Florida Atlantic University
COLLEGE OF BUSINESS

ISM 6136 – Datamining/Predictive Analytics

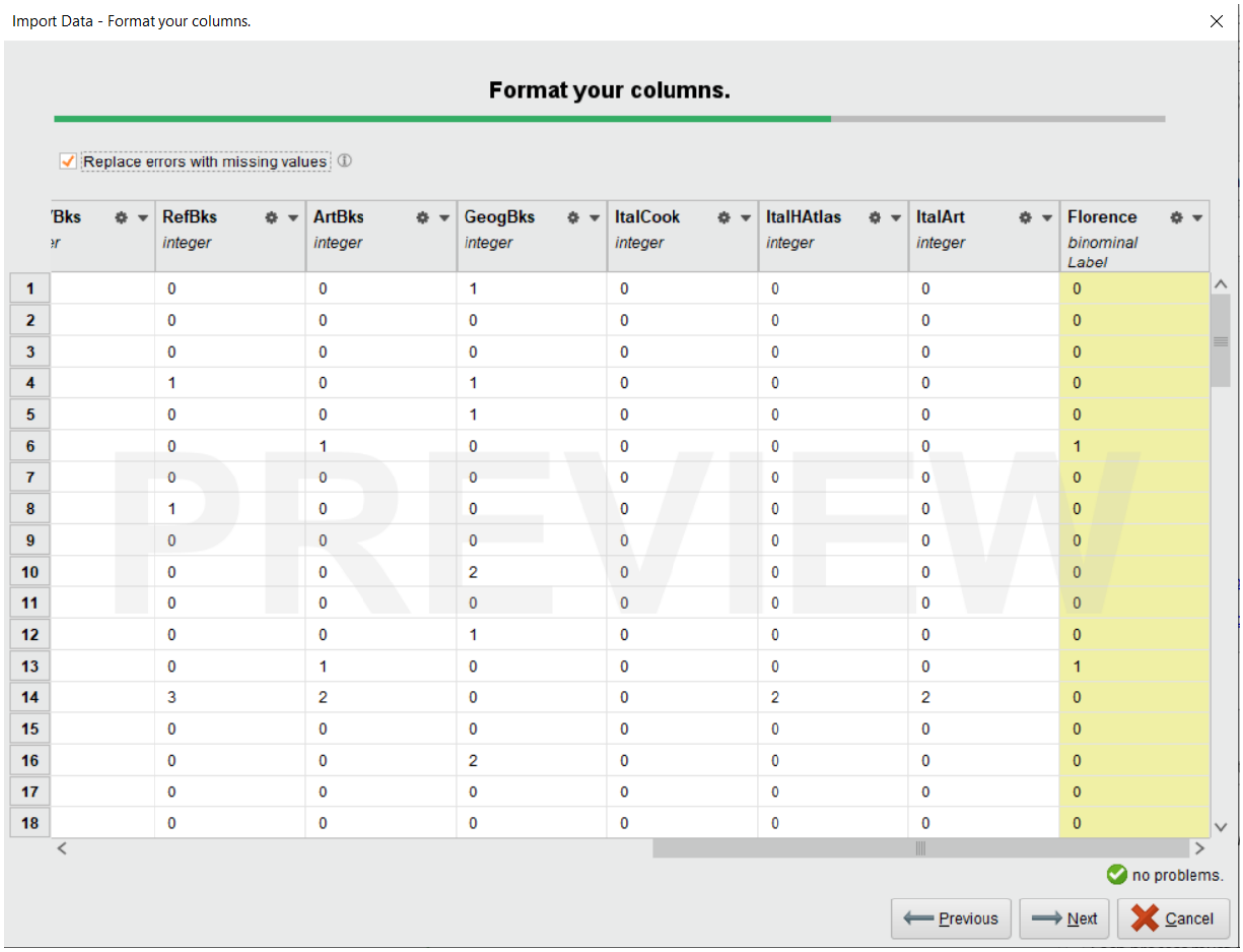
Class Assignment 6
5 points
Jacob Perrone

TASK: Performing predictive analytics using Logistic Regression in RapidMiner

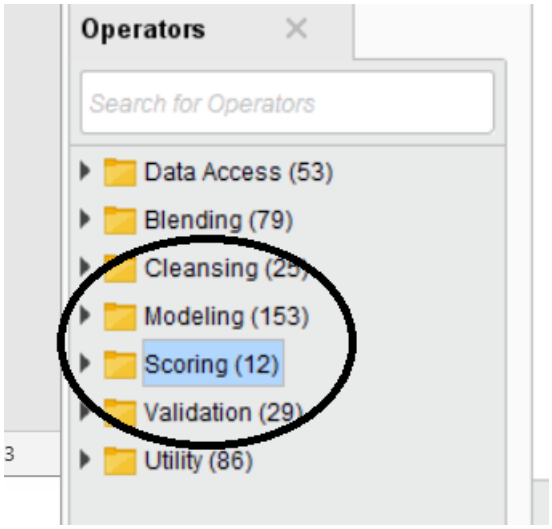
Perform the following data mining steps using Logistic Algorithm using RapidMiner and predict based on the past data which of the new readers in the Charles Book club will buy the Florence book 'Yes – 1 ' or 'No – 0' .

Follow the datamining steps below:

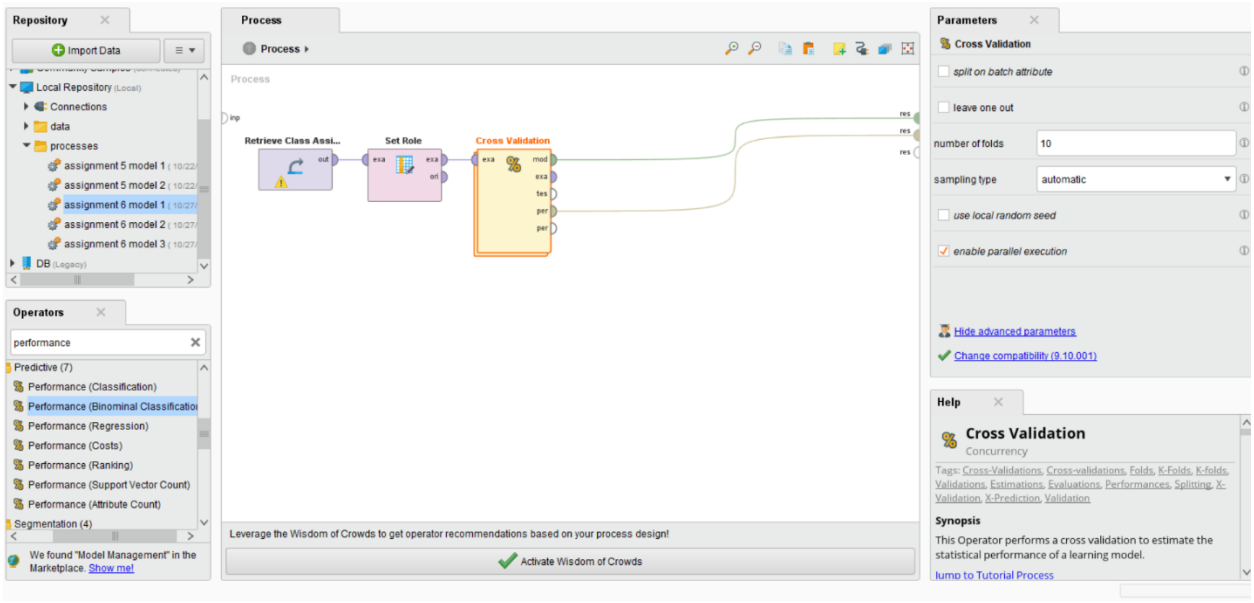
a) Understand the problem and purpose of data mining task
b) Import the dataset into RapidMiner
c) Explore, clean and preprocess data
d) Cleanup or do not select any column that is not a predictor
    1. Check 'Replace errors by missing values'
    2. Check Statistics and look for any missing values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

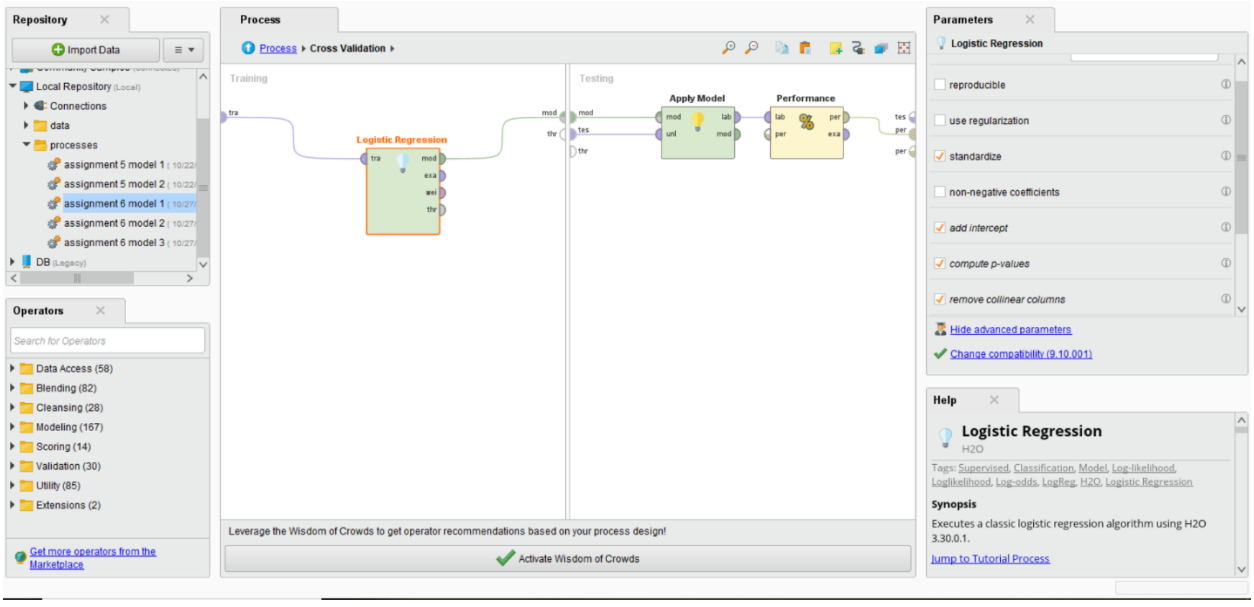


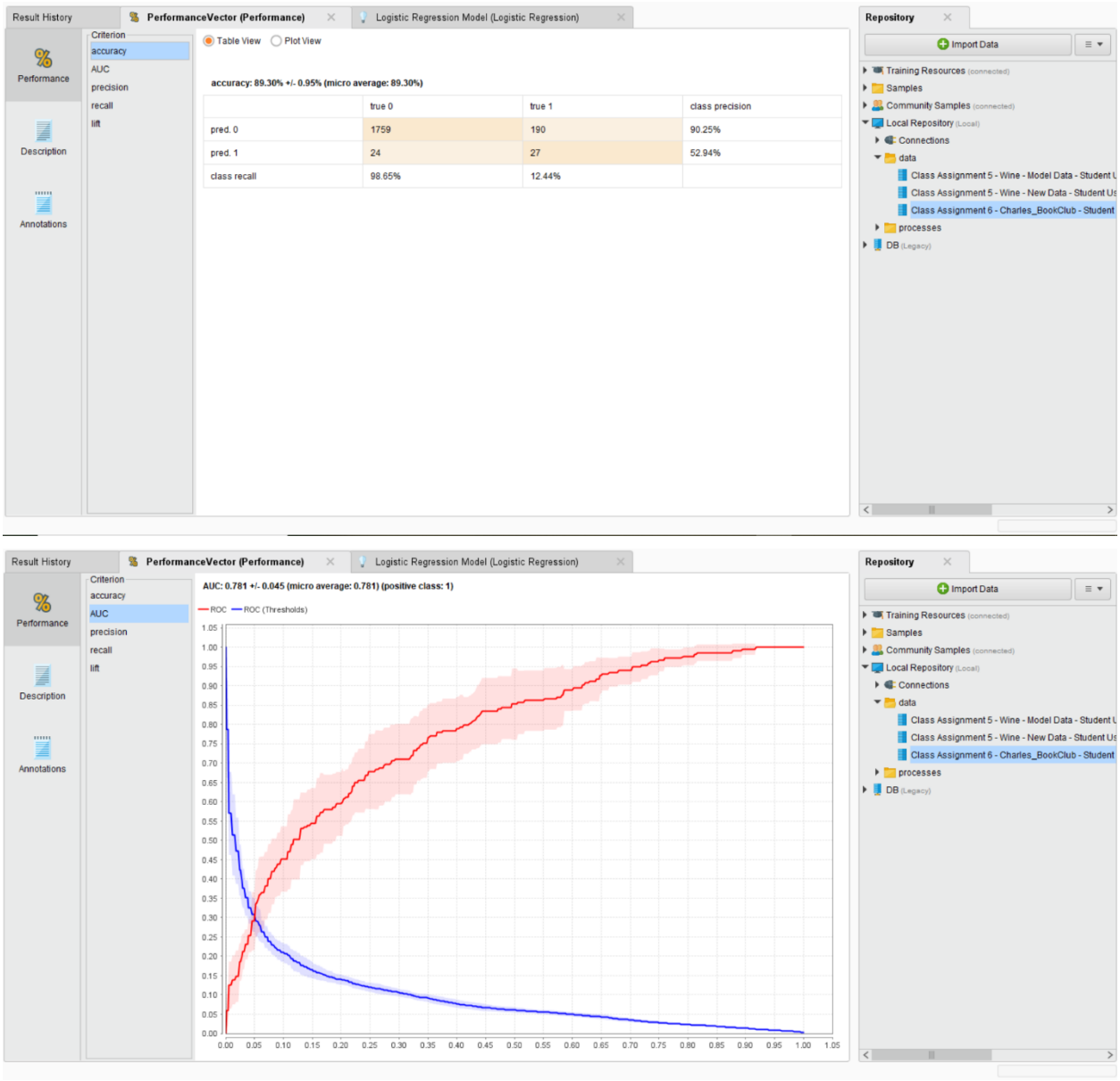
e) Design your process using appropriate operators. Provide screen shot of the overall design.



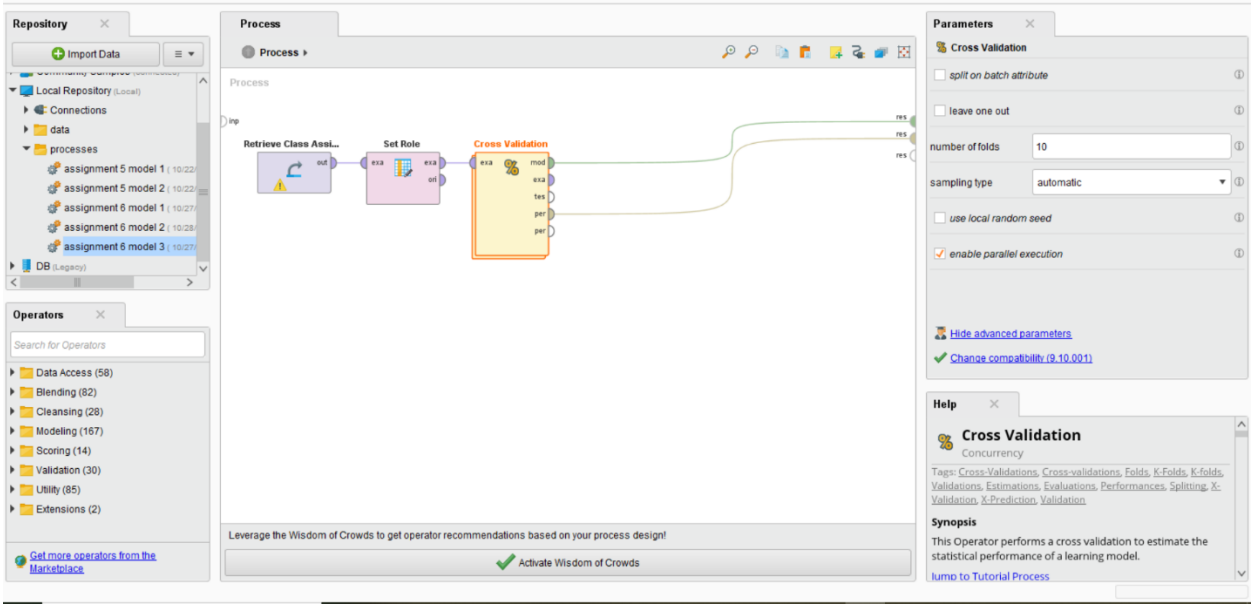
f) **Build three models** – save each of the models into a separate 'process'
Provide screen shots of each of the model settings – to show me difference in each model – You can even present this in Table form.
Model 1: model 1 used 10 fold cross validation and default selections for Logistic Regression 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.

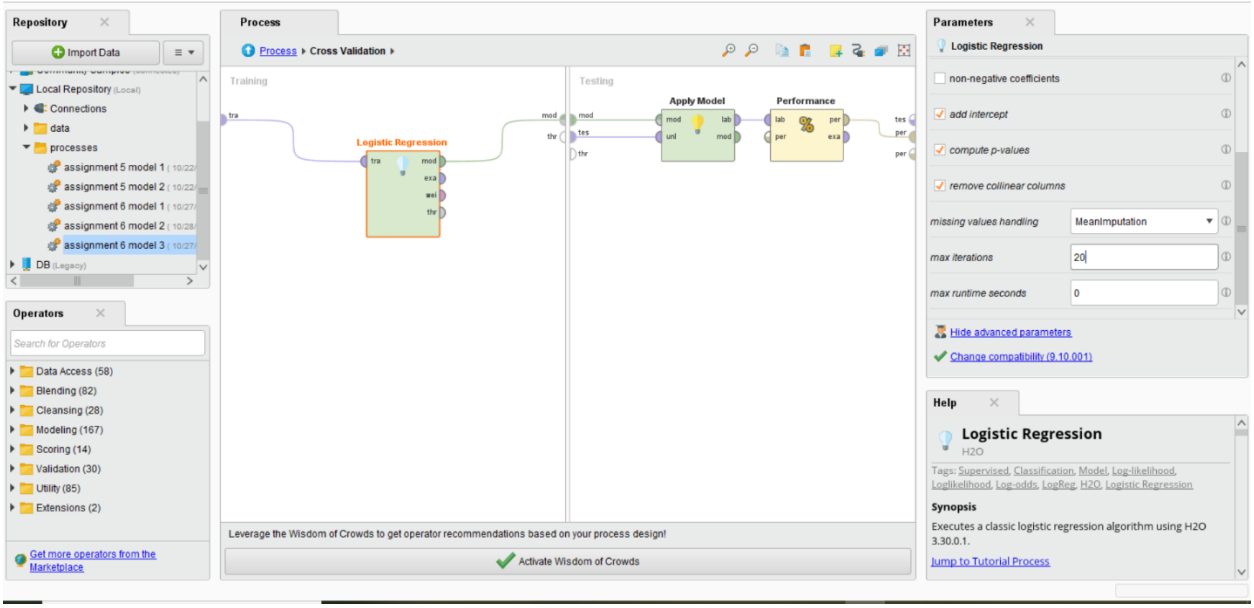


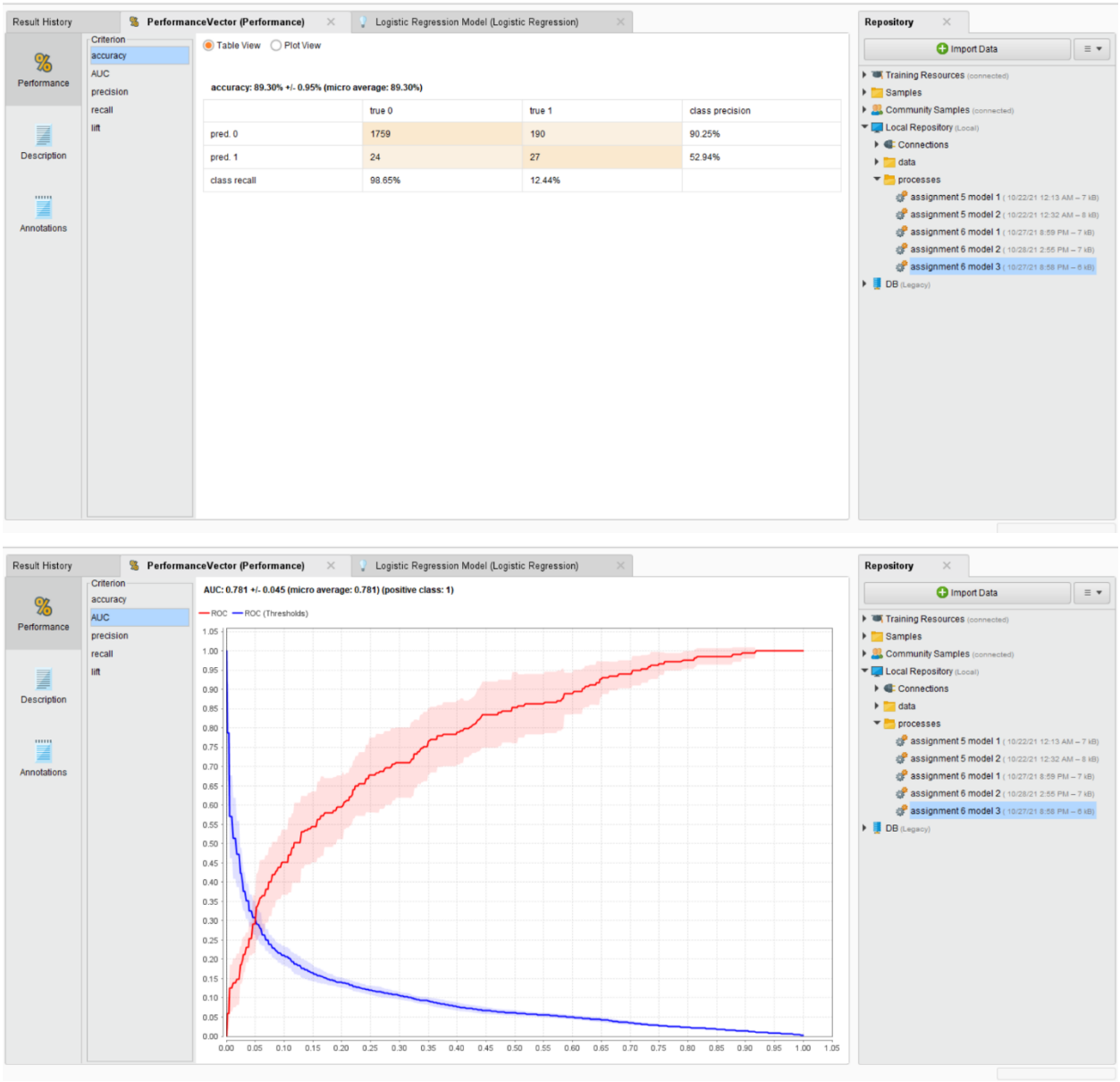




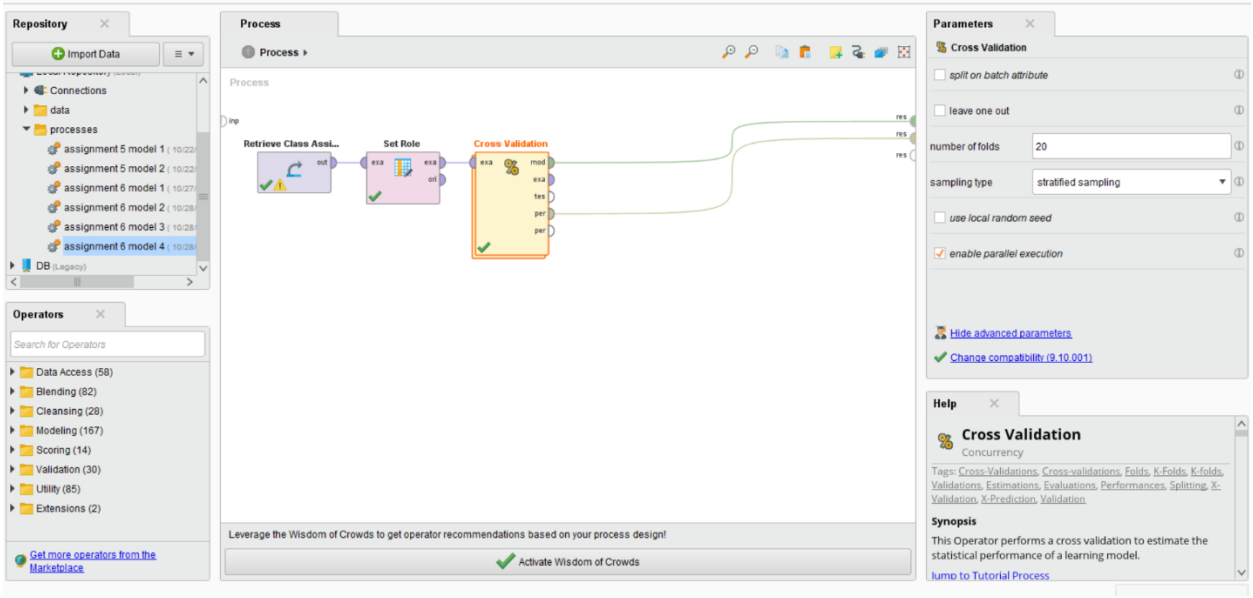
Model 2: model 2 also used 10 fold cross validation and the same default selection options for Logistic Regression except I changed the number of iterations to 20. This model had the same performance as model 1 in terms of accuracy, and AUC score.

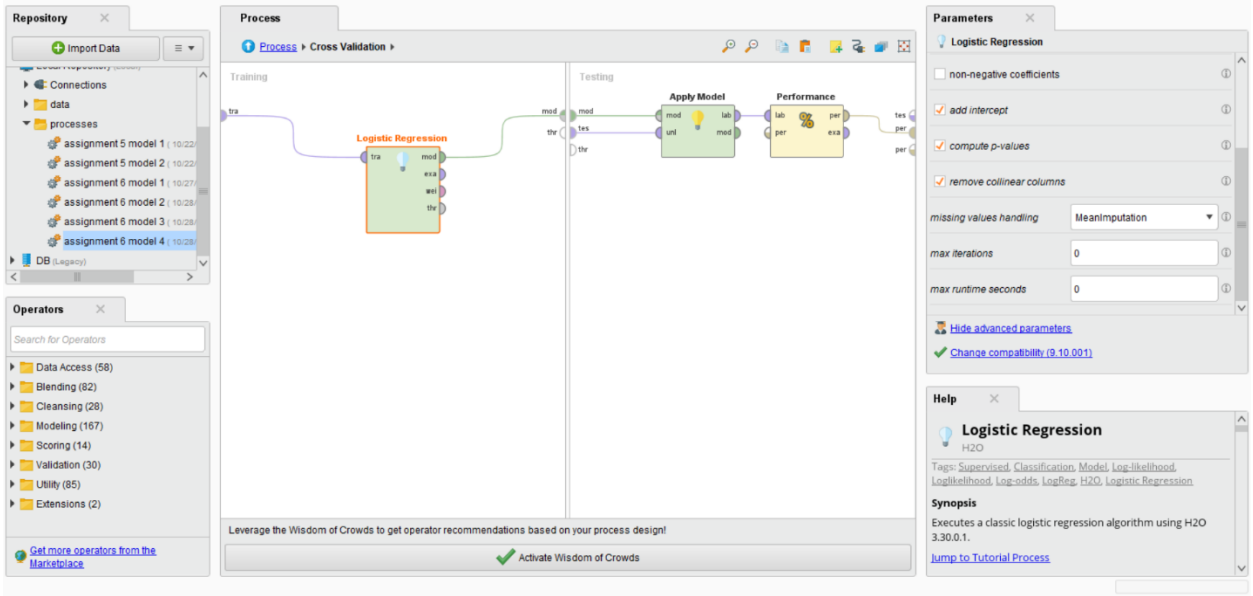

Repository
Import Data
Local Repository (Local)
Connections
data
processes
assignment 5 model 1
assignment 5 model 2
assignment 6 model 1
assignment 6 model 2
assignment 6 model 3
DB (Legacy)
Operators
Search for Operators
Data Access (58)
Blending (82)
Cleansing (28)
Modeling (167)
Scoring (14)
Validation (30)
Utility (85)
Extensions (2)
Get more operators from the Marketplace
Process
Process
Retrieve Class Assi...
Set Role
Cross Validation
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Cross Validation
split on batch attribute
leave one out
number of folds
10
sampling type
automatic
use local random seed
enable parallel execution
Hide advanced parameters
Change compatibility (9.10.001)
Help
Cross Validation
Concurrency
Tags: Cross-Validations, Cross-validations, Folds, K-Folds, K-folds, Validations, Estimations, Evaluations, Performances, Splitting, X-Validation, X-Prediction, Validation
Synopsis
This Operator performs a cross validation to estimate the statistical performance of a learning model.
Jump to Tutorial Process

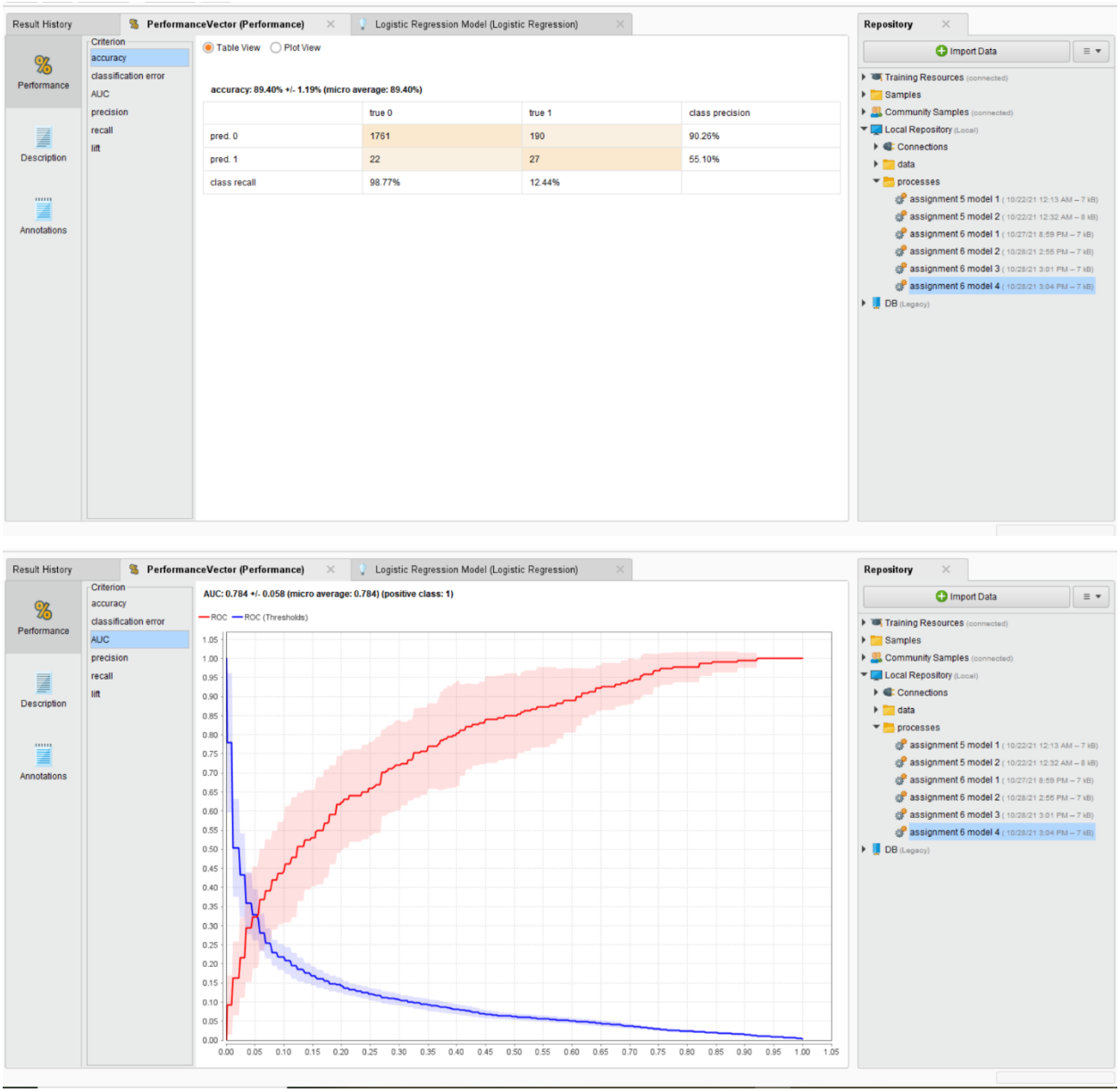

Repository
Import Data
Local Repository (Local)
Connections
data
processes
assignment 5 model 1
assignment 5 model 2
assignment 6 model 1
assignment 6 model 2
assignment 6 model 3
DB (Legacy)
Operators
Search for Operators
Data Access (58)
Blending (82)
Cleansing (28)
Modeling (167)
Scoring (14)
Validation (30)
Utility (85)
Extensions (2)
Get more operators from the Marketplace
Process
Process ▸ Cross Validation ▸
Training
Testing
Logistic Regression
Apply Model
Performance
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Logistic Regression
non-negative coefficients
add intercept
compute p-values
remove collinear columns
missing values handling
MeanImputation
max iterations
20
max runtime seconds
0
Hide advanced parameters
Change compatibility (9.10.001)
Help
Logistic Regression
H2O
Tags: Supervised, Classification, Model, Log-likelihood, Loglikelihood, Log-odds, LogReg, H2O, Logistic Regression
Synopsis
Executes a classic logistic regression algorithm using H2O 3.30.0.1.
Jump to Tutorial Process



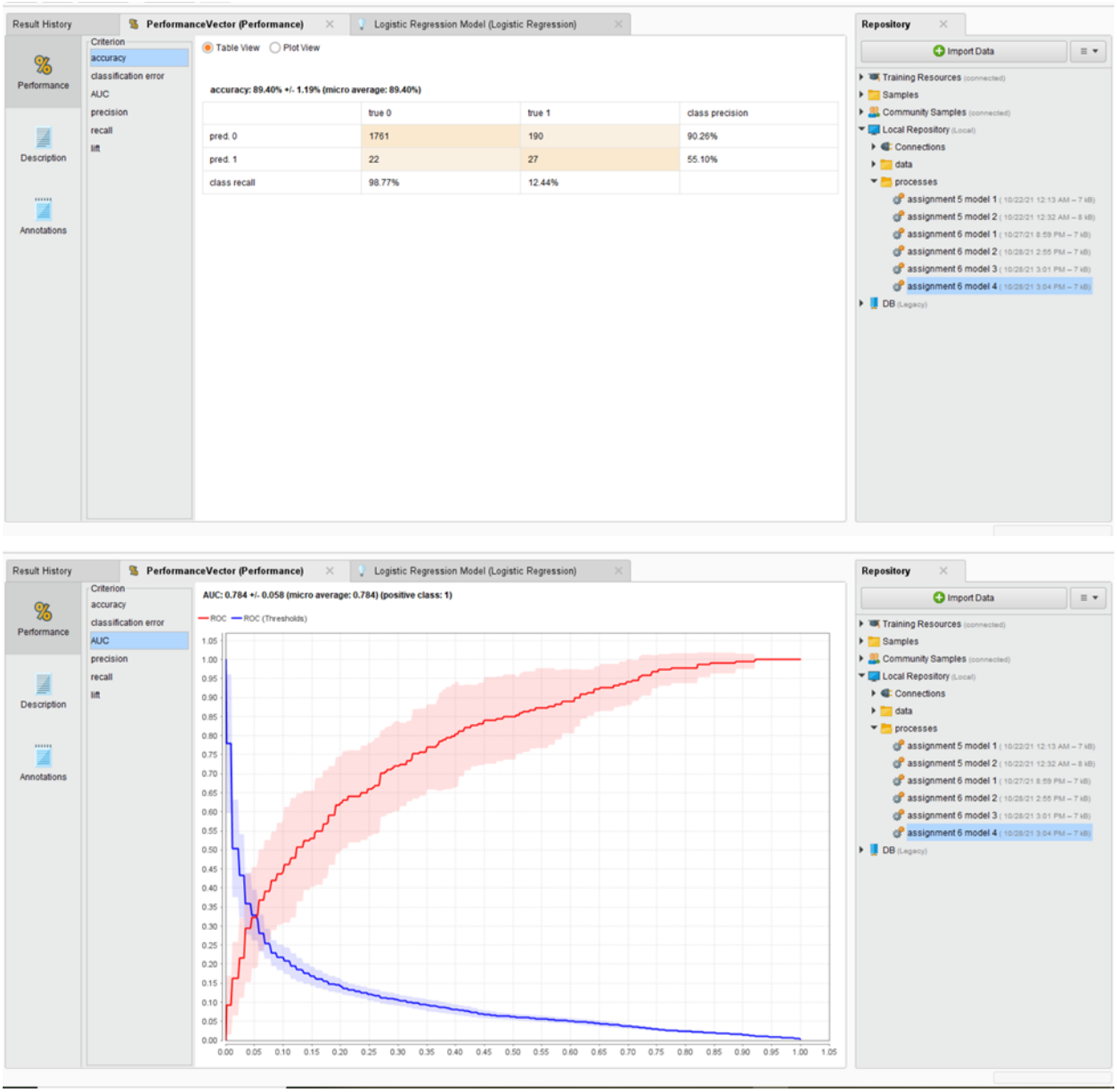
Model 3: model 3 used 20 fold cross validation and was stratified sampling. The Logistic Regression selections were set to default as changing those parameters did not seem to increase accuracy or AUC score (from testing and running those options, no screenshot shown for those test models). This model 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).


Repository
Import Data
Connections
data
processes
assignment 5 model 1
assignment 5 model 2
assignment 6 model 1
assignment 6 model 2
assignment 6 model 3
assignment 6 model 4
DB (Legacy)
Operators
Search for Operators
Data Access (58)
Blending (82)
Cleansing (28)
Modeling (167)
Scoring (14)
Validation (30)
Utility (85)
Extensions (2)
Get more operators from the Marketplace
Process
Process
Retrieve Class Assi...
Set Role
Cross Validation
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Cross Validation
split on batch attribute
leave one out
number of folds
20
sampling type
stratified sampling
use local random seed
enable parallel execution
Hide advanced parameters
Change compatibility (9.10.001)
Help
Cross Validation
Concurrency
Synopsis
This Operator performs a cross validation to estimate the statistical performance of a learning model.
Jump to Tutorial Process

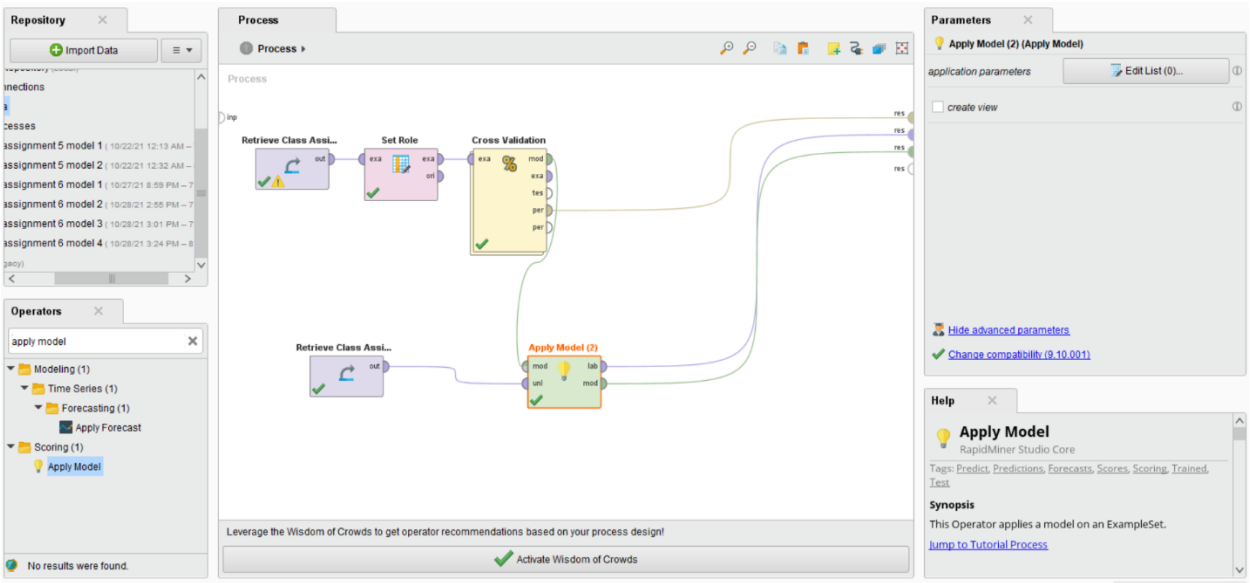

Repository
Import Data
Connections
data
processes
assignment 5 model 1
assignment 5 model 2
assignment 6 model 1
assignment 6 model 2
assignment 6 model 3
assignment 6 model 4
DB (Legacy)
Operators
Search for Operators
Data Access (58)
Blending (82)
Cleansing (28)
Modeling (167)
Scoring (14)
Validation (30)
Utility (85)
Extensions (2)
Get more operators from the Marketplace
Process
Process ▸ Cross Validation ▸
Training
Testing
Logistic Regression
Apply Model
Performance
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Logistic Regression
non-negative coefficients
add intercept
compute p-values
remove collinear columns
missing values handling
MeanImputation
max iterations
0
max runtime seconds
0
Hide advanced parameters
Change compatibility (9.10.001)
Help
Logistic Regression
H2O
Synopsis
Executes a classic logistic regression algorithm using H2O 3.30.0.1.
Jump to Tutorial Process



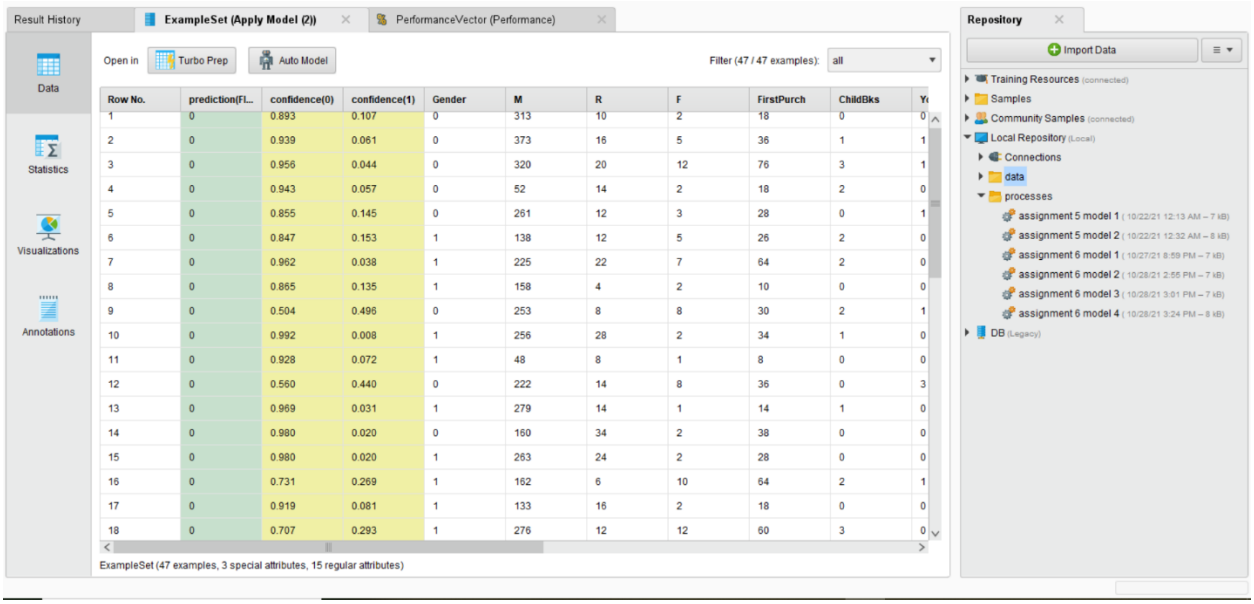
g) Select the best model - Provide screen shots of your model selection criteria – confusion matrix (accuracy, precision, recall) and AUC, ROC chart.
Model 3 was best model: As stated previously accuracy was the highest as well as AUC. Recall was higher in Model 3 compared to the other models but precision was lower 55% model 3 to 57% to model 1 and 2, but I still chose model 3 as accuracy and AUC were better.

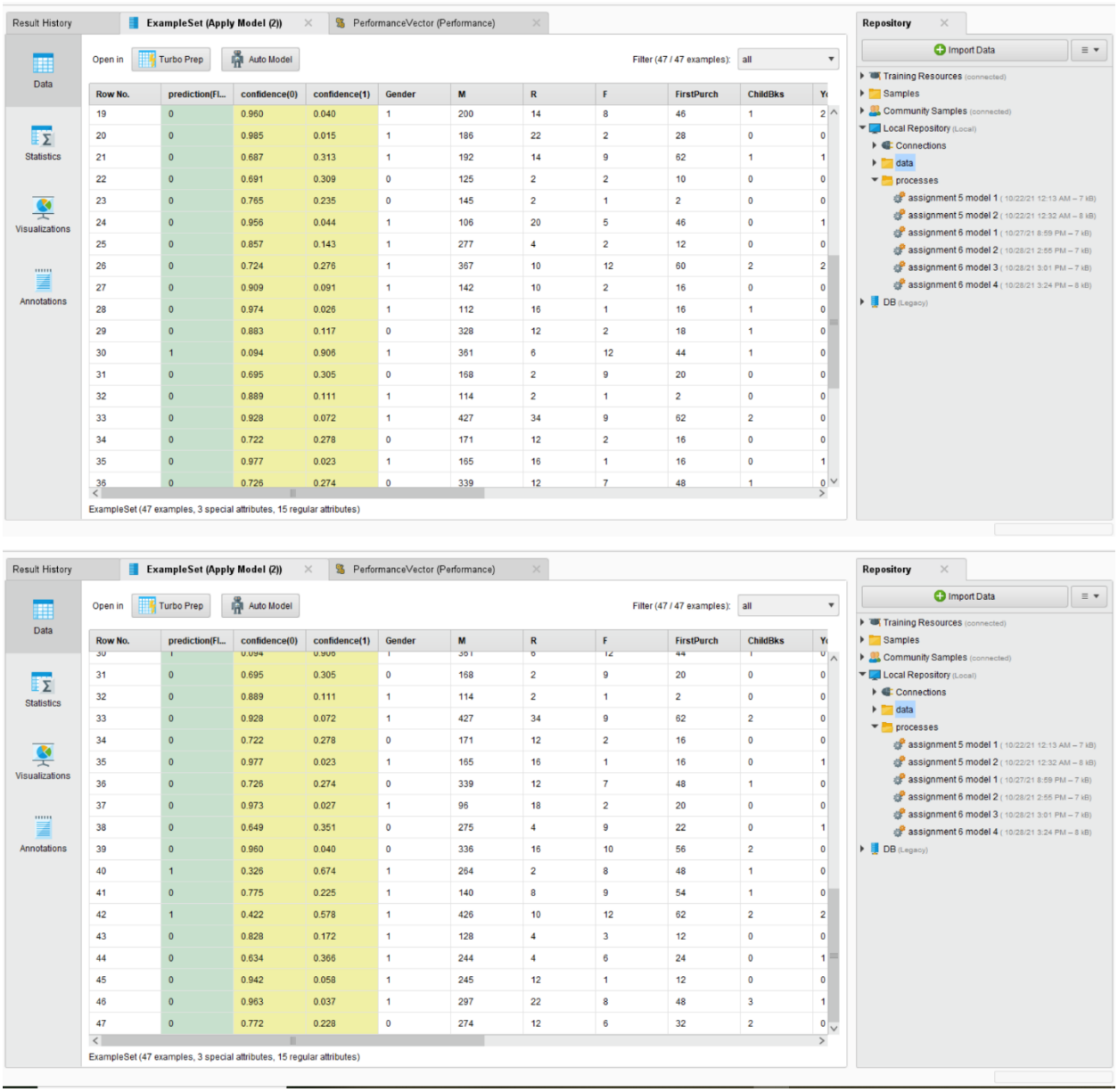


h) Apply New Data Scoring to the best model process – provide screen shot of the Design process

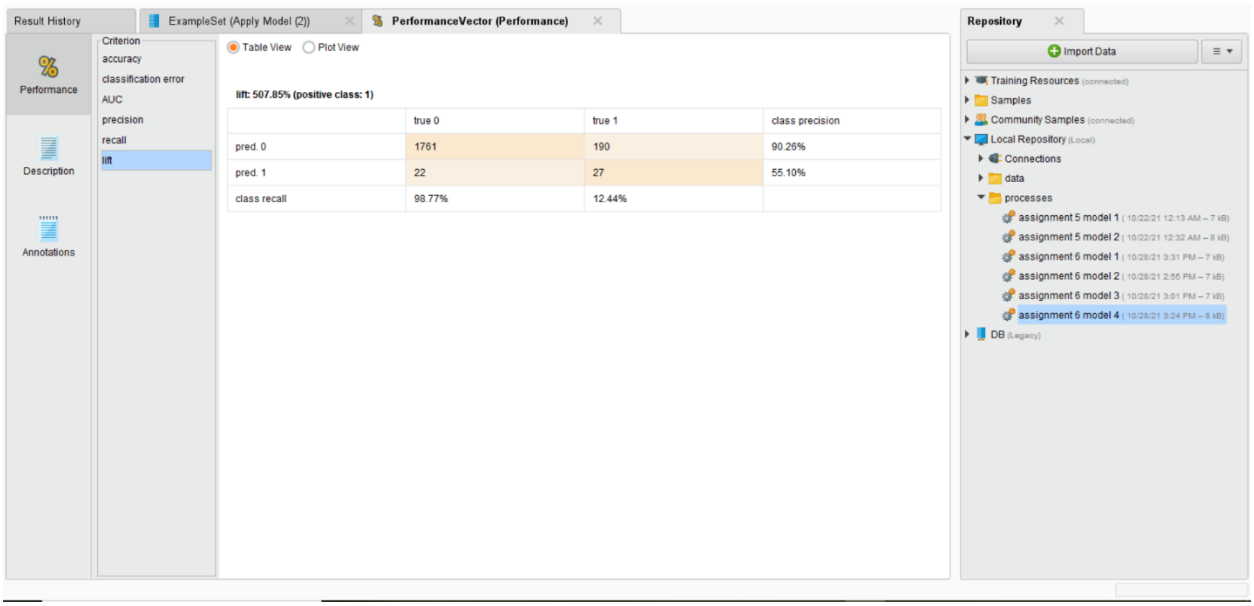


i) Provide screen shot of the prediction results on the New Data





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.



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