

**ISM 6136 – Datamining/Predictive Analytics**

**Class Assignment 6**

**5 points**

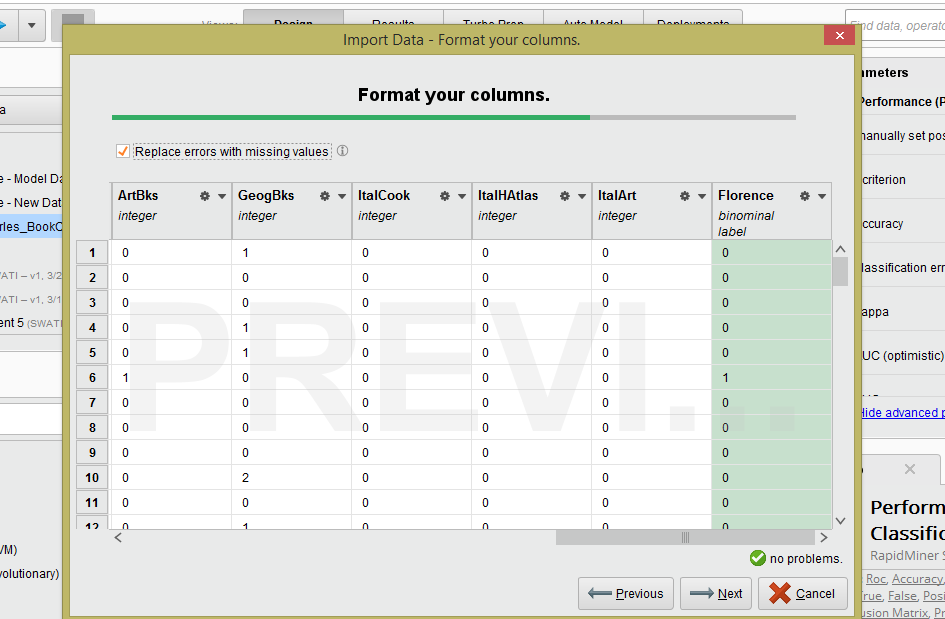
**TASK: Performing predictive analytics using Logistic Regression in RapidMiner**

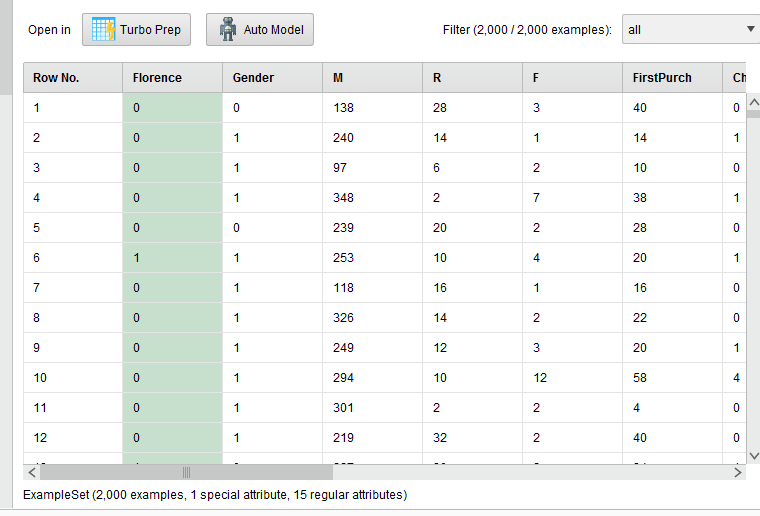
**Perform the following data mining steps using Logistic Algorithm using RapidMiner and predict based on the past data which of the new readers in the Charles Book club will buy the Florence book ‘Yes – 1 ‘ or ‘No – 0’ .**

Follow the datamining steps below:

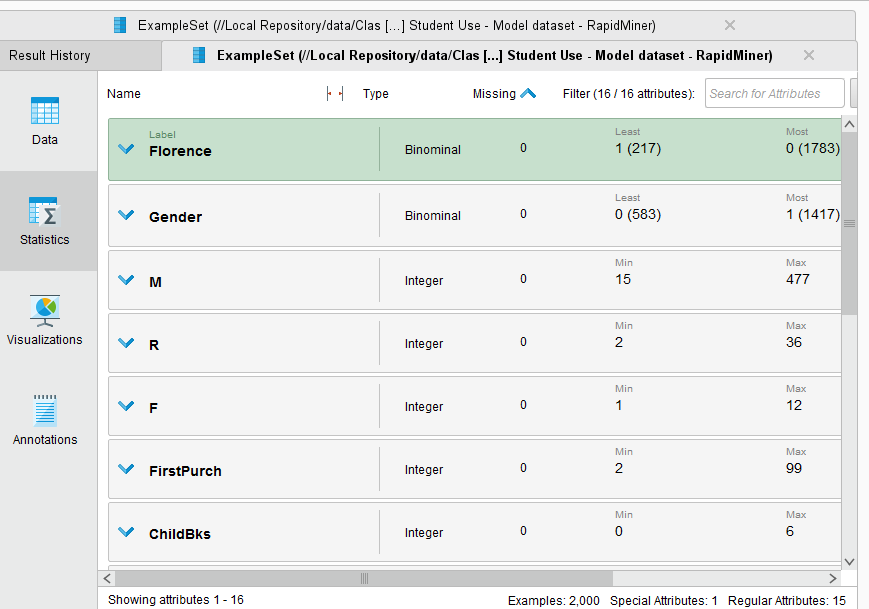
1. Understand the problem and purpose of data mining task
2. Import the dataset into RapidMiner
3. Explore, clean and preprocess data
4. Cleanup or do not select any column that is not a predictor
   1. Check ‘Replace errors by missing values’

Changed Gender + Florence to Binomial, Changed role of Florence to label. Despite this I got errors which were corrected by adding operators as per the Rapidminer’s suggestions.





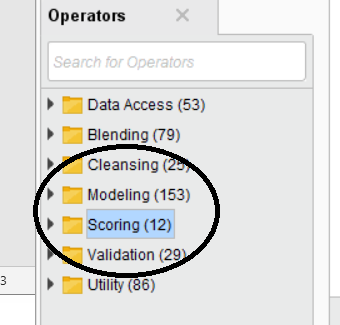
No missing values



* 1. Check Statistics and look for any missing values (if yes then you will have to add a replace with the ‘Replace Missing Values’ operator

Not required

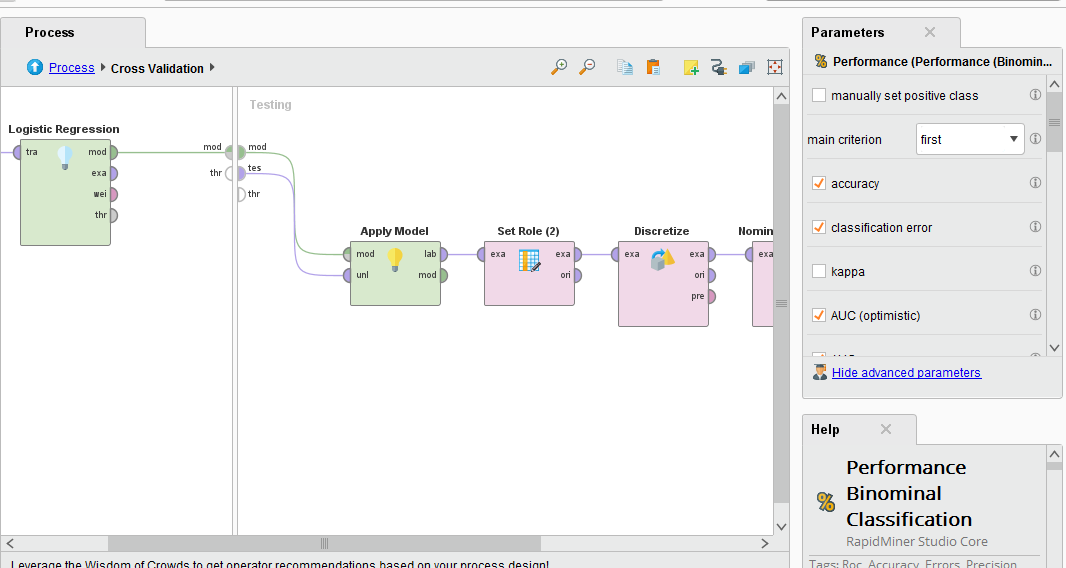
1. Design your process using appropriate operators. Provide screen shot of the overall design.

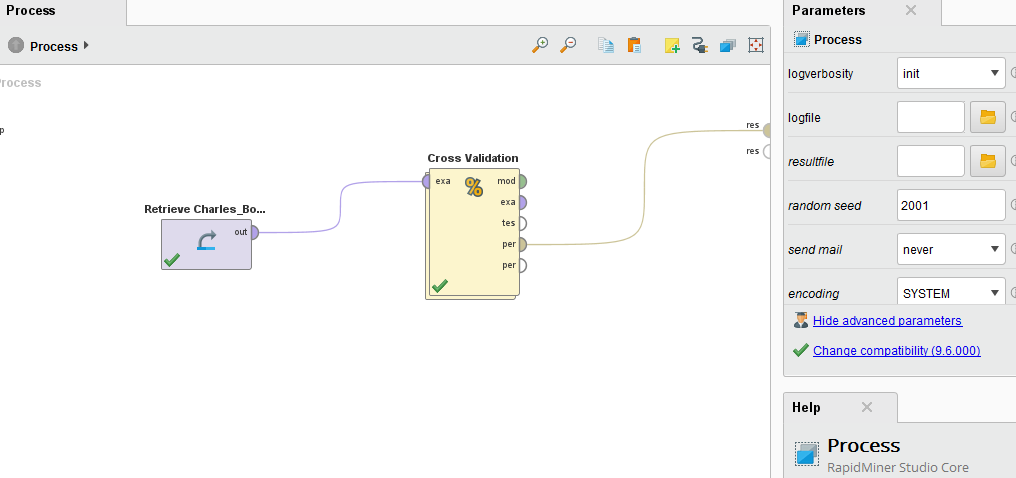


1. **Build three models** – save each of the models into a separate ‘process’

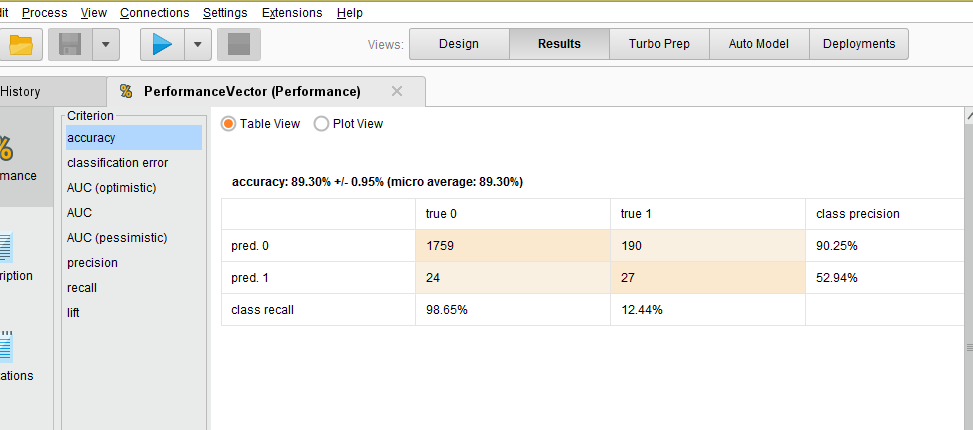
Provide screen shots of each of the model settings– to show me difference in each model – You can even present this in Table form.

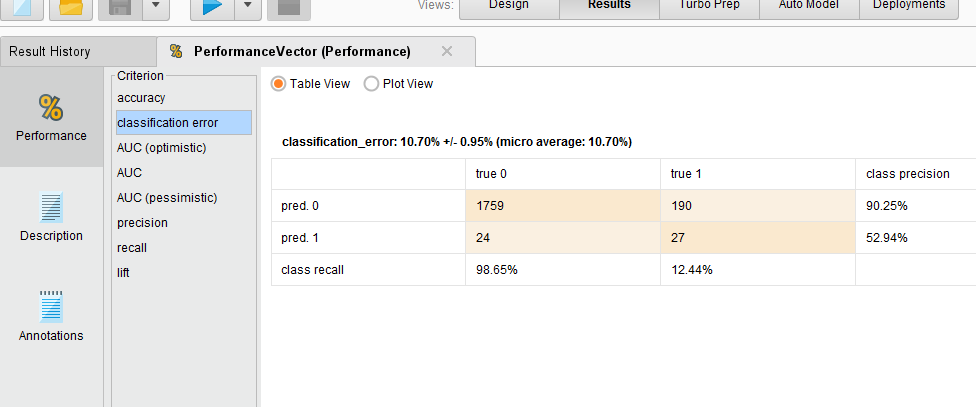
Overall Process Design

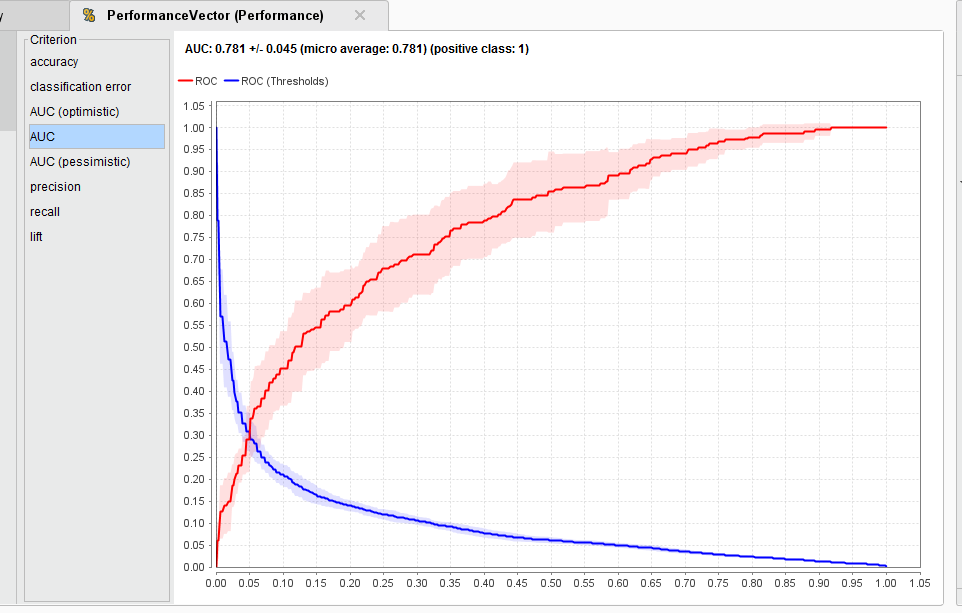


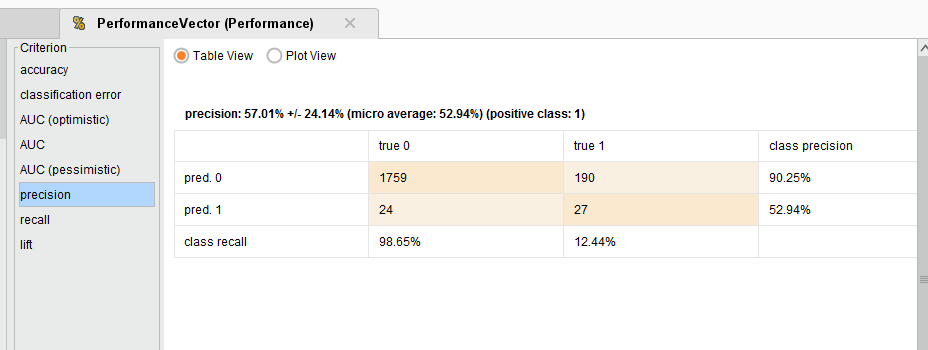


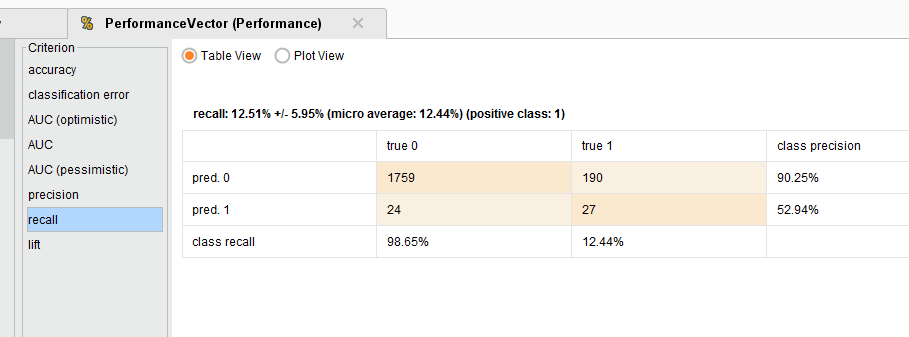
Model 1: 10 folds

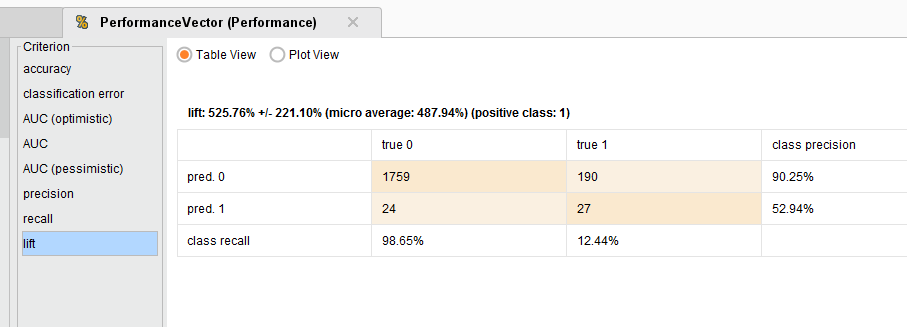




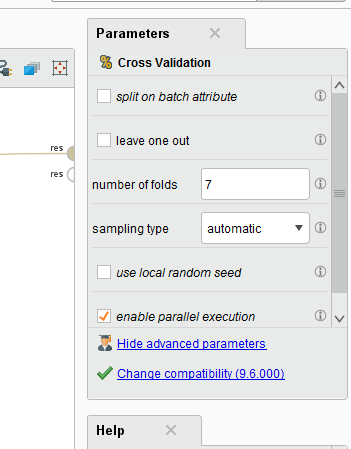


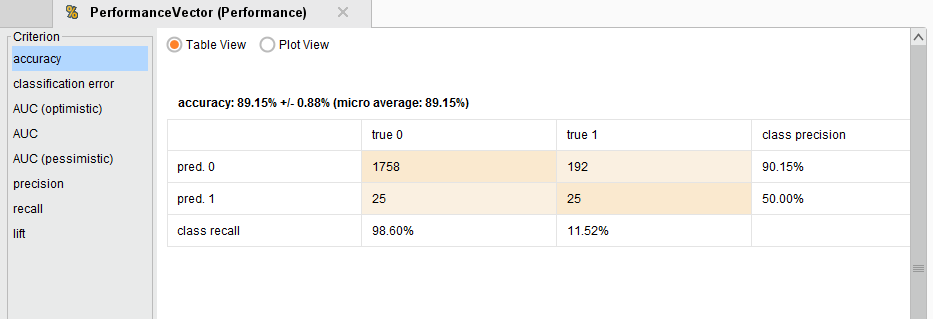


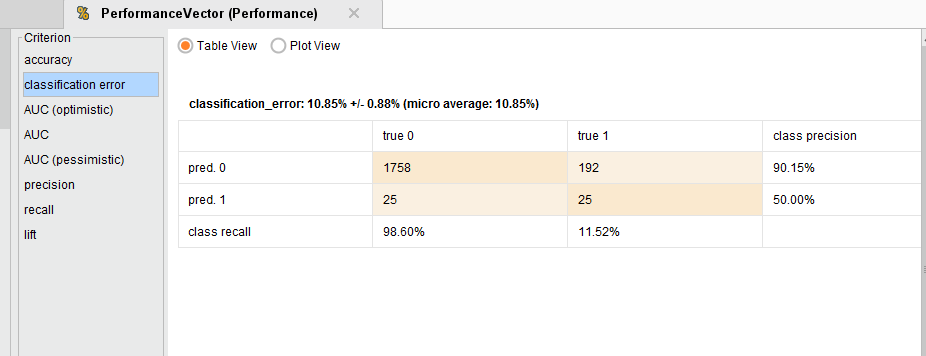


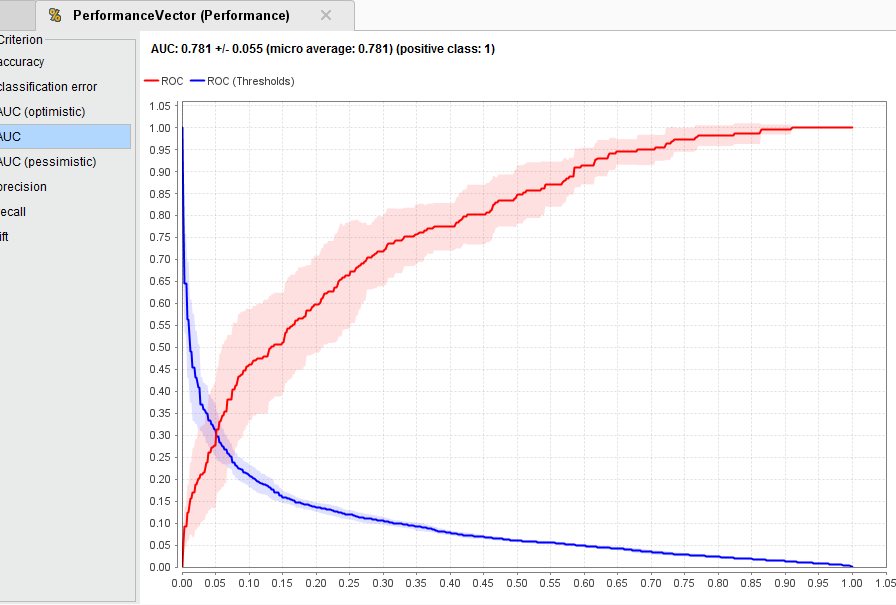


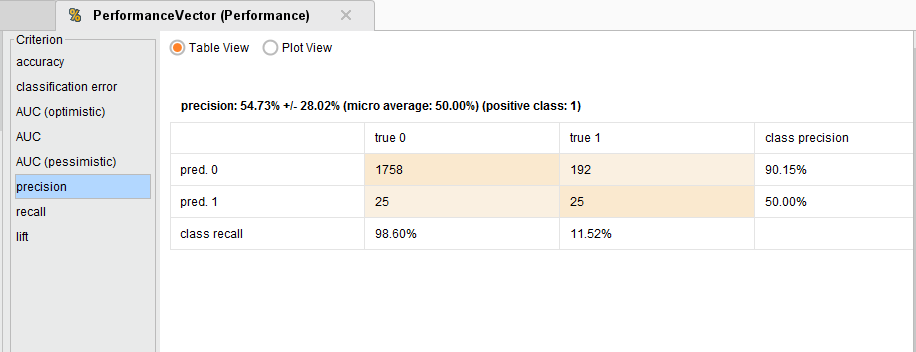
Model #2 – 7 folds

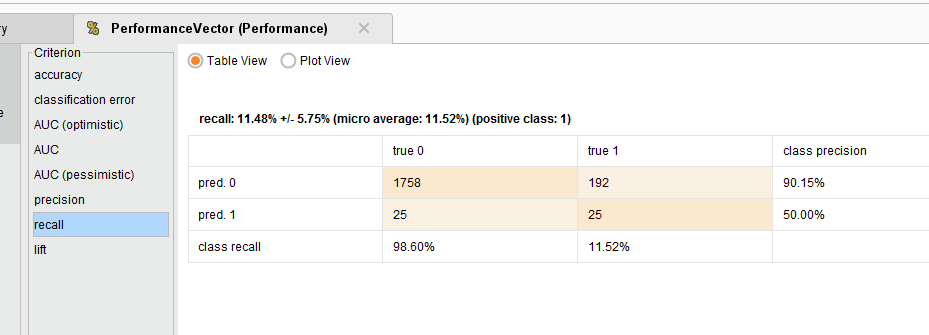


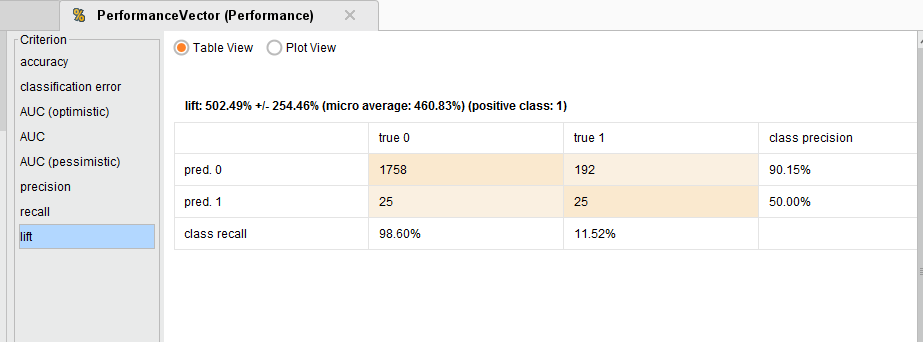




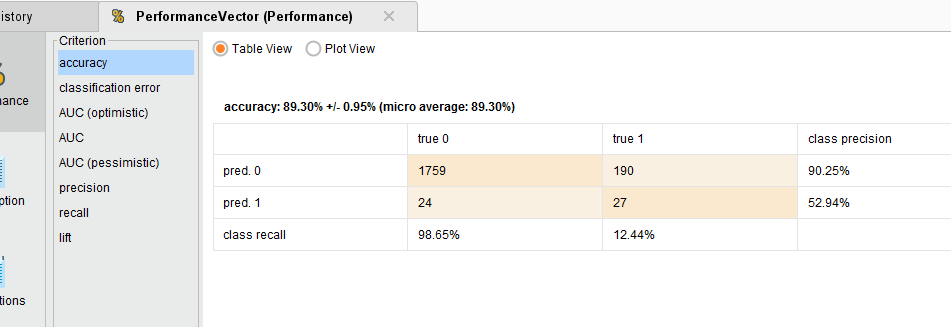


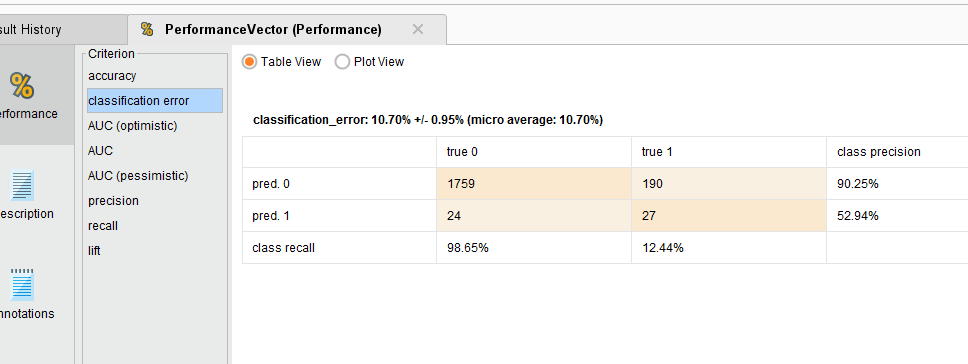


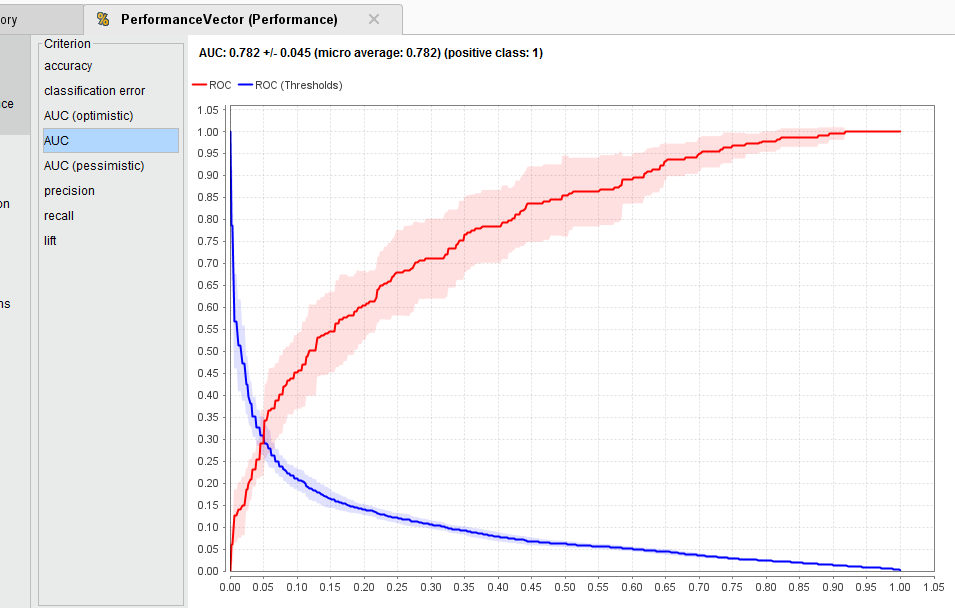


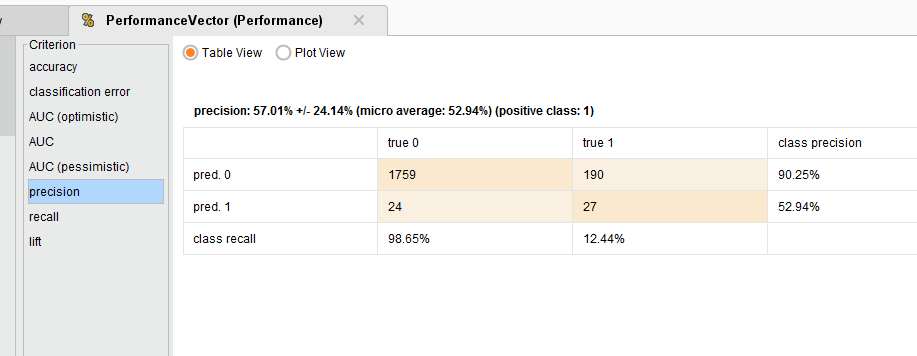


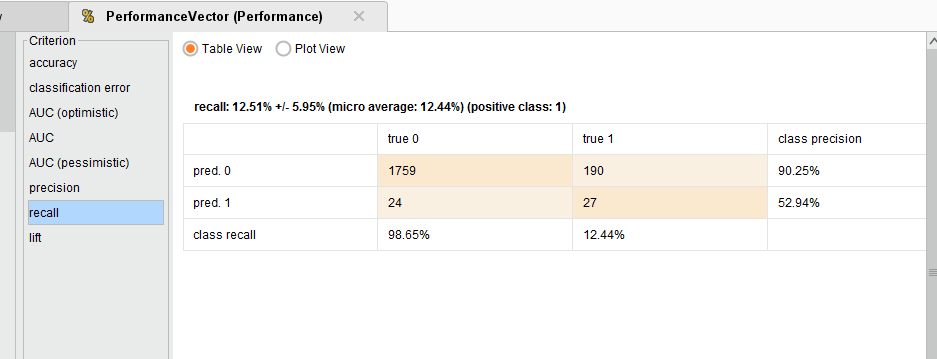
Model 3: Changed Iterations from 0 to 2, Folds: 10

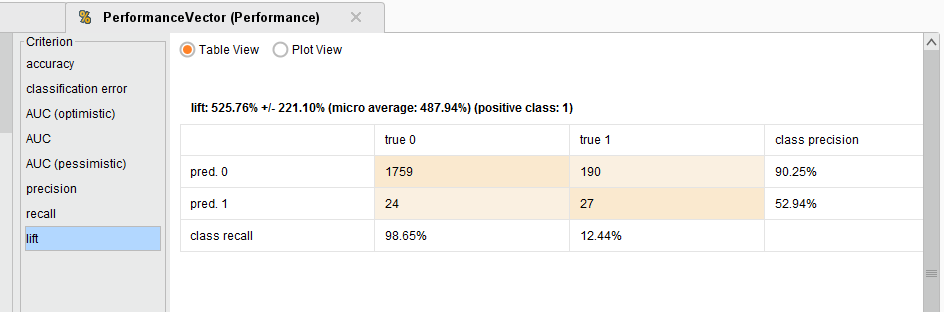




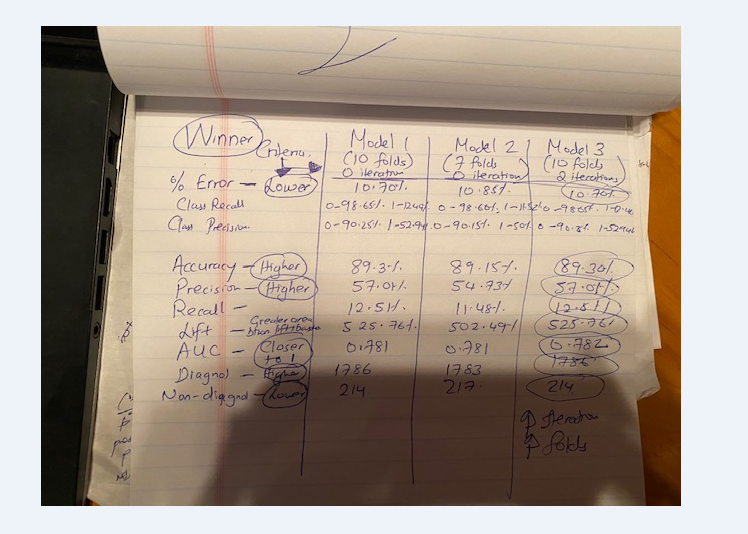








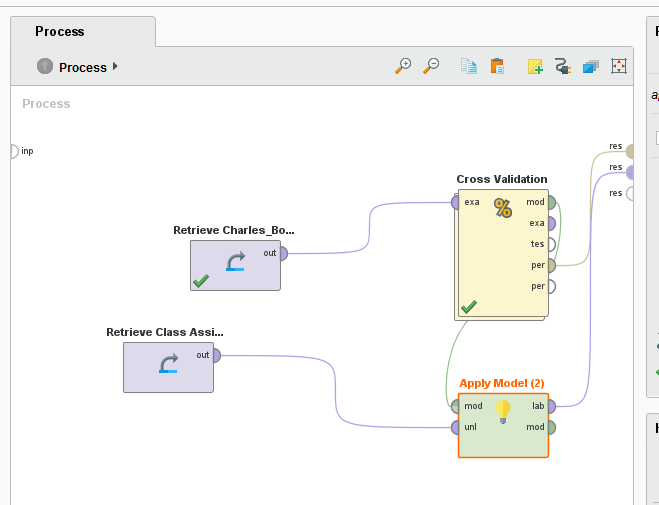
1. Select the best model - Provide screen shots of your model selection criteria – confusion matrix (accuracy, precision, recall) and AUC, ROC chart.



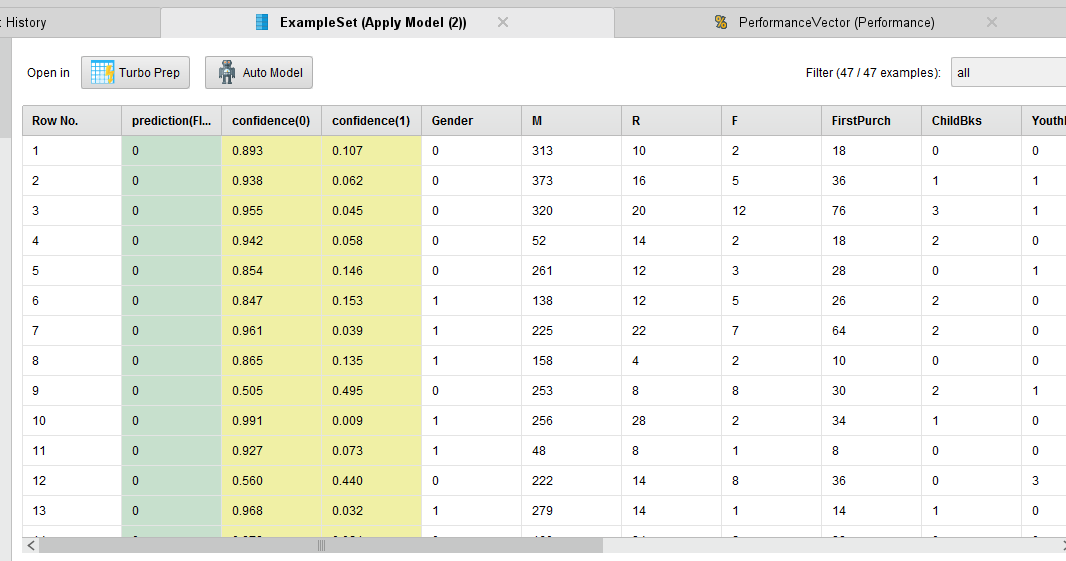
AUC Optimistic and pessimistic were same as AUC, hence only AUC was taken.

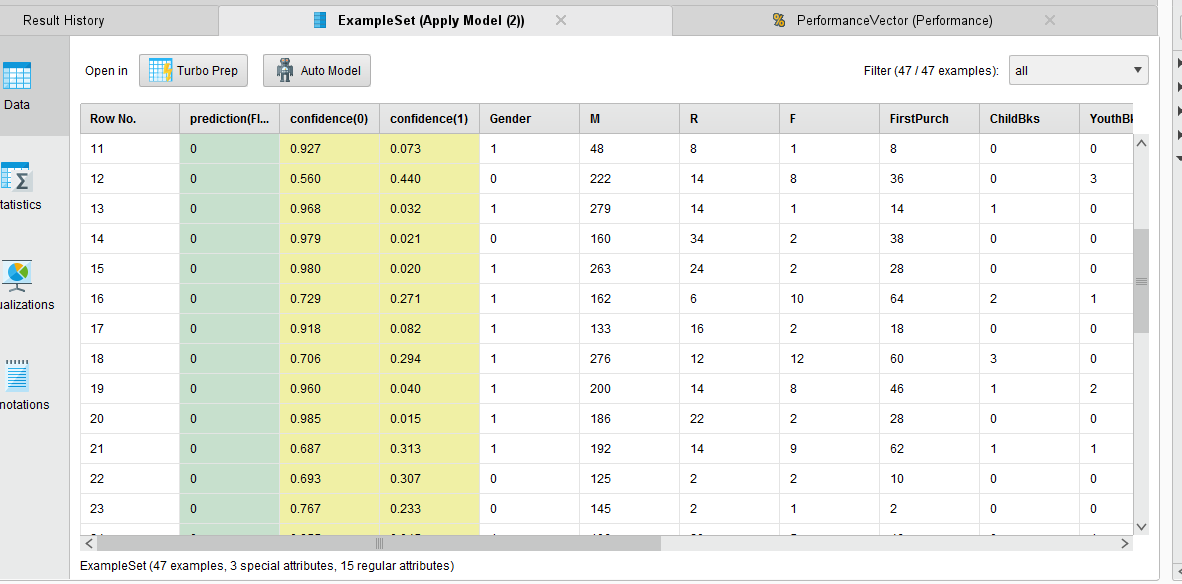
Based on the above picture comparing all three models, increasing iteration and higher folds gives better accuracy and precision and less error. Hence model 3 was chosen.

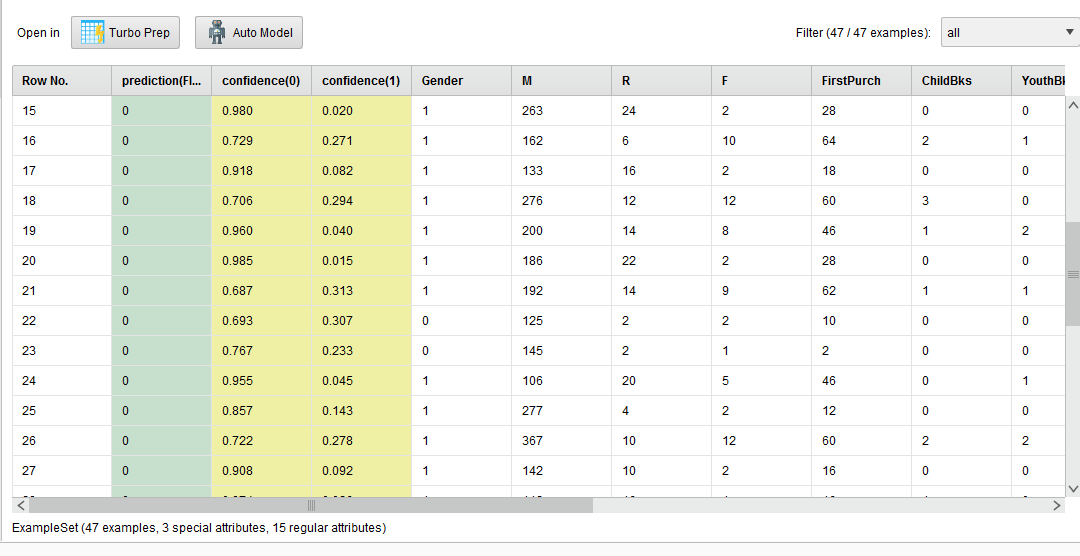
1. Apply New Data Scoring to the best model process – provide screen shot of the Design process

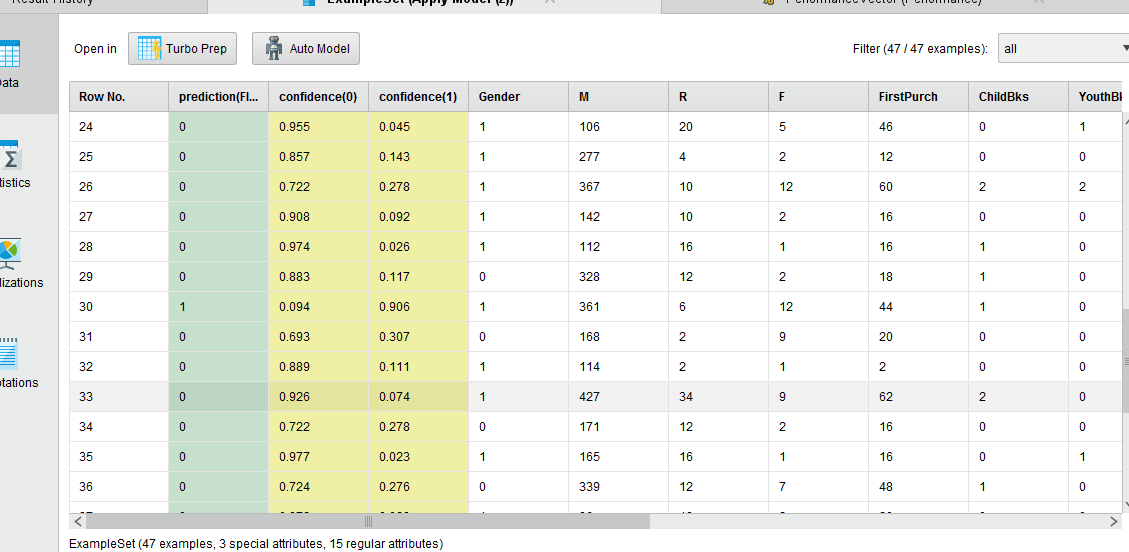


1. Provide screen shot of the prediction results on the New Data









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