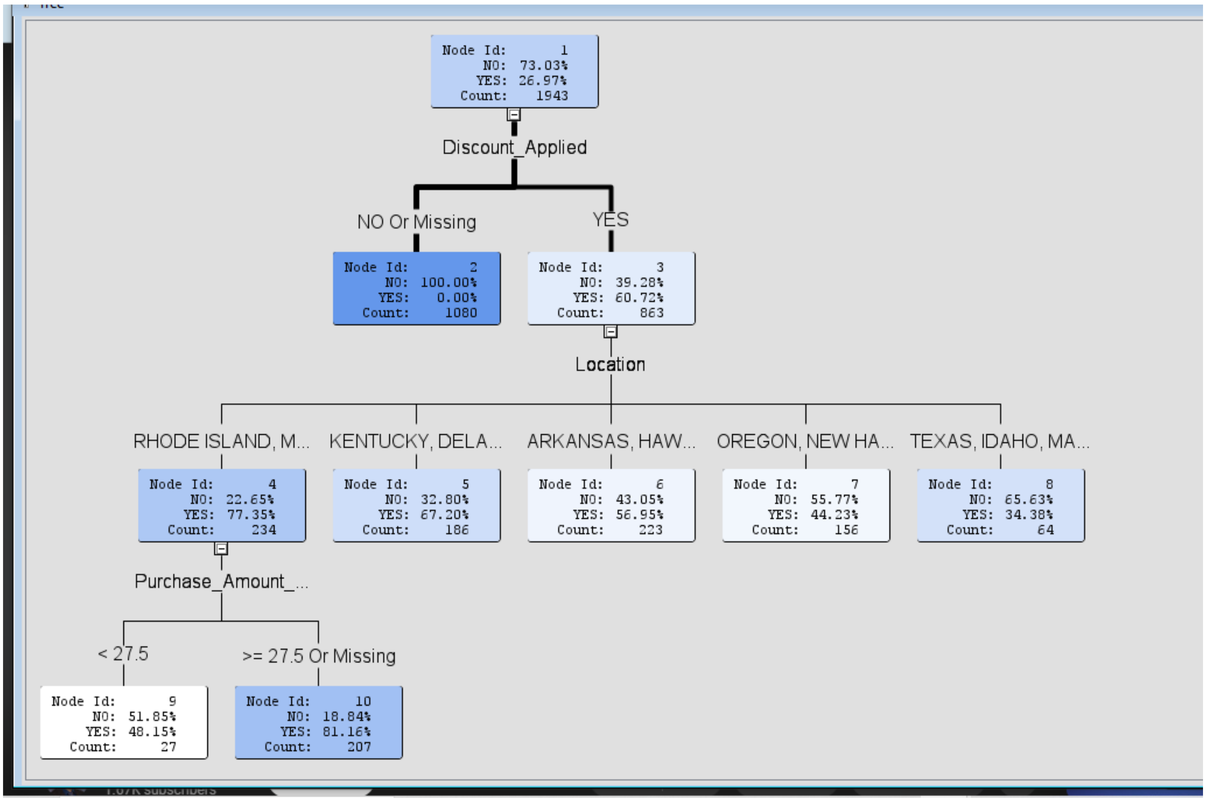
WQD7005 AA1 – Nur Aisyah Yusof (22072845)

Results and analysis

Decision Tree

A decision tree is a supervised machine learning algorithm used for both classification and regression tasks. Its main function is to make decisions based on a set of rules derived from the input features. The decision tree model organizes these rules in a tree-like structure, where each internal node represents a decision based on a feature, each branch represents the outcome of the decision, and each leaf node represents the final prediction or classification.

For decision tree analysis with 50:50 data partition, the leaf node output shown as below: -



The root node is the subscription status which is a binary classification whether yes or no. The train data has 1943 data count with 73.03% count of not subscribe the ecommerce platform and 26.97% count that subscribe the platform. If a discount were not applied, 100% of not subscribe count from the root node will not subscribe the platform with 1080 count. The leaf node is known as decision node. Meanwhile, if a discount were applied, 60.72% from the 26.97% from the root node will subscribe to the online shopping platform. From the interior node, location will be one of the features important in decision-making whether to subscribe or not. Those from Rhode Island were also affected by the purchase amount in the train prediction modelling where 14 count will subscribe if the amount more than 27.5 USD and 168 count will subscribe if the purchase amount less than 27.5 USD.

```
Event Classification Table

Data Role=TRAIN Target=Subscription_Status Target Label=' '

  False      True       False      True
Negative   Negative   Positive   Positive

   104       1223        196        420
```

From the classification table, the confusion matrix as shown below:

	Real - Positive	Real - Negative
Predicted - Positive	420	196
Predicted - Negative	104	1223

From here, the model has a quite good performance matrix, which the value as below: -

Accuracy: 84.5%

Precision: 68.1%

Recall: 80.2%

F1-Score: 73.6%

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Subscription_Status		_NOBS_	Sum of Frequencies	1943	.	1945
Subscription_Status		_MISC_	Misclassification Rate	0.1544	.	0.17635
Subscription_Status		_MAX_	Maximum Absolute Error	0.811594	.	0.811594
Subscription_Status		_SSE_	Sum of Squared Errors	373.9569	.	407.1852
Subscription_Status		_ASE_	Average Squared Error	0.096232	.	0.104675
Subscription_Status		_RASE_	Root Average Squared Error	0.310213	.	0.323535
Subscription_Status		_DIV_	Divisor for ASE	3886	.	3890
Subscription_Status		_DFT_	Total Degrees of Freedom	1943	.	.

From the fit statistic, the misclassification is 0.1544 for train model and 0.1764 for the test model. The Average Square Error for train model is 0.096 and 0.105 for the test model. Overfitting occurs when a model performs well on the training data but fails to generalize to new, unseen data. The misclassification and the average square error for both models is almost the same and low which indicates there is no overfitting in the model and the model quite good.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance
Discount_Applied		1	1.0000
Location		1	0.3032
Purchase_Amount__USD		1	0.1213

From here, the feature importance for subscription status is discount applied with 1.0 importance, location with 0.303 importance and purchase amount in USD with 0.1213 importance value. Those three features could be indicated as the variable that will affect the Status Subscription Prediction model.

From the analysis, it is shown that the decision tree shown in the model has quite a good performance. However, as ensemble method is said can improve the performance of the model, random forest modelling is used to compare both model performance.

Ensemble Method – Random Forest

Random Forest is an ensemble learning method, meaning it builds multiple individual models and combines their predictions. The base models in a Random Forest are decision trees. Decision trees are constructed by recursively splitting the data based on feature conditions. Each tree in a Random Forest is trained on a random subset of features at each split. This introduces diversity among the trees.

Here are the results and analysis from SAS EM: -

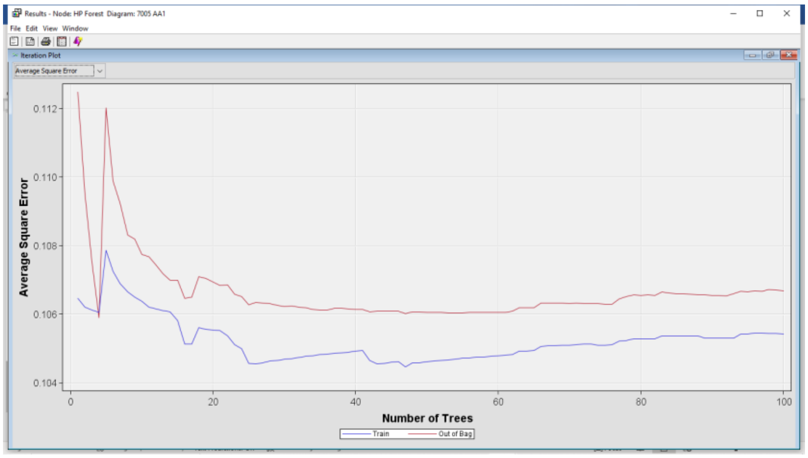
Fit Statistics

Target=Subscription_Status Target Label=' '

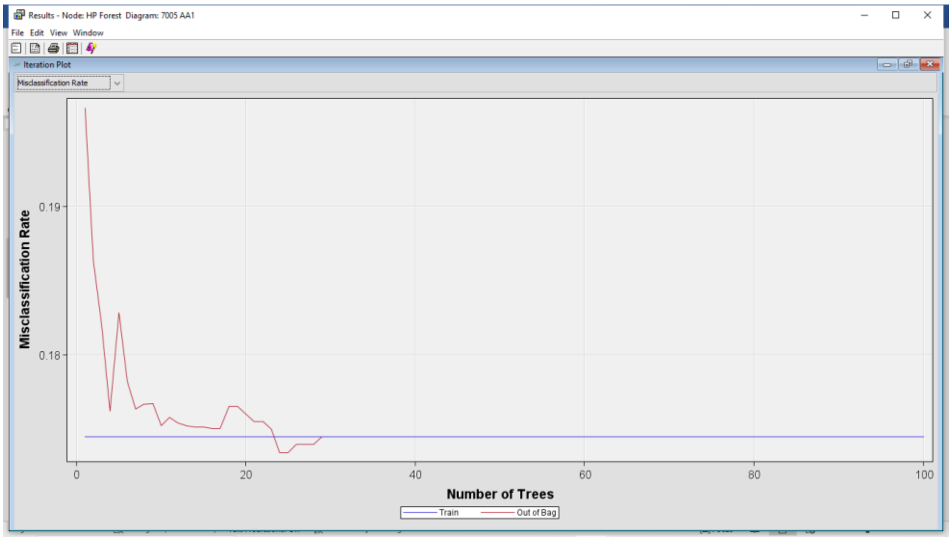
Fit Statistics	Statistics Label	Train	Test
ASE	Average Squared Error	0.11	0.10
DIV	Divisor for ASE	3886.00	3890.00
MAX	Maximum Absolute Error	0.60	0.61
NOBS	Sum of Frequencies	1943.00	1945.00
RASE	Root Average Squared Error	0.32	0.31
SSE	Sum of Squared Errors	409.72	376.47
DISF	Frequency of Classified Cases	1943.00	1945.00
MISC	Misclassification Rate	0.17	0.15
WRONG	Number of Wrong Classifications	339.00	284.00

From the fit statistic, the misclassification is 0.17 for train model and 0.15 for the test model. The Average Square Error for train model is 0.11 and 0.10 for the test model. The misclassification and the average square error for both also indicates there is no overfitting in the model and the model is quite good too.

The out-of-bag (OOB) error or accuracy graph is a useful tool to interpret the performance of a Random Forest model during training. The OOB error is an estimate of how well the model is likely to perform on new, unseen data. In a Random Forest, the OOB error is computed using the out-of-bag samples, which are instances not included in the bootstrapped sample used to train each individual tree.



From the OOB error of average square root, the graph trend is decreasing or stabilizing as the number of trees increases and follow the train model as well. This indicates that the ensemble is learning and improving its ability to generalize to new data.



From the OOB error of misclassification rate, the graph trend is also decreasing or stabilizing as the number of trees increases. This indicates no sign of overfitting from the model.

```
Event Classification Table

Data Role=TRAIN Target=Subscription_Status Target Label=' '

  False       True        False       True
Negative    Negative    Positive    Positive

    0         1080         339         524
```

From the classification table, the confusion matrix as shown below:

	Real - Positive	Real - Negative
Predicted - Positive	524	339
Predicted - Negative	0	1080

Accuracy: 82.5%

Precision: 60.7%

Recall: 100%

F1-Score: 75.5%

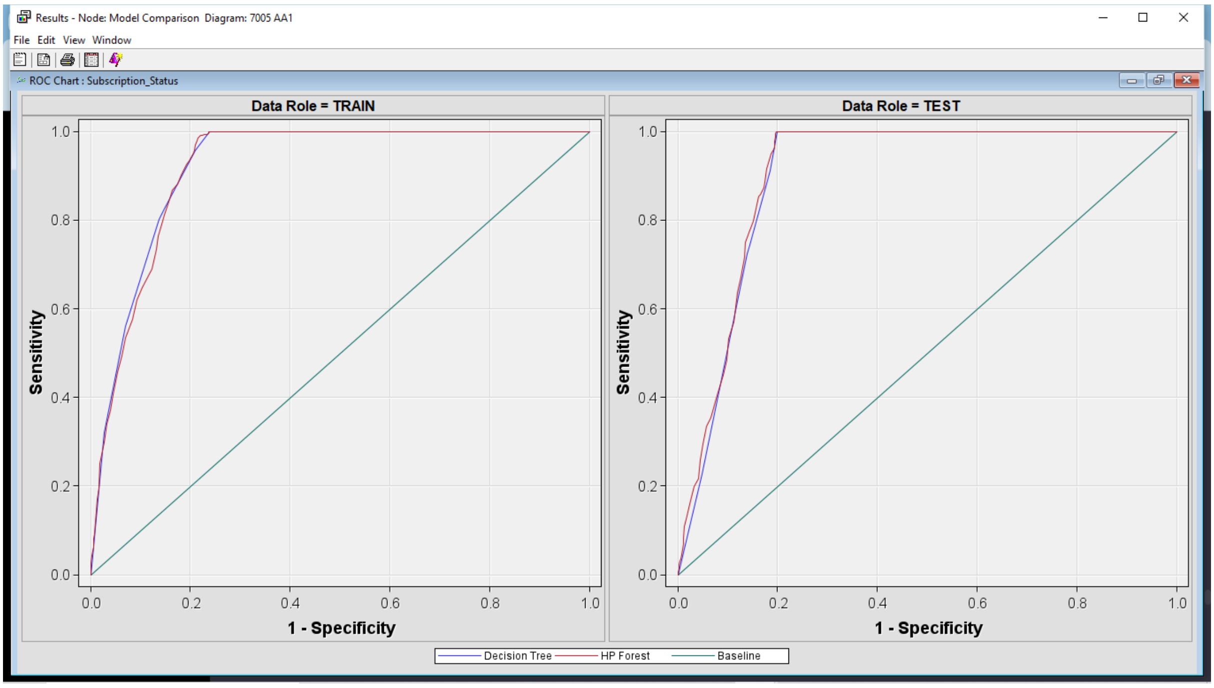
Variable Importance

Variable Name	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction	OOB: Gini Reduction	OOB: Margin Reduction	Label
Promo_Co...	55	0.086732	0.173463	0.08737	0.17493	
Discount_A...	43	0.065630	0.131259	0.06554	0.13128	
Gender	25	0.016143	0.032285	0.01598	0.03146	
Age	9	0.000341	0.000682	-0.00057	-0.00024	
Location	9	0.000871	0.001743	-0.00085	0.00060	
IMP_Revie...	7	0.000113	0.000226	-0.00009	0.00003	Imputed Re...
Previous_P...	7	0.000129	0.000258	-0.00023	-0.00010	
Shipping_T...	2	0.000082	0.000164	-0.00008	-0.00001	
Category	1	0.000030	0.000060	-0.00007	-0.00002	
Item_Purch...	1	0.000085	0.000171	-0.00012	-0.00006	
Purchase_...	1	0.000022	0.000044	-0.00004	-0.00002	
Size	1	0.000035	0.000071	-0.00005	-0.00002	
Color	0	0.000000	0.000000	0.00000	0.00000	
Frequency_...	0	0.000000	0.000000	0.00000	0.00000	
Payment_M...	0	0.000000	0.000000	0.00000	0.00000	
Season	0	0.000000	0.000000	0.00000	0.00000	

From the variable importance results, promo code used is the most important variable with splitting rules of 55 and train Gini reduction of 0.0867 and OOB Gini reduction of 0.0873. A higher Gini reduction indicates a more significant improvement in purity or homogeneity of classes after the split.

Comparison Model

Here is the performance comparison for both model: -



In ROC graph show both have the diagonal line (from (0,0) to (1,1)) represents the performance of a random classifier that makes predictions without any discrimination between classes. Points above the diagonal line indicate better-than-random performance.

From the table below, both prediction models outperform in their performance metrics, however random forest model show perfect scores in recall metrics. This indicates here that random forest model outperforms decision tree model but with small significant difference only.

it Statistics
odel Selection based on Train: Misclassification Rate (_MISC_)

elected Model	Model Node	Model Description	Train: Misclassification Rate	Train: Average Squared Error	Train: Roc Index	Train: Gini Coefficient
Y	Tree	Decision Tree	0.15440	0.09623	0.922	0.844
	HPDMForest	HP Forest	0.17447	0.10543	0.919	0.838

Results - Node: Model Comparison Diagram: 7005 AA1

File Edit View Window

Fit Statistics

Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error	Train: Sum of Squared Errors	Train: Frequency of Classified Cases	Train: Misclassification Rate	Train: Number of Wrong Classifications	Test: Average Squared Error	Test: Divisor for ASE	Test: Maximum Absolute Error	Test: Sum of Frequencies	Test: Root Average Squared Error	Test: Sum of Squared Errors	Test: Frequency of Classified Cases	Test: Misclassification Rate	Test: Number of Wrong Classifications
0.096232	3886	0.811594	1943	0.310213	373.9569	.	0.1544	.	0.104675	3890	0.811594	1945	0.323535	407.1852	.	0.17635	
0.105435	3886	0.604997	1943	0.324707	409.7185	1943	0.174472	339	0.096779	3890	0.611067	1945	0.311093	376.4709	1945	0.146015	

	Decision Tree	Random Forest
Accuracy	0.845	0.825
Precision	0.681	0.607
Recall	0.802	1.000
F1-score	0.736	0.755
Misclassification rate in Train model	0.15	0.17
Average Square Error in Train Model	0.10	0.11
Misclassification rate in Test model	0.18	0.15
Average Square Error in Test Model	0.10	0.10