

Lab Assignment 2

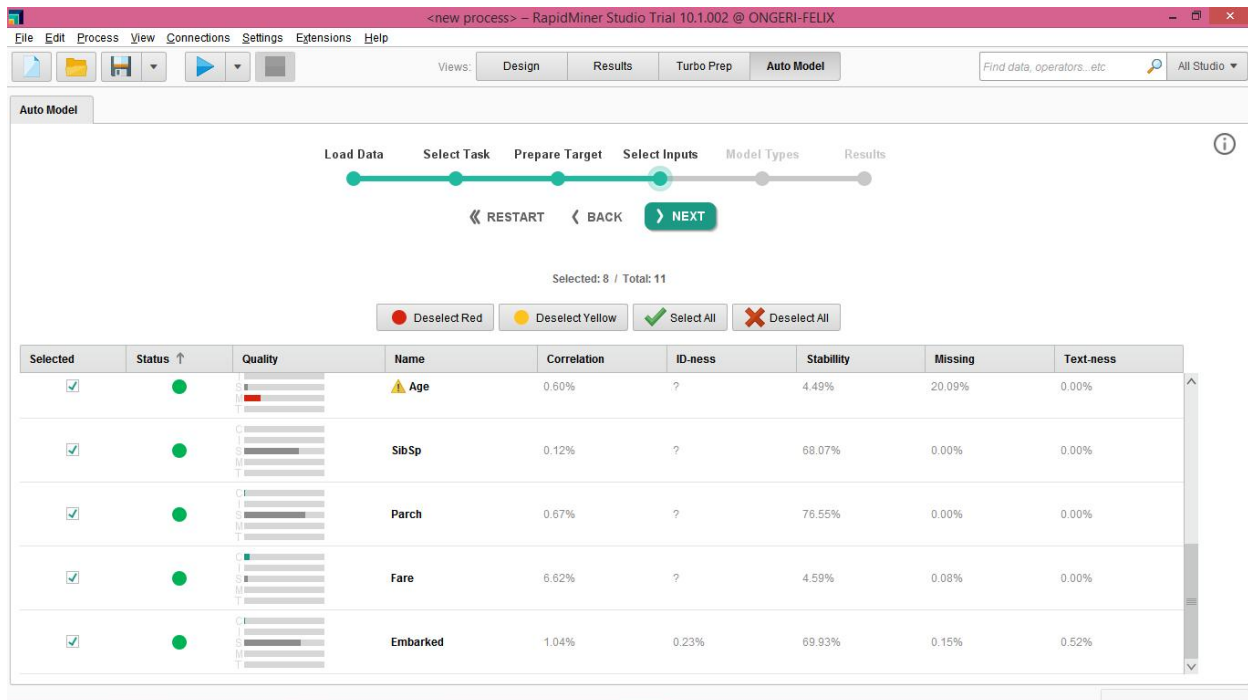
Were Vincent Ouma

BUSINESS ANALYTICS: DATA, MODELS AND DECISIONS

Homework 10






Question I

The variables included in the model as predictors were Passenger ID, sex, Pclas, Fare, Embarked, Parch, Age and SibSP. These variables were flagged with either Orange or Green colors and also marked as shown below.



Selected: 8 / Total: 11

☐ Deselect Red
 ☐ Deselect Yellow
 ☒ Select All
 ☒ Deselect All

Selected	Status	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
<input checked="" type="checkbox"/>	●		Age	0.60%	?	4.49%	20.09%	0.00%
<input checked="" type="checkbox"/>	●		SibSp	0.12%	?	68.07%	0.00%	0.00%
<input checked="" type="checkbox"/>	●		Parch	0.67%	?	76.55%	0.00%	0.00%
<input checked="" type="checkbox"/>	●		Fare	6.62%	?	4.59%	0.08%	0.00%
<input checked="" type="checkbox"/>	●		Embarked	1.04%	0.23%	69.93%	0.15%	0.52%

<new process> – RapidMiner Studio Trial 10.1.002 @ ONGERI-FELIX

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model

Find data, operators, etc. All Studio

Auto Model

Load Data Select Task Prepare Target Select Inputs Model Types Results

RESTART BACK NEXT

Selected: 8 / Total: 11

Deselect Red Deselect Yellow Select All Deselect All

Selected	Status ↑	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
<input checked="" type="checkbox"/>	●		PassengerId	0.00%	?	0.08%	0.00%	0.00%
<input checked="" type="checkbox"/>	●		Pclass	11.46%	?	54.16%	0.00%	0.00%
<input checked="" type="checkbox"/>	●		Sex	29.52%	0.15%	64.40%	0.00%	2.15%
<input checked="" type="checkbox"/>	●		Age	0.60%	?	4.49%	20.09%	0.00%

Question II

Random forest was found to be the best algorithm for the prediction. This is because it had the highest AUC of 0.903, this led to producing the best outcome.

<new process> – RapidMiner Studio Trial 10.1.002 @ ONGERI-FELIX

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model

Find data, operators, etc. All Studio

Auto Model

Load Data Select Task Prepare Target Select Inputs Model Types Results

RESTART BACK OPEN PROCESS EXPORT

Results

Comparison

Overview

ROC Comparison

Naive Bayes

Model

Weights

Simulator

Performance

Lift Chart

Predictions

Production Model

Generalized Linear Model

SAVE RESULTS

Overview

Number of Models: 205

AUC

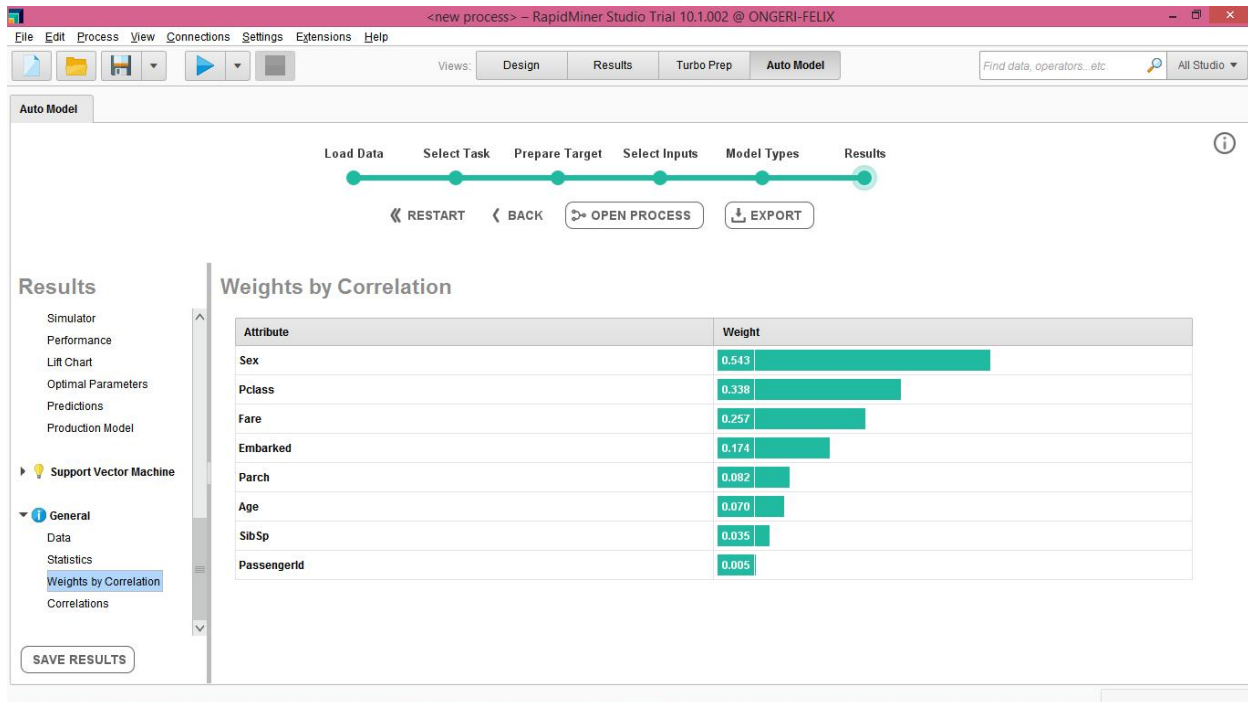
Runtimes (ms)

Model	AUC	Standard Deviation	Gains	Total Time	Trainin
Deep Learning	0.882	± 0.039	86	15 s	752 ms
Decision Tree	0.828	± 0.028	100	10 s	28 ms
Random Forest	0.903	± 0.034	100	2 min 35 s	121 ms
Gradient Boosted Trees	0.807	± 0.03	100	38 s	705 ms

Random Forest Best Performance

Question III

Sex was the most important predictor in the model. This can be determined by the weights in the correlations as shown below.



Question IV

The file for the Random forest algorithm is exported as RandomForestPredictions.xls