

**ISM 6136 – Datamining/Predictive Analytics**

**Class Assignment 6**

**5 points**

**Jacob Perrone**

**TASK: Performing predictive analytics using Logistic Regression in RapidMiner**

**Perform the following data mining steps using Logistic Algorithm using RapidMiner and predict based on the past data which of the new readers in the Charles Book club will buy the Florence book ‘Yes – 1 ‘ or ‘No – 0’ .**

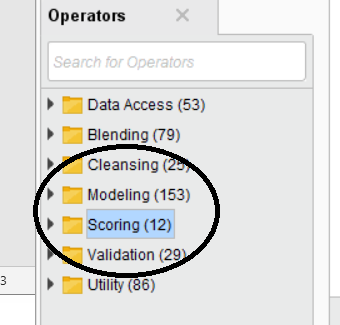
Follow the datamining steps below:

1. Understand the problem and purpose of data mining task
2. Import the dataset into RapidMiner
3. Explore, clean and preprocess data
4. Cleanup or do not select any column that is not a predictor
   1. Check ‘Replace errors by missing values’
   2. Check Statistics and look for any missing values (if yes then you will have to add a replace with the ‘Replace Missing Values’ operator

Table

Description automatically generated

1. Design your process using appropriate operators. Provide screen shot of the overall design.



1. **Build three models** – save each of the models into a separate ‘process’

Provide screen shots of each of the model settings– to show me difference in each model – You can even present this in Table form.

Model 1: model 1 used 10 fold cross validation and default selections for Logistic Regression options. The confusion matrix, accuracy, and AUC and ROC curve are shown below. Accuracy was 89.30% while AUC was rather on the low side with a value of 0.781.

A picture containing graphical user interface

Description automatically generated

A picture containing diagram

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Chart

Description automatically generated

Model 2: model 2 also used 10 fold cross validation and the same default selection options for Logistic Regression except I changed the number of iterations to 20. This model had the same performance as model 1 in terms of accuracy, and AUC score.

A picture containing diagram

Description automatically generated

Diagram

Description automatically generated with medium confidence

Graphical user interface, application, Word

Description automatically generated

Chart

Description automatically generated

Model 3: model 3 used 20 fold cross validation and was stratified sampling. The Logistic Regression selections were set to default as changing those parameters did not seem to increase accuracy or AUC score (from testing and running those options, no screenshot shown for those test models). This model had the best accuracy at 89.40% and the best AUC score with 0.784. Changing the number of folds seemed to be the only way (upon my testing) to increase the performance metrics but more options could be tested later on (not done here due to time constraints).

A picture containing diagram

Description automatically generated

A picture containing diagram

Description automatically generated

Graphical user interface, application

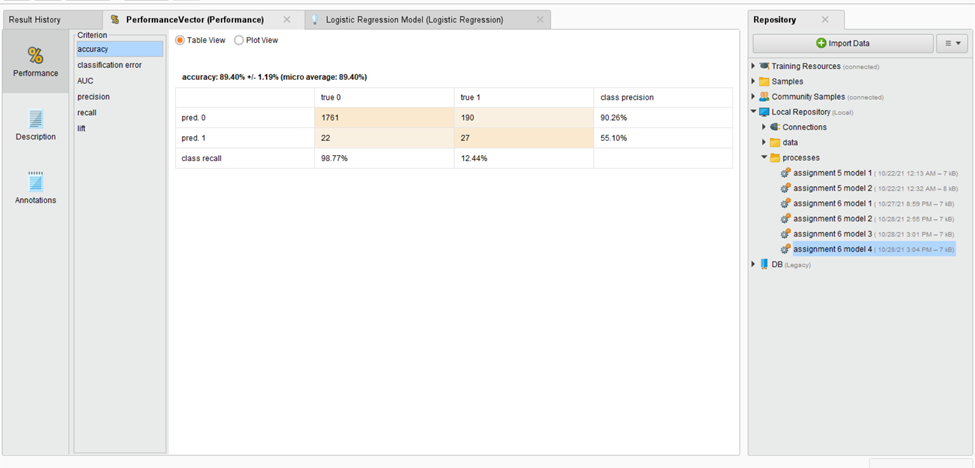
Description automatically generated

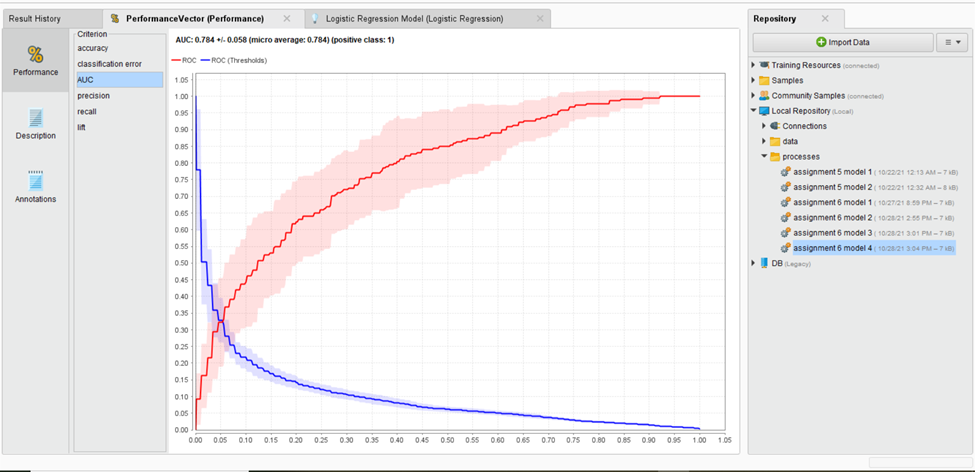
Chart

Description automatically generated

1. Select the best model - Provide screen shots of your model selection criteria – confusion matrix (accuracy, precision, recall) and AUC, ROC chart.

Model 3 was best model: As stated previously accuracy was the highest as well as AUC. Recall was higher in Model 3 compared to the other models but precision was lower 55% model 3 to 57% to model 1 and 2, but I still chose model 3 as accuracy and AUC were better.





1. Apply New Data Scoring to the best model process – provide screen shot of the Design process

Diagram

Description automatically generated

1. Provide screen shot of the prediction results on the New Data

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Generally Lift charts are shown for binomial classification, I could not find how to display a lift chart in RapidMiner but there was a lift option in the performance options that I provide a screenshot of the lift of the best model below.**

Graphical user interface

Description automatically generated with low confidence

1. **Explain and provide screen shots of steps f) through i).**