



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

Format your columns.

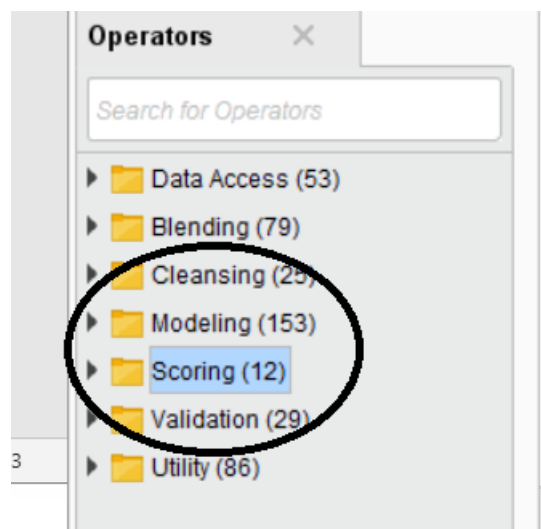
☒ Replace errors with missing values ⓘ

	Bks integer	RefBks integer	ArtBks integer	GeogBks integer	ItalCook integer	ItalHAtlas integer	ItalArt integer	Florence binominal Label
1		0	0	1	0	0	0	0
2		0	0	0	0	0	0	0
3		0	0	0	0	0	0	0
4		1	0	1	0	0	0	0
5		0	0	1	0	0	0	0
6		0	1	0	0	0	0	1
7		0	0	0	0	0	0	0
8		1	0	0	0	0	0	0
9		0	0	0	0	0	0	0
10		0	0	2	0	0	0	0
11		0	0	0	0	0	0	0
12		0	0	1	0	0	0	0
13		0	1	0	0	0	0	1
14		3	2	0	0	2	2	0
15		0	0	0	0	0	0	0
16		0	0	2	0	0	0	0
17		0	0	0	0	0	0	0
18		0	0	0	0	0	0	0

✔ no problems.

← Previous → Next ✕ Cancel

- 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.

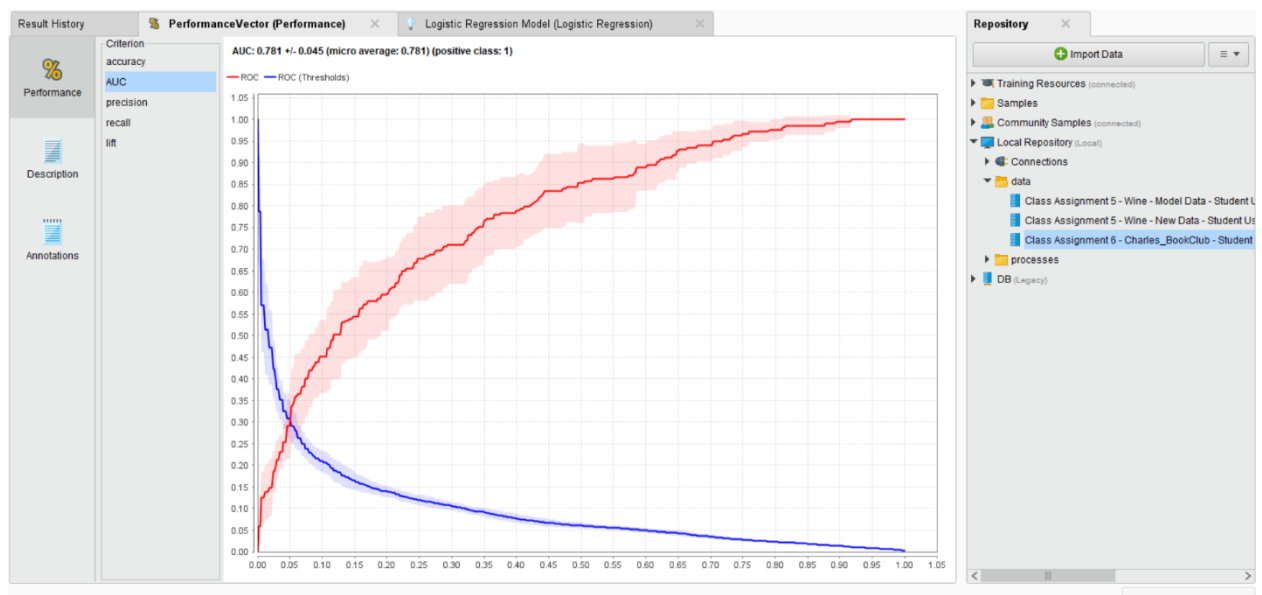
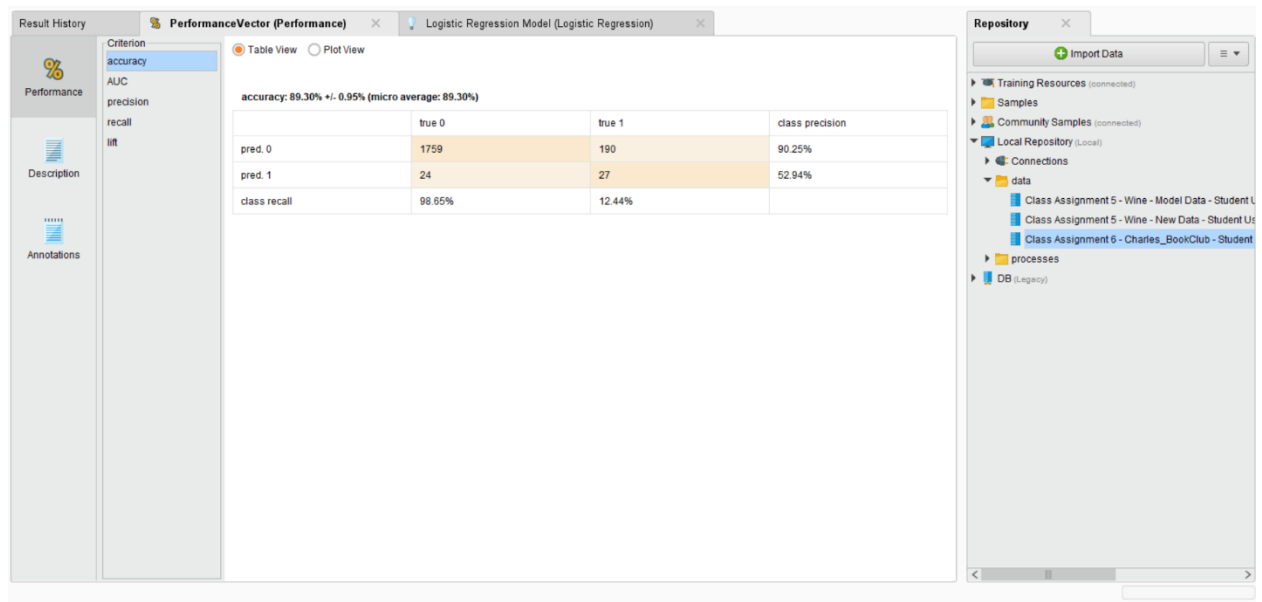
The image displays two screenshots of the H2O machine learning interface, showing the setup for Model 1.

Top Screenshot: Cross Validation Process

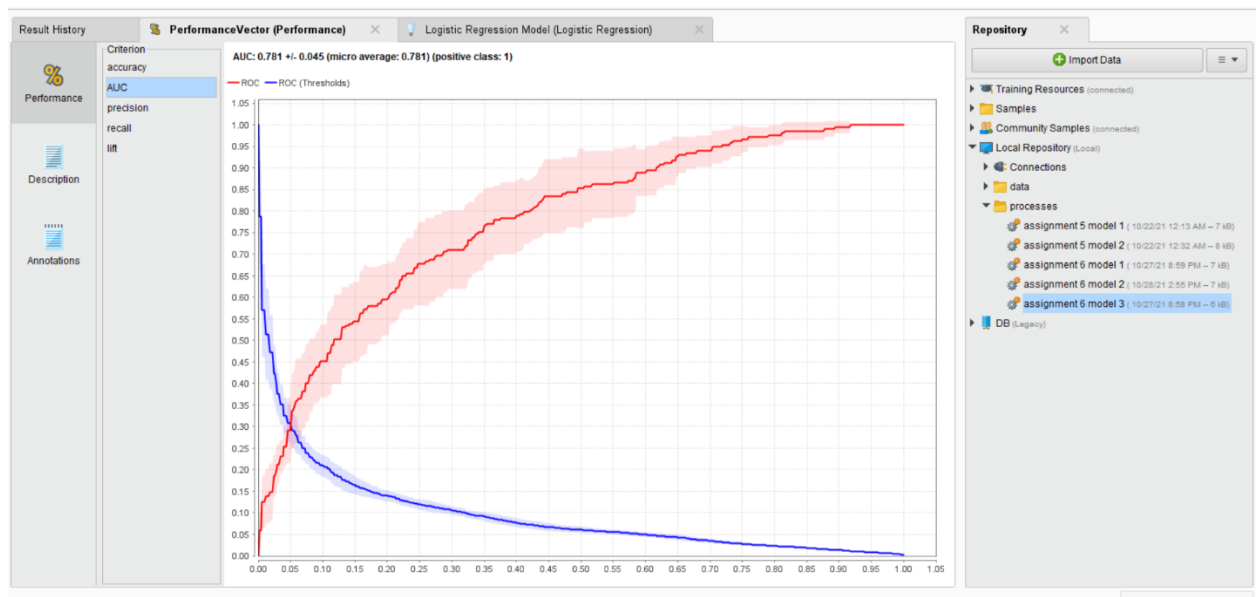
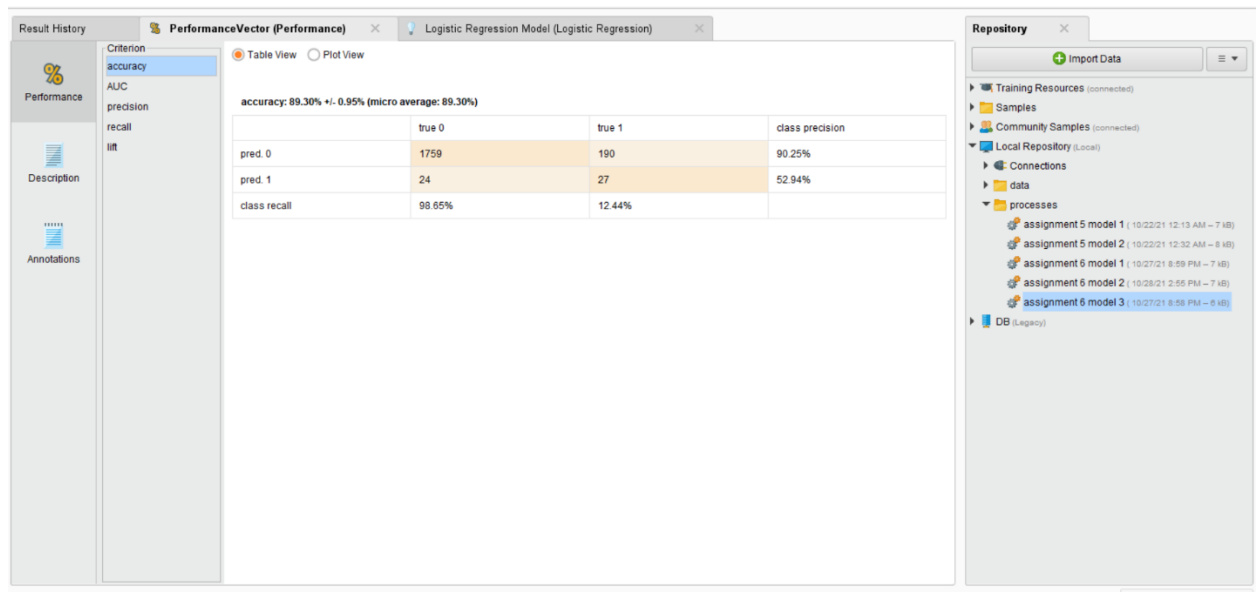
- Repository:** Shows the local repository with various processes and data sources.
- Process:** The workflow includes 'Retrieve Class Ass...', 'Set Role', and 'Cross Validation'.
- Parameters:**
 - Cross Validation:**
 - split on batch attribute: ☐
 - leave one out: ☐
 - number of folds: 10
 - sampling type: automatic
 - use local random seed: ☐
 - enable parallel execution: ☒
- Help:** Provides information about the Cross Validation operator, including its tags and synopsis.

Bottom Screenshot: Logistic Regression Process

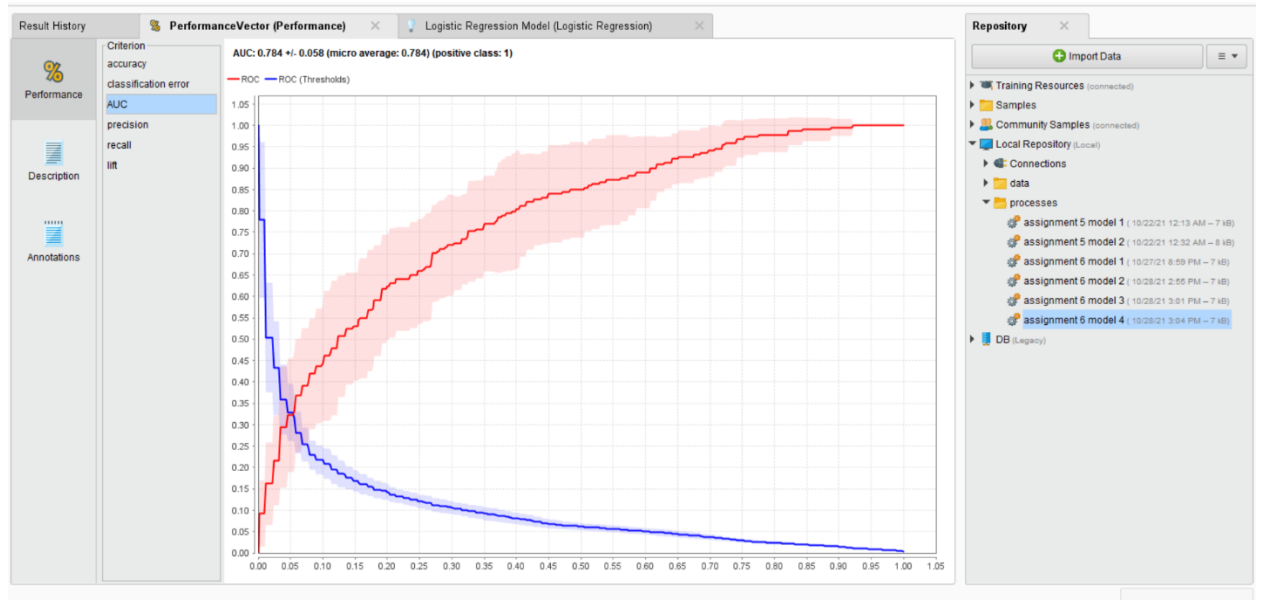
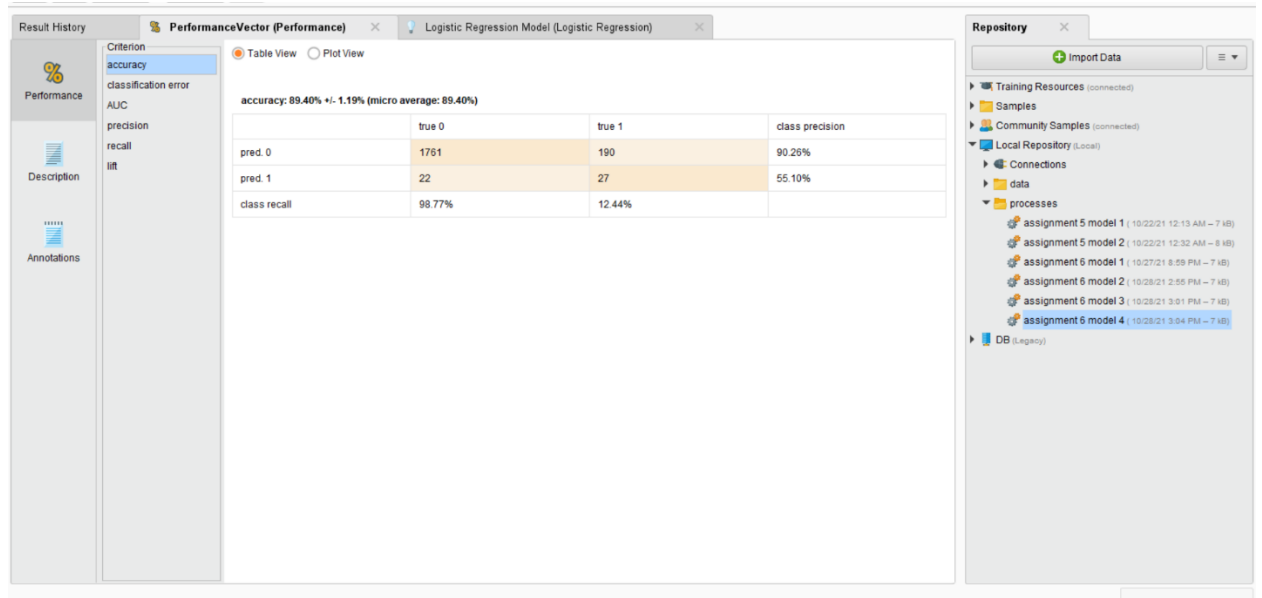
- Repository:** Shows the local repository with various processes and data sources.
- Process:** The workflow includes 'Logistic Regression', 'Apply Model', and 'Performance'.
- Parameters:**
 - Logistic Regression:**
 - reproducible: ☐
 - use regularization: ☐
 - standardize: ☒
 - non-negative coefficients: ☐
 - add intercept: ☒
 - compute p-values: ☒
 - remove collinear columns: ☒
- Help:** Provides information about the Logistic Regression operator, including its tags and synopsis.



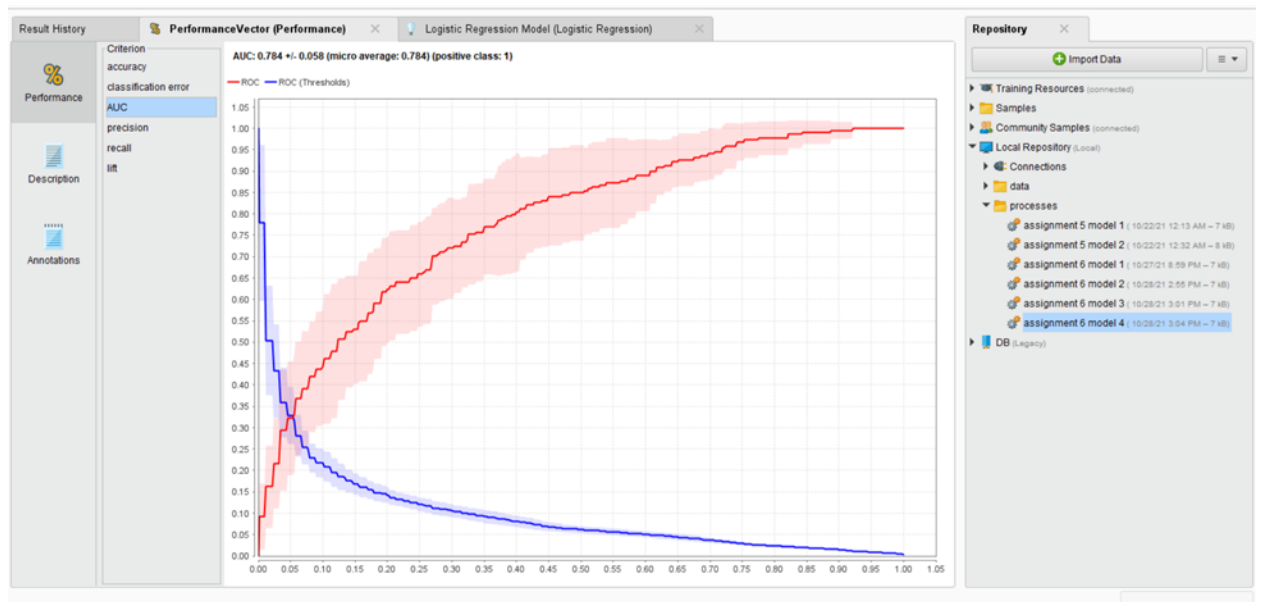
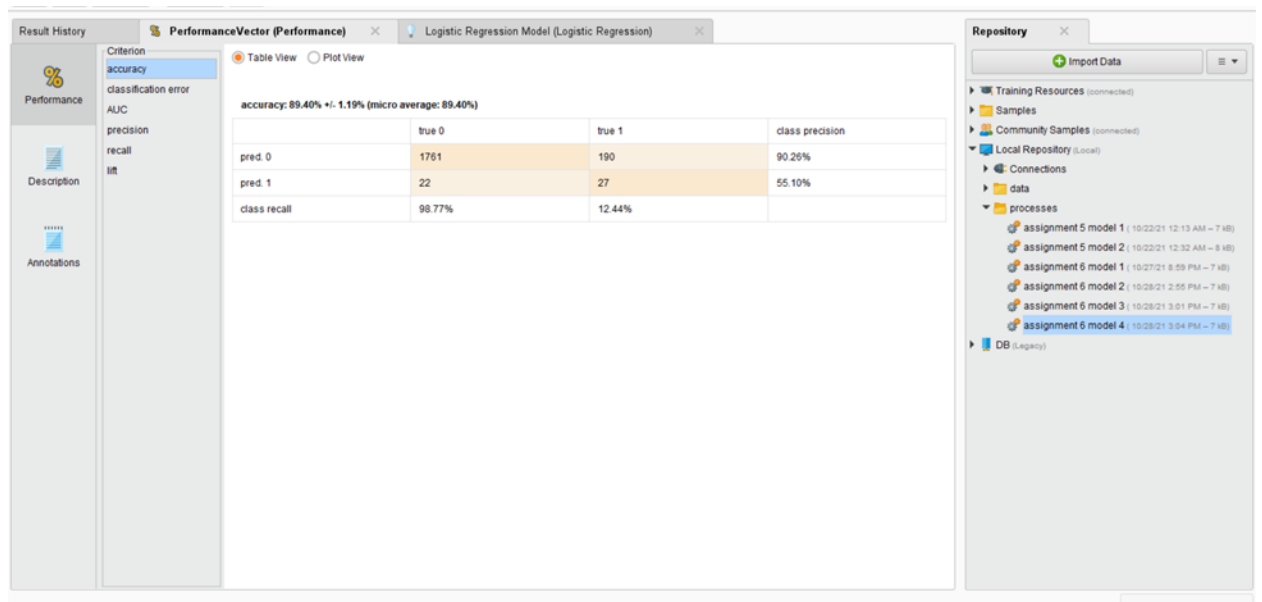
Model 2: model 2 also used 10 fold cross validation and the same default selection options for Logistic Regression except I changed the number of iterations to 20. This model had the same performance as model 1 in terms of accuracy, and AUC score.



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).



- 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

The screenshot shows the RapidMiner Studio interface. The **Repository** panel on the left lists various processes and operators. The **Process** panel in the center displays a workflow: **Retrieve Class Assi...** (input) → **Set Role** → **Cross Validation** → **Apply Model (2)** (output). The **Parameters** panel on the right shows the configuration for the **Apply Model (2)** operator, including application parameters and a **create view** checkbox. The **Help** panel provides information about the **Apply Model** operator, including its tags and a synopsis.

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

The screenshot shows the **Result History** panel in RapidMiner Studio, displaying the prediction results for the **ExampleSet (Apply Model (2))** operator. The results are shown in a table with 18 rows and 11 columns. The **Repository** panel on the right shows the project structure, including training resources and community samples.

Row No.	prediction(f1...	confidence(0)	confidence(1)	Gender	M	R	F	FirstPurch	ChildBks	Y
1	0	0.893	0.107	0	313	10	2	18	0	0
2	0	0.939	0.061	0	373	16	5	36	1	1
3	0	0.956	0.044	0	320	20	12	76	3	1
4	0	0.943	0.057	0	52	14	2	18	2	0
5	0	0.855	0.145	0	261	12	3	28	0	1
6	0	0.847	0.153	1	138	12	5	26	2	0
7	0	0.962	0.038	1	225	22	7	64	2	0
8	0	0.865	0.135	1	158	4	2	10	0	0
9	0	0.504	0.496	0	253	8	8	30	2	1
10	0	0.992	0.008	1	256	28	2	34	1	0
11	0	0.928	0.072	1	48	8	1	8	0	0
12	0	0.560	0.440	0	222	14	8	36	0	3
13	0	0.969	0.031	1	279	14	1	14	1	0
14	0	0.980	0.020	0	160	34	2	38	0	0
15	0	0.980	0.020	1	263	24	2	28	0	0
16	0	0.731	0.269	1	162	6	10	64	2	1
17	0	0.919	0.081	1	133	16	2	18	0	0
18	0	0.707	0.293	1	276	12	12	60	3	0

ExampleSet (47 examples, 3 special attributes, 15 regular attributes)

Result History | ExampleSet (Apply Model (2)) | PerformanceVector (Performance)

Open in Turbo Prep Auto Model Filter (47 / 47 examples): all

Row No.	prediction(F...	confidence(0)	confidence(1)	Gender	M	R	F	FirstPurch	ChildBks	Y
19	0	0.960	0.040	1	200	14	8	46	1	2
20	0	0.985	0.015	1	186	22	2	28	0	0
21	0	0.687	0.313	1	192	14	9	62	1	1
22	0	0.691	0.309	0	125	2	2	10	0	0
23	0	0.785	0.235	0	145	2	1	2	0	0
24	0	0.956	0.044	1	106	20	5	46	0	1
25	0	0.857	0.143	1	277	4	2	12	0	0
26	0	0.724	0.276	1	367	10	12	60	2	2
27	0	0.909	0.091	1	142	10	2	16	0	0
28	0	0.974	0.026	1	112	16	1	16	1	0
29	0	0.883	0.117	0	328	12	2	18	1	0
30	1	0.094	0.906	1	361	6	12	44	1	0
31	0	0.695	0.305	0	168	2	9	20	0	0
32	0	0.889	0.111	1	114	2	1	2	0	0
33	0	0.928	0.072	1	427	34	9	62	2	0
34	0	0.722	0.278	0	171	12	2	16	0	0
35	0	0.977	0.023	1	165	16	1	16	0	1
36	0	0.726	0.274	0	339	12	7	48	1	0

ExampleSet (47 examples, 3 special attributes, 15 regular attributes)

Repository

- Import Data
- Training Resources (connected)
- Samples
- Community Samples (connected)
- Local Repository (Local)
 - Connections
 - data
 - processes
 - assignment 5 model 1 (10/22/21 12:13 AM - 7 MB)
 - assignment 5 model 2 (10/22/21 12:32 AM - 8 MB)
 - assignment 6 model 1 (10/27/21 8:59 PM - 7 MB)
 - assignment 6 model 2 (10/28/21 2:55 PM - 7 MB)
 - assignment 6 model 3 (10/28/21 3:01 PM - 7 MB)
 - assignment 6 model 4 (10/28/21 3:24 PM - 8 MB)
- DB (Legacy)

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.

Result History

ExampleSet (Apply Model (2))

PerformanceVector (Performance)

Performance

Description

Annotations

Criterion

accuracy

classification error

AUC

precision

recall

lift

Table View Plot View

lift: 507.85% (positive class: 1)

	true 0	true 1	class precision
pred. 0	1761	190	90.26%
pred. 1	22	27	55.10%
class recall	98.77%	12.44%	

Repository

Import Data

Training Resources (connected)

Samples

Community Samples (connected)

Local Repository (Local)

Connections

data

processes

assignment 5 model 1 (10/22/21 12:13 AM - 7 kB)

assignment 5 model 2 (10/22/21 12:32 AM - 8 kB)

assignment 6 model 1 (10/26/21 3:31 PM - 7 kB)

assignment 6 model 2 (10/26/21 2:55 PM - 7 kB)

assignment 6 model 3 (10/26/21 3:01 PM - 7 kB)

assignment 6 model 4 (10/26/21 3:24 PM - 8 kB)

DB (Legacy)

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