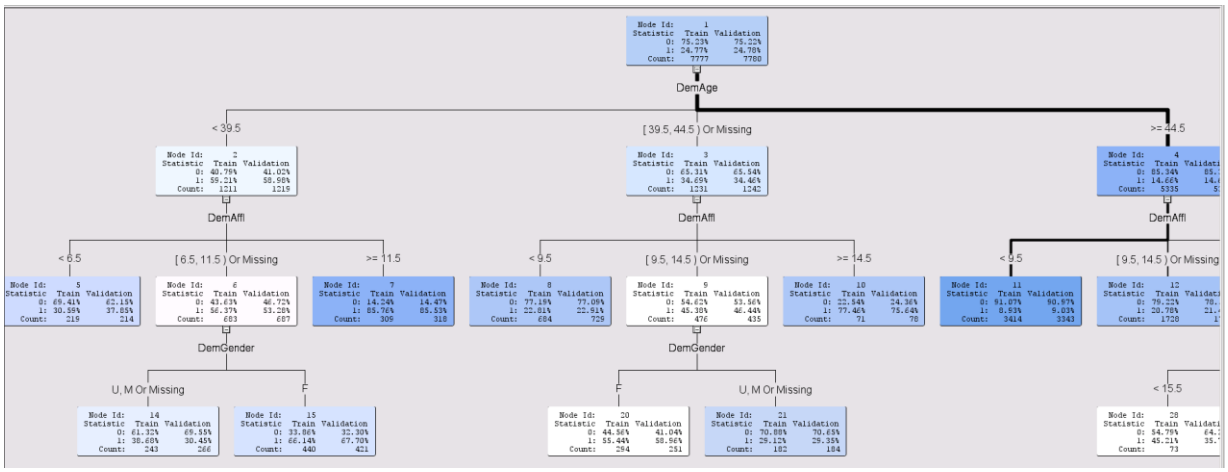
Target Variable: TargetBuy

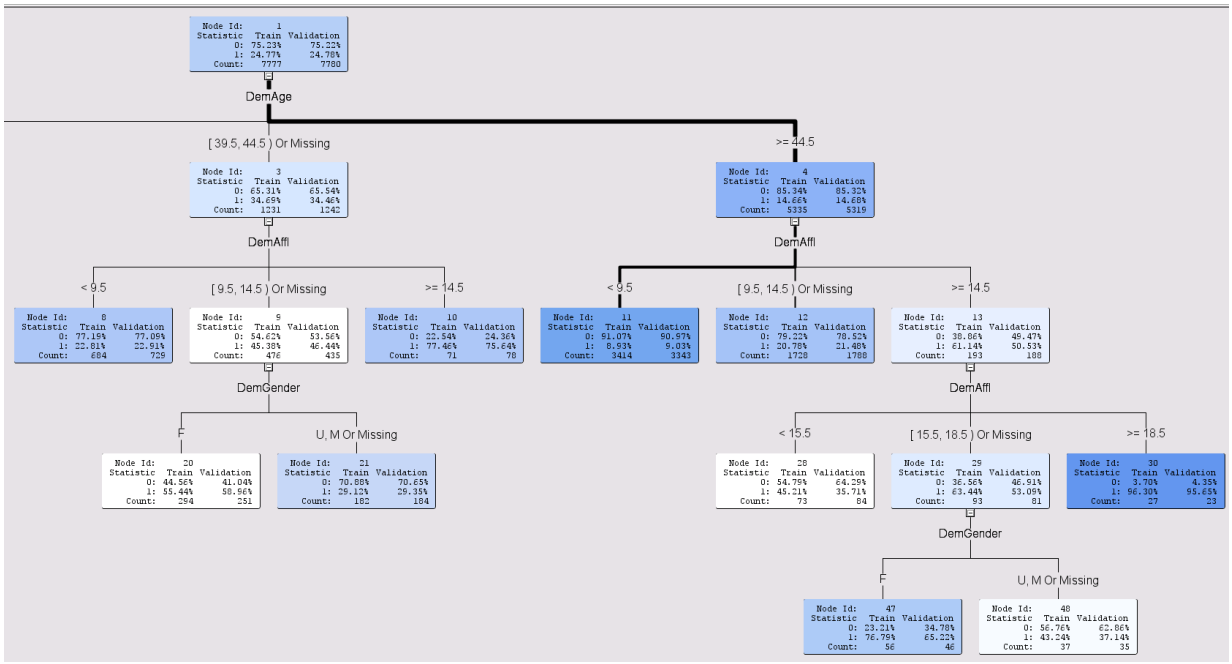
Variable	Variable Description	-Log(p)	Branches
DemAge	DemAge	221.5414	2
DemAffl	DemAffl	134.1687	2
DemGender	DemGender	89.9081	2
PromSpend	PromSpend	18.2721	2
PromClass	PromClass	14.7367	2

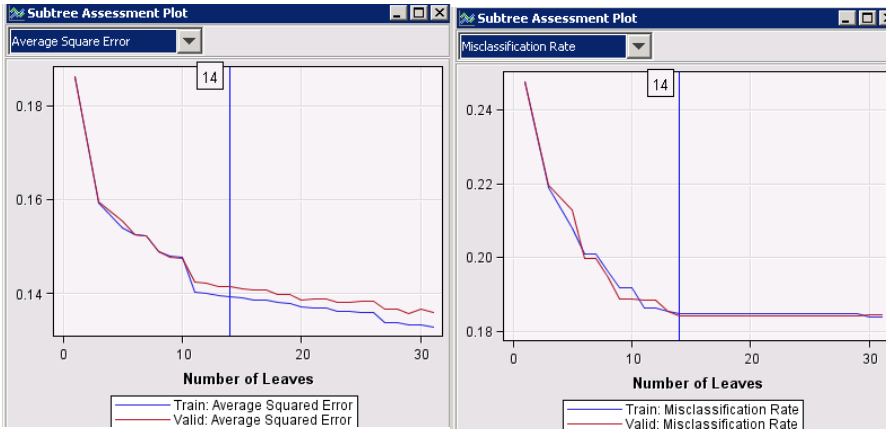
Part 2b: 3-branch decision tree

Tree 3: Decision tree node is added with keeping 'Decision' as the default option for Assessment Measure as the number of tree branches is to be changed. To allow 3-way splits, the 'Maximum Branch' row in under Splitting Rules should be inputted as '3'. From Tree 3 as shown below, along with the subtree plots using Average Square Error and Misclassification Rate, the optimal number of leaves are 14 because more leaves do not improve the model's performance. DemAge is still the variable being used for the first split, breaking the tree into 3 branches with DemAge being smaller than 39.5, between 39.5 to 44.5 (or missing), and greater than or equal to 44.5.



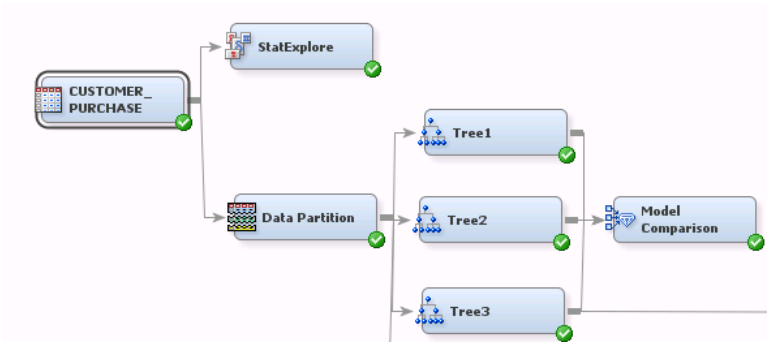






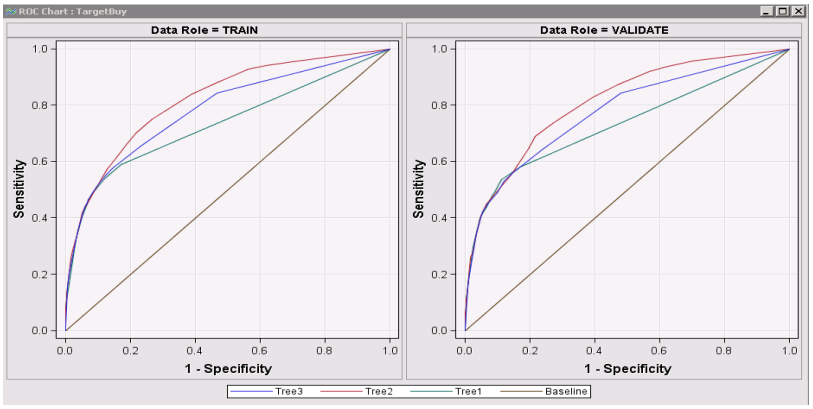
Part 2c: Model comparison

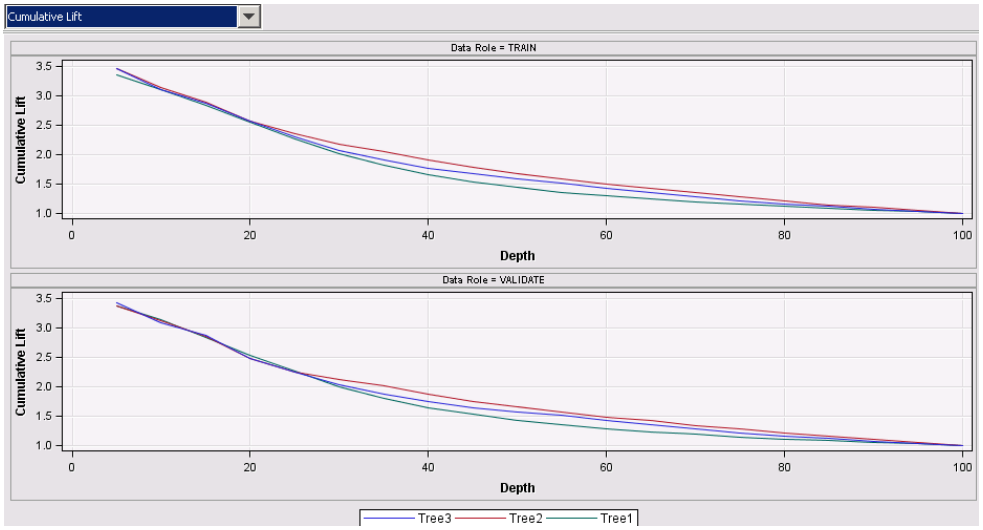
To compare the performance of 3 decision tree models, the 'Model Comparison' node is added and connected to Tree 1, Tree 2 and Tree 3.



The results are given with ROC chart, Cumulative Lift and Fit Statistics as below. Tree 3 appears to be better and hence, is the selected one because of three reasons:

- The use of 3 branches at the first split allows the Tree 3 to classify inputs in a less misclassification-prone way and hence, the misclassification rate of Tree 3 is the lowest of 0.18419.
- With the target variable (i.e. TargetBuy) being a binary variable, the 'Model Comparison' node in SAS selects Misclassification Rate as the preferred direct metric to evaluate the model performance with this binary target variable and hence, Tree 3 is then selected with the lowest misclassification rate.
- It is worth noting that ASE of Tree 2 is the smallest, and ROC and Cumulative lifts both indicate that Tree 2 shows a more favorable performance than Tree 3. However, tracing back to the model building step where model assessment statistic of Tree 2 is set as Average Square Error, it is understood that ASE method prioritizes minimization of errors in prediction and hence, breaks this tree down into lots of leaves (i.e. 22 leaves). Therefore, the model becomes more complex to reduce the prediction error, so ROC and Cumulative lifts show more favorable performance. With nearly doubled the number of leaves, the model is in fact, more prone to overfitting issue and hence, Tree 3 is still more preferred in this case.





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	Tree3	Tree3	0.18419	0.13935	0.18478	0.14141
	Tree	Tree1	0.18483	0.14310	0.18619	0.14348
	Tree2	Tree2	0.18522	0.13364	0.18349	0.13638

PART 3: REGRESSION MODEL ANALYSIS

Part 3a: Imputation decision

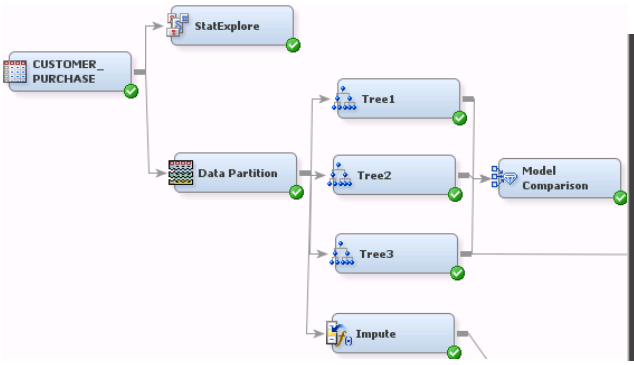
As shown in the summary statistic below, there are some missing values existing in the dataset. Before building regression model, the imputation of missing values is needed because they might produce biased results if they exist and are not being handled properly. As discussed in part 1c, there are some missing values in variables such as DemClusterGroup, DemGender, DemReg, DemTVReg, PromTime, DemAffl, and DemAge, which all need to be handled through imputation.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	Number of Levels	Percent Missing
DemAffl	Input	Interval	No		No	.	.	.	4.962396
DemAge	Input	Interval	No		No	.	.	.	6.910073
DemClusterGrou	Input	Nominal	No		No	.	.	7	3.072572
DemGender	Input	Nominal	No		No	.	.	3	10.8697
DemReg	Input	Nominal	No		No	.	.	5	2.101948
DemTVReg	Input	Nominal	No		No	.	.	13	2.101948
ID	ID	Interval	No		No	.	.	.	0
PromClass	Input	Nominal	No		No	.	.	4	0
PromSpend	Input	Interval	No		No	.	.	.	0
PromTime	Input	Interval	No		No	.	.	.	1.253455
TargetAmt	Rejected	Interval	No		No	.	.	4	0
TargetBuy	Target	Binary	No		No	.	.	2	0

For decision trees, imputation is not necessary because decision trees can handle missing values by itself through surrogating the splits with alternative variables or simply by ignoring the missing values.

Part 3b: Variable imputation

To do imputation, the 'Impute' node is added and connected after the 'Partition' node because this can help SAS use the imputation statistics derived from the training data and apply them consistently to the validation data, ensuring the two data sets are comparable. For class and interval variables, the imputed values are 'U' and overall means respectively.



Class Variables	
Default Input Method	Default Constant Value
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	U
Default Number Value	.
Method Options	

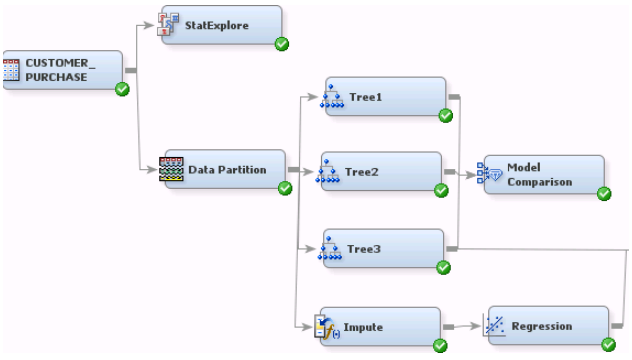
Imputation indicators are created for the imputed inputs as below:

Imputation Summary

Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
DemAffl	MEAN	IMP_DemAffl	M_DemAffl	8.6824580401	INPUT	INTERVAL		389
DemAge	MEAN	IMP_DemAge	M_DemAge	53.884127859	INPUT	INTERVAL		519
DemClusterGroup	CONSTANT	IMP_DemClusterGroup	M_DemClusterGroup	U	INPUT	NOMINAL		250
DemGender	CONSTANT	IMP_DemGender	M_DemGender	U	INPUT	NOMINAL		863
DemReg	CONSTANT	IMP_DemReg	M_DemReg	U	INPUT	NOMINAL		156
DemTVReg	CONSTANT	IMP_DemTVReg	M_DemTVReg	U	INPUT	NOMINAL		156
PromTime	MEAN	IMP_PromTime	M_PromTime	6.5462950905	INPUT	INTERVAL		98

Part 3c: Regression model

The regression model is created by adding 'Regression' node following the 'Impute' node' one. In the 'Regression' node's property, setting selection model as 'Stepwise' and selection criterion as 'Validation Error' gives the model as below:



Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes

Part 3d: Regression analysis

As shown in the output below, the selected model is model in step 3 and the variables included in the final model are IMP_DemAffl (i.e. affluence grades), IMP_DemAge (i.e. customer ages), IMP_Gender (gender).

The selected model, based on the error rate for the validation data, is the model trained in Step 3. It consists of the following effects:

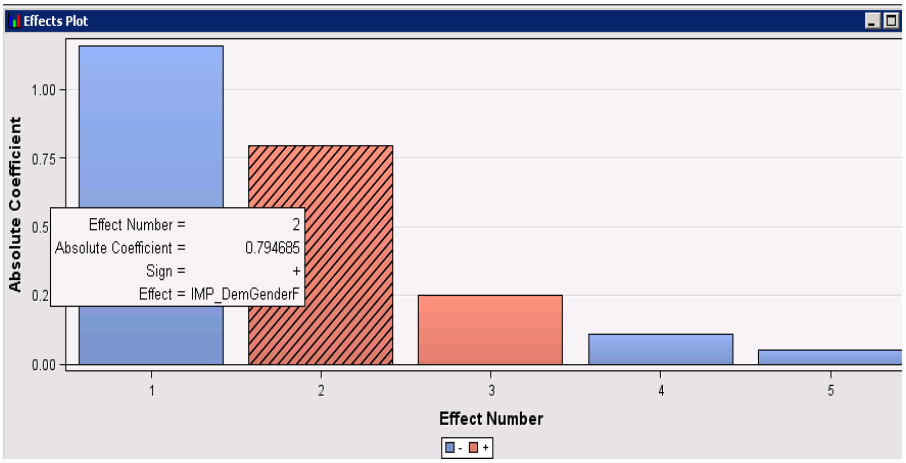
Intercept IMP_DemAffl IMP_DemAge IMP_DemGender

Analysis of Maximum Likelihood Estimates

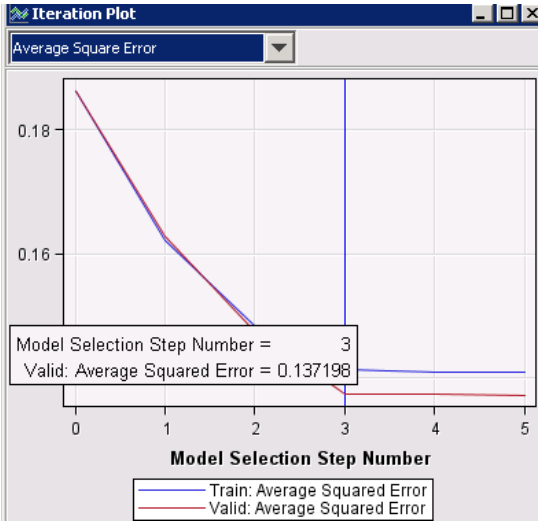
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-1.1621	0.1588	53.55	<.0001		0.313
IMP_DemAffl		1	0.2488	0.00996	623.76	<.0001	0.4583	1.283
IMP_DemAge		1	-0.0506	0.00247	417.99	<.0001	-0.3550	0.951
IMP_DemGender	F	1	0.7947	0.0453	307.94	<.0001		2.214
IMP_DemGender	M	1	-0.1092	0.0551	3.92	0.0476		0.897

In the context of supermarket management, this means that the likelihood of purchasing by customers is strongly affected by affluence grade, ages and genders. With positive estimate values, customers with higher affluence grade or being identified as female likely to make the purchase, so the supermarket management should look out for targeted marketing strategies aimed at affluent female customers to promote the new clothing line Conversely, customers with higher age or being identified as males less likely to make the purchase, indicating that the management may need to consider alternative product lines or marketing approaches for these segments.

The importance of each variable in the model can be evaluated with Effect Plot below and the magnitude of the coefficient estimates above. As can be seen, the variables to the left are having the highest rankings in terms of the absolute magnitude of their impact on predicting TargetBuy. The colors of the bars represent the signs of the relationship between a particular variable and target variable, with the blue one being negative and the red one being positive. In this case, besides Intercept (where all predictors are 0), IMP_GenderF is the most important variable due to the highest value of absolute coefficient (i.e. 0.795) and exponential estimate of 2.214. A positive sign of IMP_GenderF indicates that if the customer is identified as Female, they are 2.214 times more likely to make the purchase (TargetBuy =1) than other gender groups. Similarly, the 3rd rank belongs to IMP_DemAffl with exponential estimate of 1.283, meaning that 1 point increase in affluence grade can increase the likelihood of purchasing by customers by 1.283 times. IMP_GenderM is ranked 4th with estimate of -0.109, meaning that if the customer is identified as Male, they are 0.897 times less likely to make the purchase than other gender groups. Finally, the last ranked variable is DemAge with estimate of -0.05, implying that an additional year of age will reduce the likelihood of purchase by 0.951 times. To sum up, the higher the effect number of variables with transformed values, the more important that variable is to the regression model.



Creating an Iteration plot as below shows the value of average squared errors (ASE) for both training and validation sets of the model. In this case, the smallest validation ASE occurs in model step 3 and this model is selected. The valid ASE is 0.137198, meaning that this is the point where no further improvement can be made to the model for achieving lower ASE and that the model 3 reaches the optimal performance level for making prediction for the target.



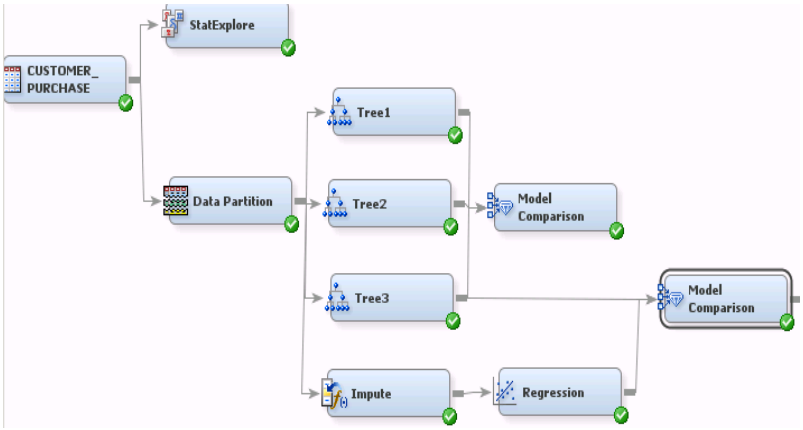
PART 4: MODEL COMPARISON AND SCORING

Part 4a: Decision tree and regression comparison

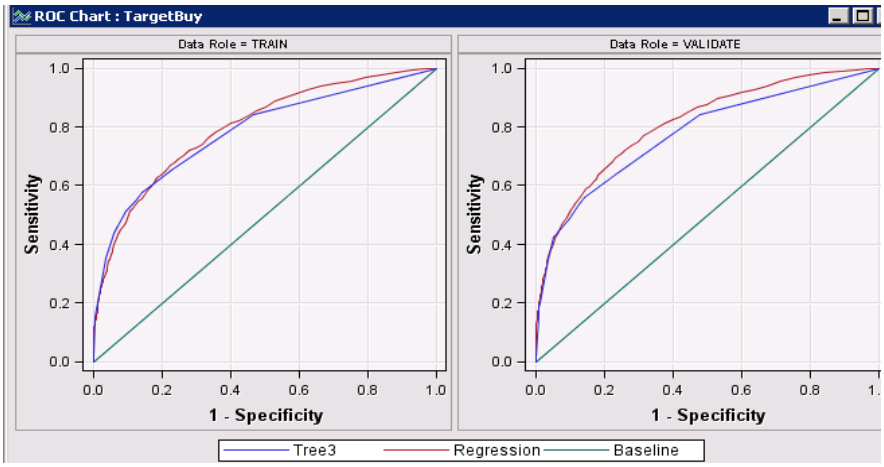
Based on the results from part 2 and 3, the first difference between decision tree and regression analysis is the determination of important variables:

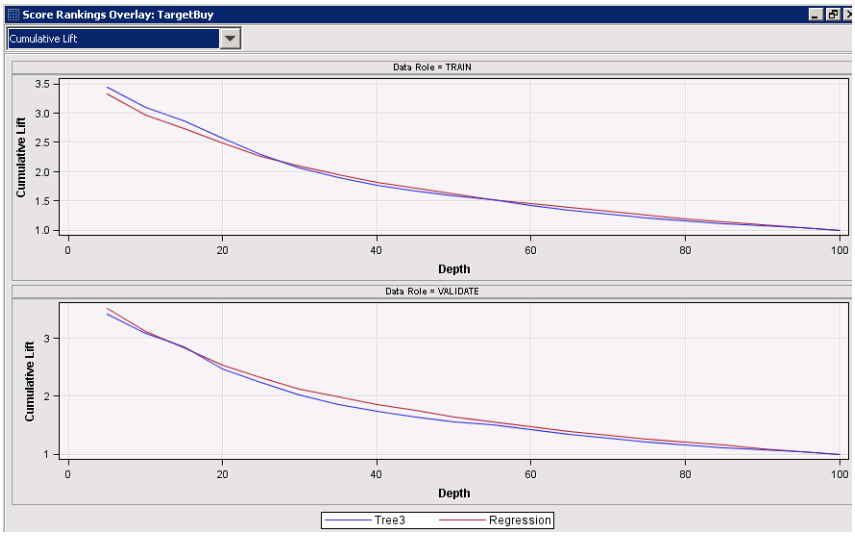
- Decision tree considers the variable in the first split as the most important variable in a non-linear manner and with the case of Tree 3, Age is the first split which is regarded as the most important variable.
- Regression determines important variables based on the size of the absolute coefficient estimates and, in this case, with a logistic relationship established, regression model regards Gender as the most important variable.

To further contrast and compare the performance of decision tree and regression models, the 'Model Comparison' node is added and connected to Regression and Tree 3, as shown below. Since Tree 3 performs better than Tree 1 and 2 as explained in part 2c, they both can be disregarded from this comparison.



As shown below, ROC chart and Cumulative lifts show that Regression model performs better than Tree 3. Also, the validation ASE of Regression (0.13720) is also lower than that of Tree 3 (0.14141), indicating the highest accuracy in predicting the TargetBuy variable, which can be an expected outcome as the regression model's natural algorithm is to minimize the prediction error. However, with the target variable (i.e. TargetBuy) being a binary variable, the 'Model Comparison' node in SAS selects Misclassification Rate as the preferred direct metric to evaluate the model performance with this binary target variable and hence, Tree 3 is then selected with the lower valid misclassification rate (0.18419), compared to the one of regression (0.18676). Extending this finding further to the business use case context, if the goal is to enhance predictive performance to identify customers, regression can be more tangible in attaining that goal. However, in this case, decision tree (Tree 3) is more suitable to use for customer classification and targeting.





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

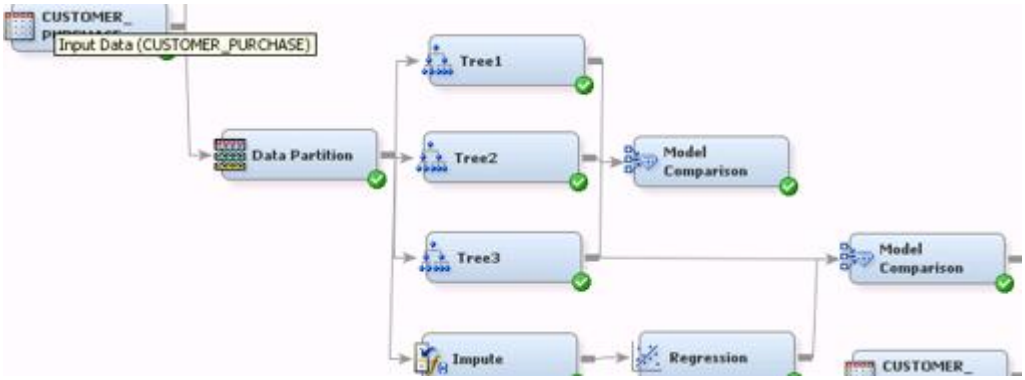
Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree3	Tree3	0.18419	0.13935	0.18478	0.14141
	Reg	Regression	0.18676	0.14126	0.19661	0.13720

Part 4b: Modeling techniques

It would not be sufficient to solely use decision tree or regression because each modeling technique will have its own strengths and weaknesses that might lead to different results in certain performance metrics.

- Decision tree: A non-parametric model splitting data based on conditions to predict the target variable (i.e. TargetBuy), which works well with missing values and classification, but being prone to overfitting issues.
- Regression: A parametric model that estimates the relationship between the predictor variables and TargetBuy using logistic method, which works well in minimizing the prediction errors and estimating the magnitude of the relationships, but might not be suitable for dataset with missing values and classification task

That being said, based on the data's complexity and the underlying algorithm, each model might give a different result and metrics such as ASE, misclassification, etc. as seen in part 4a. Using the Model Comparison node, as shown below, it is necessary to perform 2 models and compare the results across both.



Sometimes, the outcome metrics might give a contradiction such as in this case, because each model is having its own strengths and weaknesses. For example, in the case as shown below, if the preferred metric is Valid ASE, Regression model might be more preferred than Decision Tree. On the other hand, as Valid Misclassification Rate is the preferred metric, Decision tree is chosen. This means that comparing different models will give better insights into the performance and depending on the area of interest and the dataset, one model can be selected as more preferred than the other.

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	Tree3	Tree3	0.18419	0.13935	0.18478	0.14141
	Reg	Regression	0.18676	0.14126	0.19661	0.13720

Part 4c: Model advantages

The advantages of decision trees include but are not limited to:

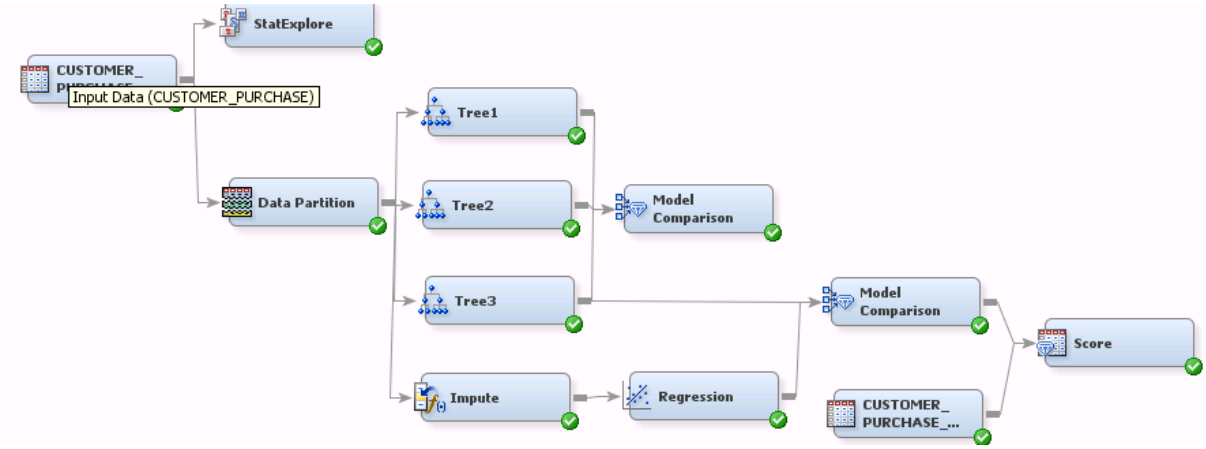
- Interpretability and variable importance: the structure of decision tree makes it easy to understand and interpret how the final decision is made through the thresholds set by different explanatory variables. Most importantly, decision tree makes make it clear that the variable being used in the first split is the most important one. For example, in the three decision trees in part 2, Age is the first split and hence, is the most important variable in all trees.
- Non-linear relationship requirement: decision tree allows prediction to be made without the need to assume any linear relationships between the explanatory variables and the target ones. In this case, no linear relationship assumptions need to be investigated between TargetBuy and explanatory variables.
- Less data cleaning effort: decision tree is less likely to be affected by outliers and more importantly, as mentioned in part 3a, decision trees innately handle the missing values itself and no imputation is required, leading to lower effort needed to clean data before using the model.

The advantages of regression models include but are not limited to:

- Quantitative measure of variable importance: the relationship between the target variable and the explanatory variables are quantified and modelized through the coefficient estimates. This allows a quantification of the relationship and based on the coefficient estimates, important variables as well as the signs of the relationships can be determined. For example, in part 3d, the coefficient estimates of Gender and Affluence inform the changes in the likelihood of TargetBuy variable.
- Handle multicollinearity: the usage of regression model requires the check for multicollinearity and dimension reduction if applicable, helping to handle this issue. For example, in part 3c, before creating regression model, ‘Variable clustering’ node might need to be added to investigate multicollinearity and improve the model’s stability.

Part 4d: Model scoring

To score the best model, the dataset Customer_Purchase_Score is added to the diagram. The 'Score' node is also added and connected to the score dataset and 'Model Comparison' node. The final diagram is as follows:



After running the Score node, selecting ellipsi next to the 'Exported Data' in the property pane and choosing score data. Clicking 'Explore', selecting the 'Plot' option with Histogram and setting P_TargetBuy1 as X will give a histogram as below.

The histogram showcases the distribution of predicted probability of purchasing from the new product line by customers (i.e. TargetBuy =1). As can be seen, the frequencies of low probabilities dominate with lots of tall bars to the left, implying that the probability of purchasing by most customers is relatively low, ranging from 0.0893 to 0.2641. To the highlighted right side of the histogram, there are some customers showing a high likelihood of making a purchase, ranging from 0.7009 to 0.9630. Bringing this insight back to the business context, this indicates that an opportunity for the department store management to take targeted marketing actions and selling effort (i.e. offers, discounts) to convert these high potential customers. For the remaining customers who show less interest in buying, the store management might also want to take a closer look at the areas for improvement or what these customers are looking for, since they account for a large amount of customer base.