Dikson Rajbanshi

Student ID: 77202796

Assessment 1

INTELLIGENT SYSTEMS

# Pre-processing Techniques Applied

### Filter Examples

Filter out rows with missing data.

### Remove Duplicates

Removes duplicate data from the dataset

### Select Attribute

Selects only the required attributes.

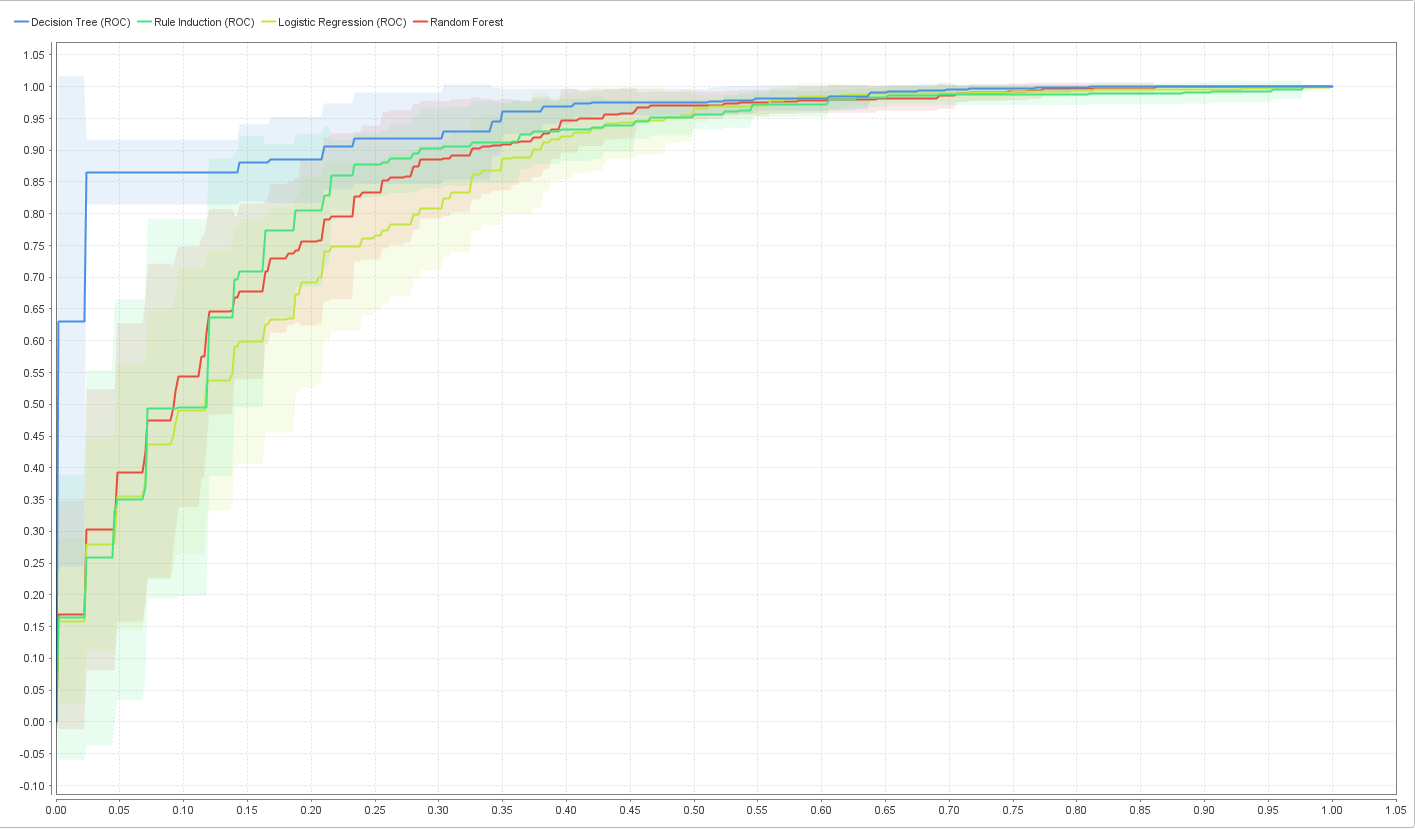
### Normalize

Normalizes all the required attributes to a range of 0-1

### Set Role

Sets the role of the output attribute as the label

# ROC Curve



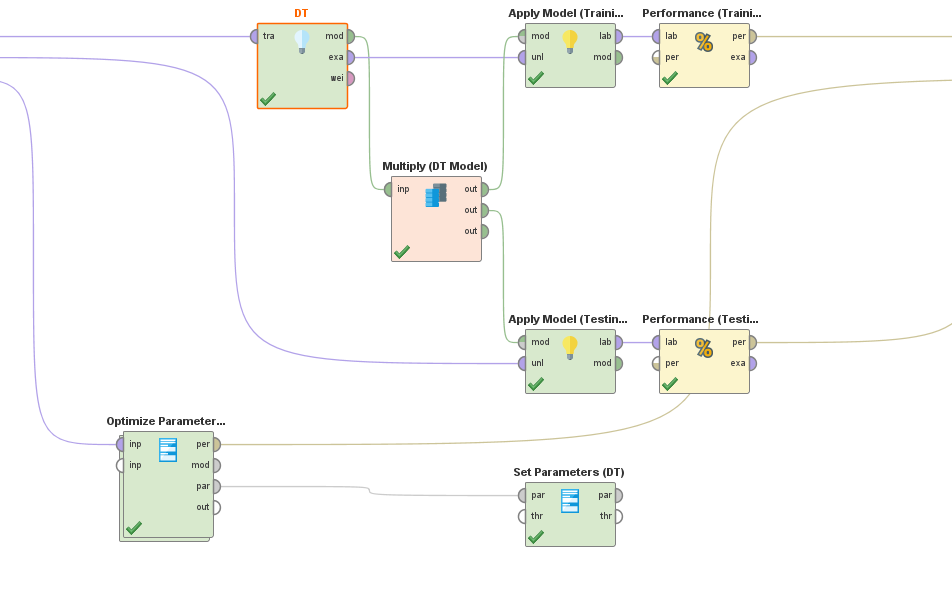
*(fig 1: ROC curve)*

# Technique 1: Decision Tree

### Motivation

* Dataset has binomial output/ class i.e., discrete data. So, classification algorithms are best for the dataset and decision tree is a classification technique.
* Best Area under the ROC Curve (fig 1)

### Snapshot

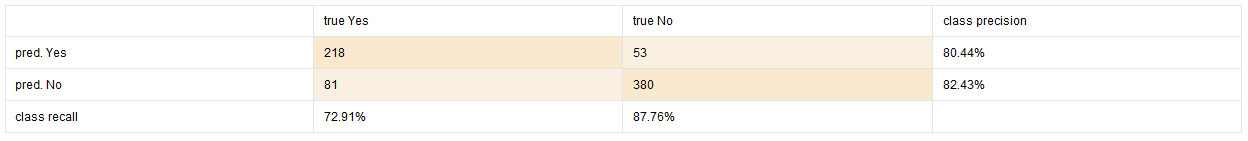


### Parameter Settings

* Criterion: gini\_index, favors larger partition and binary splits
* Maximal depth: 4, showed the best performance
* Apply pruning: false
* Apply prepruning: false

|  |  |
| --- | --- |
| accuracy | 81.69% |
| precision | 82.43% (positive class: No) |
| recall | 87.76% (positive class: No) |
| f-measure | 85.01% (positive class: No) |

###### Confusion Matrix:

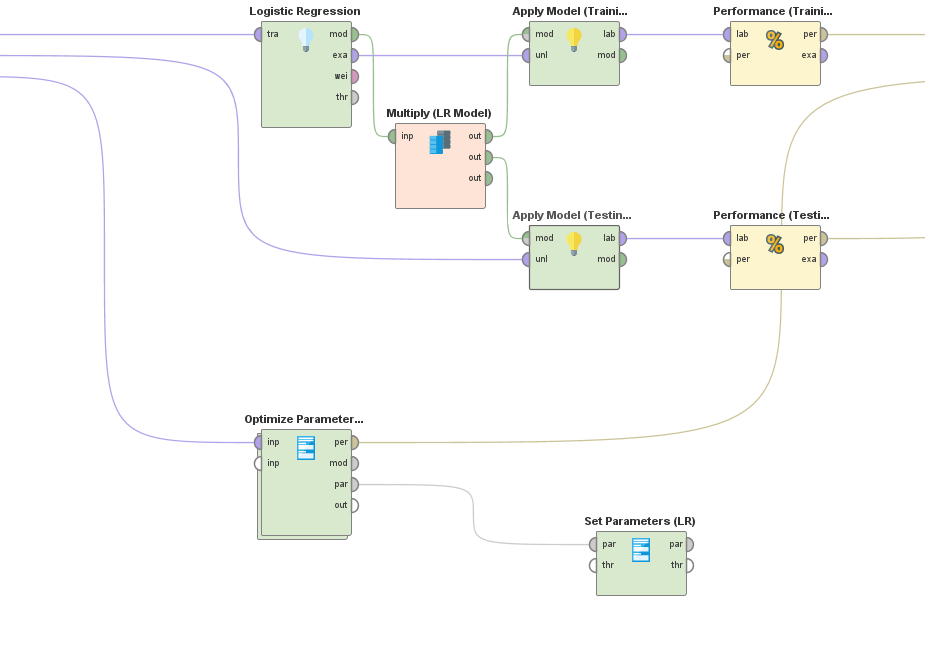


# Technique 2: Logistic Regression

### Motivation

* Logistic Regression is a classification technique and RiskyJournyCO dataset has binomial output.
* Dataset has low arguments/ parameters count

### Snapshot

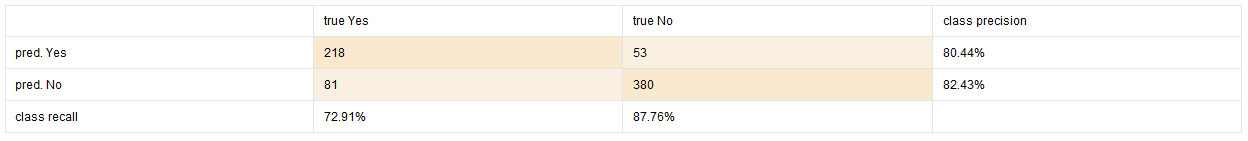


### Parameter Settings

* Solver: auto (default)
* Reproducible: false (default)
* Use regularization: false (default), no overfitting
* Standardize: true (default)
* Non-negative coefficients: false (default)
* Add intercept: true (default)
* Compute p-values: true (default)
* Remove collinear column: true (default)

|  |  |
| --- | --- |
| accuracy | 78.01% |
| precision | 80.09% (positive class: No) |
| recall | 83.60% (positive class: No) |
| f-measure | 81.81% (positive class: No) |

###### Confusion Matrix:



# Comparison of Testing Performance

### Testing Performance Report

### Cross-validation Performance Report

# Final Recommendation of Best Modal

Decision Tree shows the best performance (accuracy, precision, recall and f-measure).

Being a simple technique, Decision Tree has lower performance requirement from the machine than other models. This lowers the cost and time for training the model and using the model for actual labelling task.

Also, Decision Tree provides inner working of how the model reached its decision for the output.

Hence, I would recommend Decision Tree for RiskyJournyCO as it provides the best performance along with lower cost.