ASSESSMENT 1

INTELLIGENT SYSTEMS

Dikson Rajbanshi
Student ID: 77202796

Pre-processing Techniques Applied

1. ### Filter Examples
   Filter out rows with missing data.

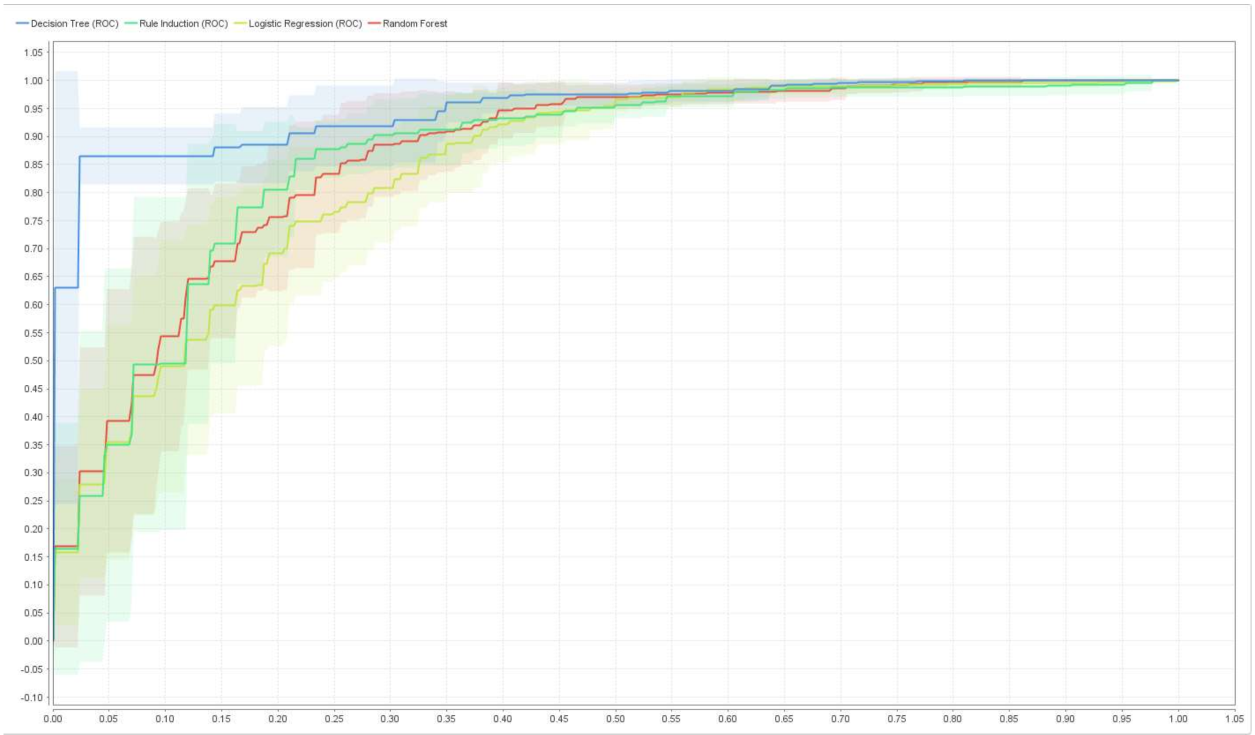
2. ### Remove Duplicates
   Removes duplicate data from the dataset

3. ### Select Attribute
   Selects only the required attributes.

4. ### Normalize
   Normalizes all the required attributes to a range of 0-1

5. ### Set Role
   Sets the role of the output attribute as the label

ROC Curve



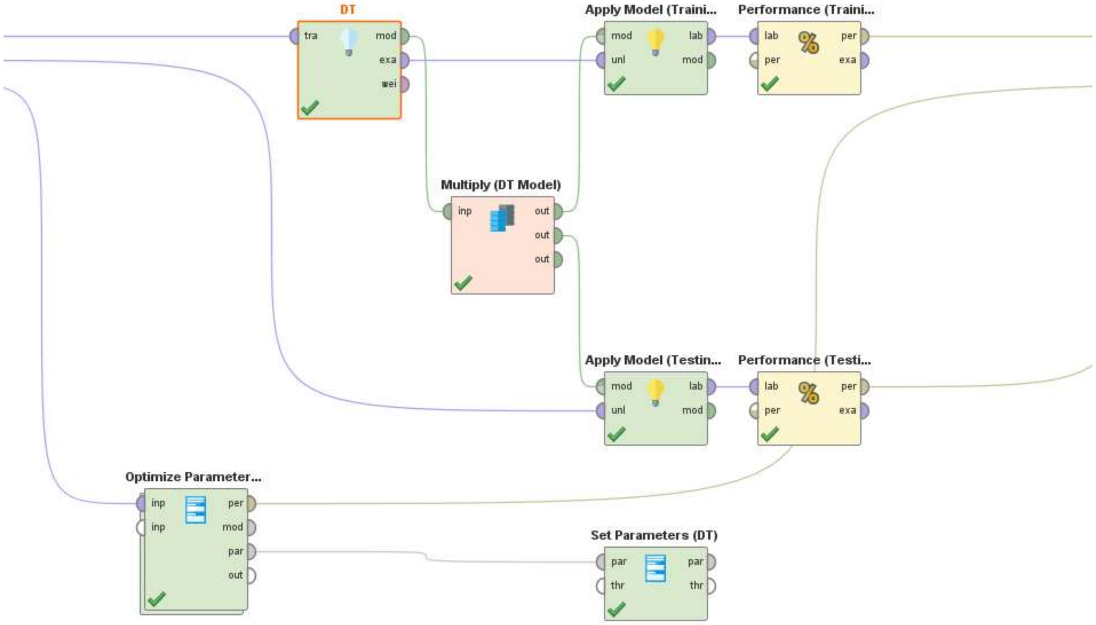
(fig 1: ROC curve)

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

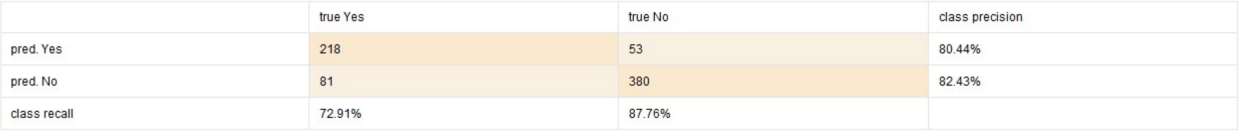


Parameter Settings

- 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:



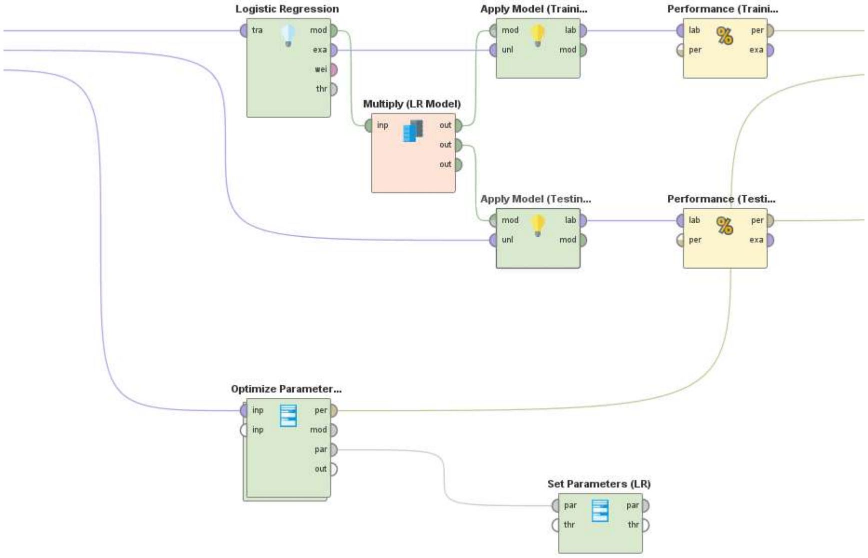
	true Yes	true No	class precision
pred. Yes	218	53	80.44%
pred. No	81	380	82.43%
class recall	72.91%	87.76%	

Technique 2: Logistic Regression

Motivation

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

Snapshot

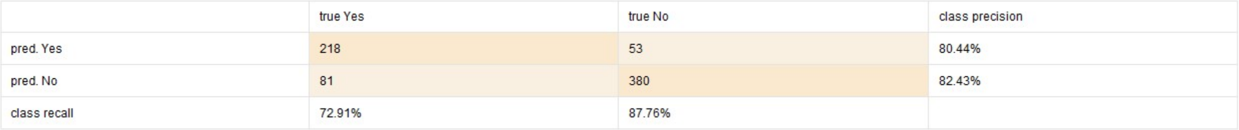


Parameter Settings

- 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)

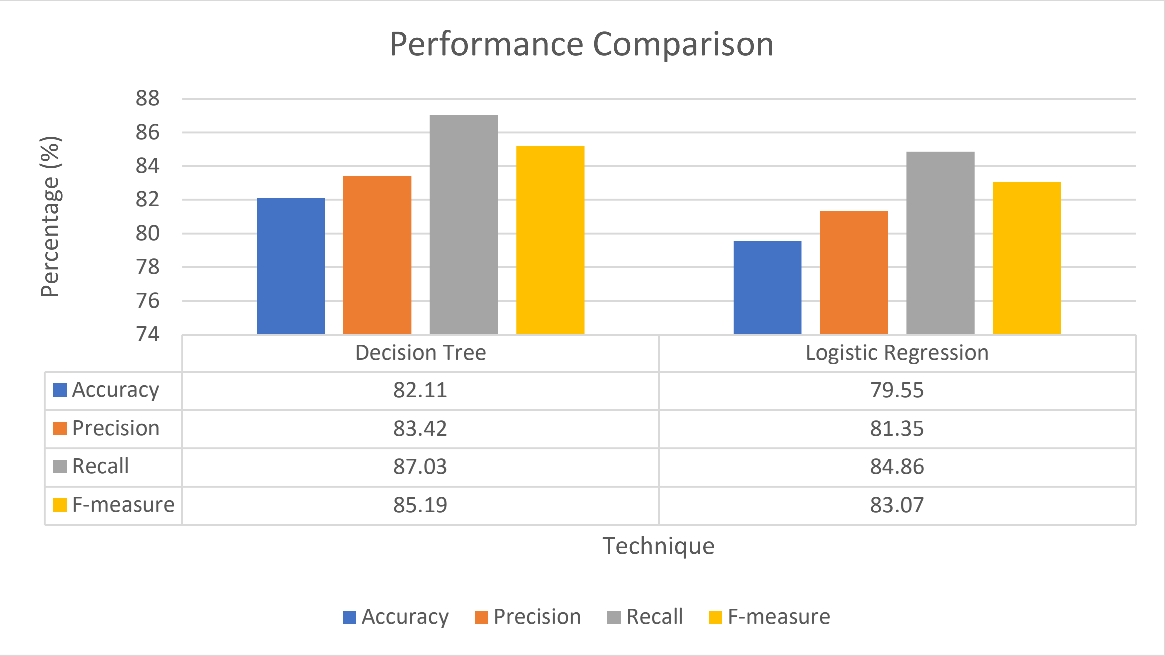
Confusion Matrix:



	true Yes	true No	class precision
pred. Yes	218	53	80.44%
pred. No	81	380	82.43%
class recall	72.91%	87.76%	

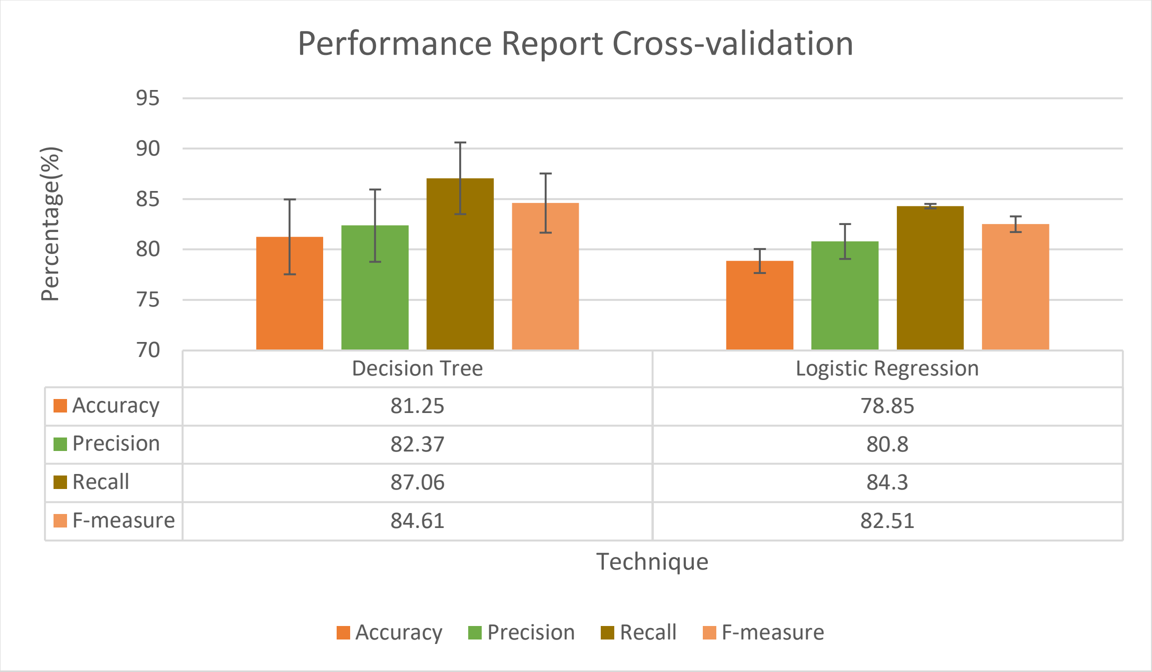
Comparison of Testing Performance

Testing Performance Report



	Decision Tree	Logistic Regression
🟦 Accuracy	82.11	79.55
🟧 Precision	83.42	81.35
⬜ Recall	87.03	84.86
🟨 F-measure	85.19	83.07

Cross-validation Performance Report



	Decision Tree	Logistic Regression
🟧 Accuracy	81.25	78.85
🟩 Precision	82.37	80.8
🟫 Recall	87.06	84.3
🟧 F-measure	84.61	82.51

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.