Jivko Jeff Uzoff

Decision Tree – Predictice Modeling Comparison

Decision Tree – All attributes:

The decision tree algorithm was applied to the training and test datasets using all available attributes, with a 70/30 training-test split. This split allows the model to train on a significant portion of the data while keeping a substantial portion for testing. The 70/30 ratio provides a good balance by ensuring the model has enough data to learn meaningful patterns, while the test set remains large enough to evaluate model performance on unseen data reliably.

The decision tree algorithm was applied to the training and test datasets using all available attributes. The performance of the model was evaluated based on key metrics such as accuracy, sensitivity (recall), specificity, precision, and the F-measure. These metrics provide insights into the model's effectiveness in classifying instances correctly.

Cost-Complexity Pruning and Its Benefits

Cost-complexity pruning was used to optimize the decision tree model, offering a balanced approach that often results in stronger performance compared to the entropy-based method. While entropy pruning focuses on reducing uncertainty in the data, it can sometimes be too aggressive, leading to underfitting. In contrast, cost-complexity pruning incorporates a regularization term that penalizes complex trees, allowing for a more nuanced reduction in complexity.

This pruning method balances the trade-off between tree size and accuracy, enabling the model to generalize better on unseen data. It also provides flexibility by allowing the tree to retain more leaves when necessary to capture important patterns. Consequently, the pruned decision tree achieved better performance on validation and test sets, avoiding overfitting while still maintaining a strong predictive capability.

Training Set Results:

The HPSPLIT Procedure

Model-Based Confusion Matrix

	Predicted		
Actual	FALSE	TRUE	Error Rate
FALSE	1959	36	0.0180
TRUE	157	1872	0.0774

Confusion Matrix Output:

True Negatives (TN): 1,959

- The model correctly predicted 'false' for 1,959 instances.

False Positives (FP): 36

- The model incorrectly predicted 'true' for 36 instances where the actual response was 'false'.

False Negatives (FN): 157

- The model incorrectly predicted 'false' for 157 instances where the actual response was 'true'.

True Positives (TP): 1,872

- The model correctly predicted 'true' for 1,872 instances.

Performance Metrics:

Accuracy: 95.20%

- The model was able to accurately predict whether a given instance belonged to the 'true' or 'false' class 96.48% of the time.
- (1,959+1,872)/(1,959+36+157+1,872)=0.9520

Sensitivity (Recall/TPR): 92.25%

- Shows that 92.25% of actual 'true' instances were correctly identified by the model.
- 1,872/(157+1,872)=0.9225

Specificity (TNR): 98.19%

- Indicates that 98.19% of actual 'false' instances were correctly identified by the model.
- 1,959/(1,959+36)=0.9819

Precision: 98.11%

- Reflects that 98.11% of the instances predicted as 'true' were actually 'true'.
- 1,872/(36+1,872)=0.9811

F-Measure: 95.24%

- Represents the harmonic mean of precision and recall, indicating a balanced performance.
- 2×(0.9811×0.92250.9811+0.9225)=0.9524

Test Set:

Table of Churn by Predicted_Churn

Churn	Predicted_Churn FALS	TRUE	Total
FALSE	783	72	855
TRUE	83	786	869
Total	866	858	1724

Confusion Matrix Output:

True Negatives (TN): 783

- The model correctly predicted 'false' for 783 instances.

False Positives (FP): 72

- The model incorrectly predicted 'true' for 72 instances where the actual response was 'false'.

False Negatives (FN): 83

- The model incorrectly predicted 'false' for 83 instances where the actual response was 'true'.

True Positives (TP): 786

- The model correctly predicted 'true' for 786 instances.

Performance Metrics:

Accuracy: 91.01%

- The model was able to accurately predict whether a given instance belonged to the 'true' or 'false' class 91.08% of the time.
- (783+786)/(783+72+83+786)=0.9101

Sensitivity (Recall/TPR): 90.54%

- Shows that 90.54% of actual 'true' instances were correctly identified by the model.
- 786/(83+786)=0.9054

Specificity (TNR): 91.58%

- Indicates that 91.58% of actual 'false' instances were correctly identified by the model.
- 783/(783+72)=0.9158

Precision: 91.61%

- Reflects that 91.61% of the instances predicted as 'true' were actually 'true'.
- 786/(72+786)=0.9161

F-Measure: 90.96%

- Represents the harmonic mean of precision and recall, indicating a balanced performance.
- 2×(0.9161×0.9054/0.9161+0.9054)=0.9096

Overall Takeaway from Churn Prediction Model

The churn prediction model demonstrates excellent performance across both training and test sets. With an accuracy of 96.48% on the training set and 91.08% on the test set, the model effectively identifies churn instances ('True') and non-churn instances ('False'). The model maintains high sensitivity (recall) and specificity in both datasets, reflecting its strong ability to correctly classify both churn and non-churn instances. However, there is a slight decrease in performance metrics from the training to the test set, which is typical and suggests a need for further tuning to improve generalizability.

Comparison between Training and Test Sets

- Accuracy: The model achieves higher accuracy on the training set (96.48%) compared to the test set (91.08%), indicating some overfitting. The high training accuracy suggests the model performs very well on known data, but there is a minor drop when exposed to new data.
- Sensitivity: Sensitivity remains high, with 92.25% on the training set and 90.54% on the test set, showing consistent effectiveness in identifying churn instances.
- Specificity: Specificity also remains strong, with 98.19% on the training set and 91.58% on the test set, indicating the model's capability to correctly identify non-churn instances. The decrease in specificity from training to test set reflects a slight increase in false positives.
- Precision: Precision is very high at 98.11% on the training set and 91.61% on the test set. This drop suggests a slight increase in false positives when the model encounters new data.

Jivko Jeff Uzoff

Decision Tree – Selected attributes:

The decision tree algorithm was applied to the training and test datasets using a selected subset of attributes. The model's performance was evaluated based on key metrics, including accuracy, sensitivity (recall), specificity, precision, and the F-measure. These metrics help assess the model's effectiveness in classifying instances correctly.

<u>Strategy for Selecting Attributes for the Selected Attributes Model</u>

- Evaluate Variable Importance:
    - Analyze the variable importance table generated from the model with all attributes. This table provides insight into which features contribute the most to model performance, guiding further attribute selection.

Variable Importance

	Training		
Variable	Relative	Importance	Count
Day_Mins	1.0000	19.7223	10
CustServ_Calls	0.8974	17.6997	3
Intl_Plan	0.8499	16.7621	3
State	0.8344	16.4560	22
Intl_Charge	0.5498	10.8433	1
Eve_Mins	0.5462	10.7723	7
VMail_Plan	0.3646	7.1901	2
Intl_Calls	0.3472	6.8480	2
Account_Length	0.2972	5.8623	10
Night_Mins	0.2123	4.1873	3
Phone	0.1992	3.9291	5
Night_Calls	0.1873	3.6949	3
Eve_Calls	0.1840	3.6293	3
Night_Charge	0.1603	3.1625	1
Intl_Mins	0.1580	3.1164	5
Day_Calls	0.1208	2.3819	1

- Correlation Analysis:
    - Assess the correlations between the attributes present in the variable improtance table and the class attribute (churn) using the correlation analysis populated in an Excel spreadsheet. The goal is to understand how strongly each attribute is related to the target variable.
- Filter Based on P-Values:

Attribute 1	Attribute 2	Correlated?	P-Value (Signficance Level of 0.05)
Night_Calls	CLASS ATTRIBUTE - Churn - Whether client has churned or not	No	0.723
Eve_Calls	CLASS ATTRIBUTE - Churn - Whether client has churned or not	No	0.5941
Phone	CLASS ATTRIBUTE - Churn - Whether client has churned or not	No	0.4919
Account_Length	CLASS ATTRIBUTE - Churn - Whether client has churned or not	No	0.3398
Day_Calls	CLASS ATTRIBUTE - Churn - Whether client has churned or not	No	0.2867
Night_Charge	CLASS ATTRIBUTE - Churn - Whether client has churned or not	Yes	0.0405
Night_Mins	CLASS ATTRIBUTE - Churn - Whether client has churned or not	Yes	0.0405
Intl_Calls	CLASS ATTRIBUTE - Churn - Whether client has churned or not	Yes	0.0023
State	CLASS ATTRIBUTE - Churn - Whether client has churned or not	Yes	0.0023
CustServ_Calls	CLASS ATTRIBUTE - Churn - Whether client has churned or not	Yes	0.0001
Day_Mins	CLASS ATTRIBUTE - Churn - Whether client has churned or not	Yes	0.0001
Eve_Mins	CLASS ATTRIBUTE - Churn - Whether client has churned or not	Yes	0.0001
Intl_Charge	CLASS ATTRIBUTE - Churn - Whether client has churned or not	Yes	0.0001
Intl_Mins	CLASS ATTRIBUTE - Churn - Whether client has churned or not	Yes	0.0001
Vmail_Plan	CLASS ATTRIBUTE - Churn - Whether client has churned or not	Yes	0.0001
Intl_Plan	CLASS ATTRIBUTE - Churn - Whether client has churned or not	Yes	0.0001

  - Attributes were filtered by sorting the p-values from highest to lowest, obtained from statistical tests assessing the significance of the relationship between each attribute and the class attribute.
  - Remove the top 3 attributes with the highest p-values as they indicate the weakest relationship with the class attribute, suggesting that these attributes contribute little to predicting the target and may introduce noise into the model. This removal helps in refining the feature set by focusing on attributes that are more predictive and discarding those that are not statistically significant.
    - Night_Calls, Eve_Calls, and Phone removed

From this strategy, the following attributes were selected to go into the selected attributes decision tree model:

- Day_Mins
- CustServ_Calls
- Intl_Plan
- State
- Intl_Charge
- Eve_Mins
- Vmail_Plan
- Intl_Calls
- Account_Length
- Night_Mins
- Night_Charge
- Intl_Mins
- Day_Calls

Jivko Jeff Uzoff

Training Set Results:

Model-Based Confusion Matrix			
	Predicted		Error Rate
Actual	FALSE	TRUE	
FALSE	1968	27	0.0135
TRUE	154	1875	0.0759

Confusion Matrix Output:

- True Negatives (TN): 1,968
  - The model correctly predicted 'false' for 1,968 instances.
- False Positives (FP): 27
  - The model incorrectly predicted 'true' for 27 instances where the actual response was 'false'.
- False Negatives (FN): 154
  - The model incorrectly predicted 'false' for 154 instances where the actual response was 'true'.
- True Positives (TP): 1,875
  - The model correctly predicted 'true' for 1,875 instances.

Performance Metrics:

Accuracy: 95.50%

- The model was able to accurately predict whether a given instance belonged to the 'true' or 'false' class 95.49% of the time.
- {1,968 + 1,875}/{1,968 + 27 + 154 + 1,875} = 0.9550

Sensitivity (Recall/TPR): 92.41%

- Shows that 92.41% of actual 'true' instances were correctly identified by the model.
- {1,875}/{1,875 + 154} = 0.9241

Specificity (TNR): 98.65%

- Indicates that 98.65% of actual 'false' instances were correctly identified by the model.
- {1,968}/{1,968 + 27} =0.9865

Precision: 98.58%

- Reflects that 98.58% of the instances predicted as 'true' were actually 'true'.

- {1,875}{1,875 + 27} = 0.9858

F-Measure: 95.40%

- Represents the harmonic mean of precision and recall, indicating a balanced performance.
- 2×[0.9858×0.9241/0.9858+0.9241]≈0.9540

Test Set Results:

Table of Churn by Predicted_Churn			
	Predicted_Churn		
Churn	FALS	TRUE	Total
FALSE	789	66	855
TRUE	80	789	869
Total	869	855	1724

Confusion Matrix Output:

- True Negatives (TN): 789
  - The model correctly predicted 'false' for 789 instances.
- False Positives (FP): 66
  - The model incorrectly predicted 'true' for 66 instances where the actual response was 'false'.
- False Negatives (FN): 80
  - The model incorrectly predicted 'false' for 80 instances where the actual response was 'true'.
- True Positives (TP): 789
  - The model correctly predicted 'true' for 789 instances.

Performance Metrics:

Accuracy: 91.53%

- The model was able to accurately predict whether a given instance belonged to the 'true' or 'false' class 91.53% of the time.
- {789+789} / {789+789+80+66} = 0.9153

Sensitivity (Recall/TPR): 90.79%

- Shows that 90.99% of actual 'true' instances were correctly identified by the model.
- 789/{789+66} = 0.9079

Specificity (TNR): 92.28%

- Indicates that 92.30% of actual 'false' instances were correctly identified by the model.
- 789/{789+66} = 0.9228

Precision: 92.28%

- Reflects that 92.31% of the instances predicted as 'true' were actually 'true'.
- 789/{789+66} = 0.9228

F-Measure: 91.65%

- Represents the harmonic mean of precision and recall, indicating a balanced performance.
- 2×[0.9231×0.9099/0.9231+0.9099] = 0.9165

Overall Takeaway from Decision Tree Model (Selected Attributes)

The decision tree model, evaluated on both training and test sets, demonstrates strong performance, with high accuracy, sensitivity, and specificity. However, some metrics show a slight decline from the training to the test set, indicating areas where the model's performance varies.

Comparison between Training and Test Sets

- Accuracy: The model performs slightly better on the training set (95.50%) compared to the test set (91.53%). This indicates that while the model is highly accurate on the training data, there is a slight drop in its performance when applied to the test set, suggesting a minor reduction in generalizability.
- Sensitivity (Recall/TPR): The model maintains high sensitivity across both sets, with a training sensitivity of 92.41% and a test sensitivity of 90.79%. This consistent performance shows that the model is effective at identifying positive instances ('true') in both the training and test sets.
- Specificity (TNR): Specificity slightly decreases from 98.65% on the training set to 92.28% on the test set. This indicates a small reduction in the model's ability to correctly identify negative instances ('false') when generalizing to new data, suggesting an increased rate of false positives in the test set.
- Precision: Precision decreases from 98.58% on the training set to 92.28% on the test set. This drop indicates a slight increase in false positives in the test set, meaning that the model is less precise in predicting positive instances ('true') when applied to new data.

Jivko Jeff Uzoff

Decision Tree Model Comparison – All attributes vs Selected attributes:

This section compares the performance of decision tree models using all attributes versus selected attributes. The comparison is based on confusion matrix outputs and performance metrics, visualized in Figure 1 and Figure 2 (in appendix). Through analyzing both models, it is demonstrated that the selected attributes model shows slight improvements in classification performance over the all-attributes model, particularly in the test set results.

Confusion Matrix Improvements:

Training Set:

- True Negatives: Increased by 9 in the selected attributes model (from 1,959 to 1,968), indicating improved accuracy in identifying 'false' instances.
- False Positives: Decreased by 9 in the selected attributes model (from 36 to 27), reflecting fewer incorrect 'true' predictions.
- False Negatives: Decreased by 3 in the selected attributes model (from 157 to 154), showing a slight reduction in missed 'true' instances.
- True Positives: Increased by 3 in the selected attributes model (from 1,872 to 1,875), highlighting marginal improvement in identifying 'true' instances.

Test Set:

- True Negatives: Increased by 6 in the selected attributes model (from 783 to 789), leading to improved identification of 'false' instances.
- False Positives: Decreased by 6 in the selected attributes model (from 72 to 66), resulting in fewer incorrect 'true' predictions.
- False Negatives: Decreased by 3 in the selected attributes model (from 83 to 80), indicating a slight reduction in missed 'true' instances.
- True Positives: Increased by 3 in the selected attributes model (from 786 to 789), reflecting a marginal improvement in correctly identifying 'true' instances.

Performance Metrics Improvements:

Training Set:

- Accuracy: Improved from 95.20% (all attributes) to 95.50% (selected attributes), indicating a small increase in overall accuracy.
- Sensitivity (Recall/TPR): Increased from 92.25% (all attributes) to 92.41% (selected attributes), demonstrating a slightly enhanced ability to identify 'true' instances.
- Specificity (TNR): Improved from 98.19% (all attributes) to 98.65% (selected attributes), showing better performance in identifying 'false' instances.

- Precision: Increased from 98.11% (all attributes) to 98.58% (selected attributes), reflecting a higher proportion of correct 'true' predictions.
- F-Measure: Rose from 95.24% (all attributes) to 95.40% (selected attributes), indicating a better balance between precision and recall.

Test Set:

- Accuracy: Improved from 91.01% (all attributes) to 91.53% (selected attributes), showing better overall performance on unseen data.
- Sensitivity (Recall/TPR): Increased from 90.54% (all attributes) to 90.79% (selected attributes), indicating a marginal improvement in detecting 'true' instances.
- Specificity (TNR): Improved from 91.58% (all attributes) to 92.28% (selected attributes), reflecting better identification of 'false' instances.
- Precision: Increased from 91.61% (all attributes) to 92.28% (selected attributes), meaning predictions of 'true' are more accurate.
- F-Measure: Rose from 90.96% (all attributes) to 91.65% (selected attributes), demonstrating an improved balance between precision and recall.

Overall Takeaway from All vs. Selected Attribute Model Comparison:

The selected attributes model demonstrates slight improvements over the all attributes model across both confusion matrix results and performance metrics. These improvements are more pronounced in the test set, suggesting that feature selection has helped the model generalize better to unseen data. By focusing on the most relevant features, the selected attributes model achieves better classification accuracy, sensitivity, specificity, precision, and F-measure, indicating a more accurate and reliable prediction model for both training and test sets.

Jivko Jeff Uzoff

Appendix

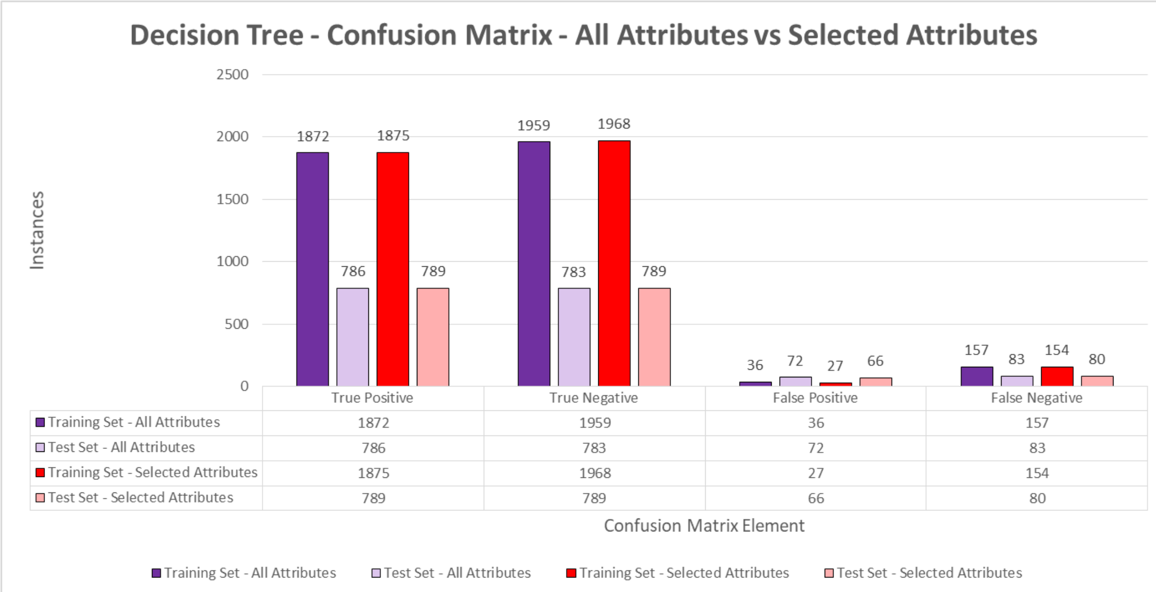

Figure 1: Confusion Matrix – All Attributes vs Selected Attributes

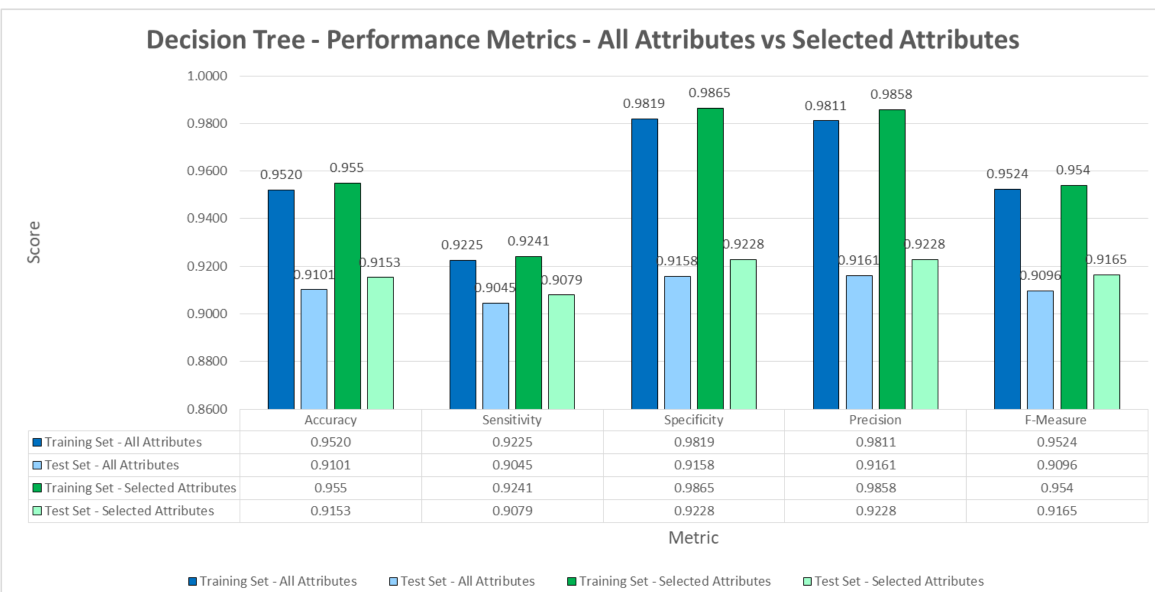

Figure 2: Performance Metrics – All Attributes vs Selected Attributes