Assignment # 2: Decisions Trees and Classification Targeted Marketing Campaign
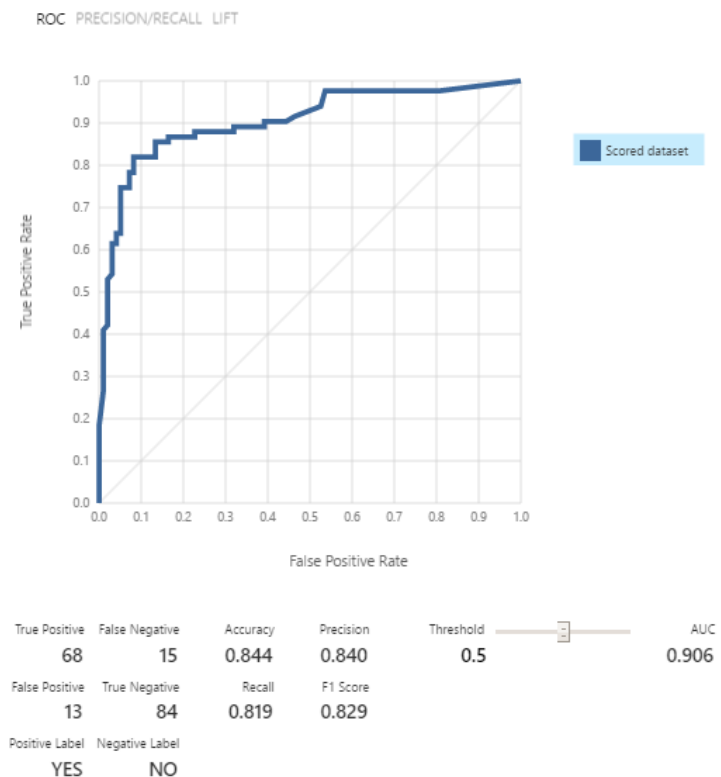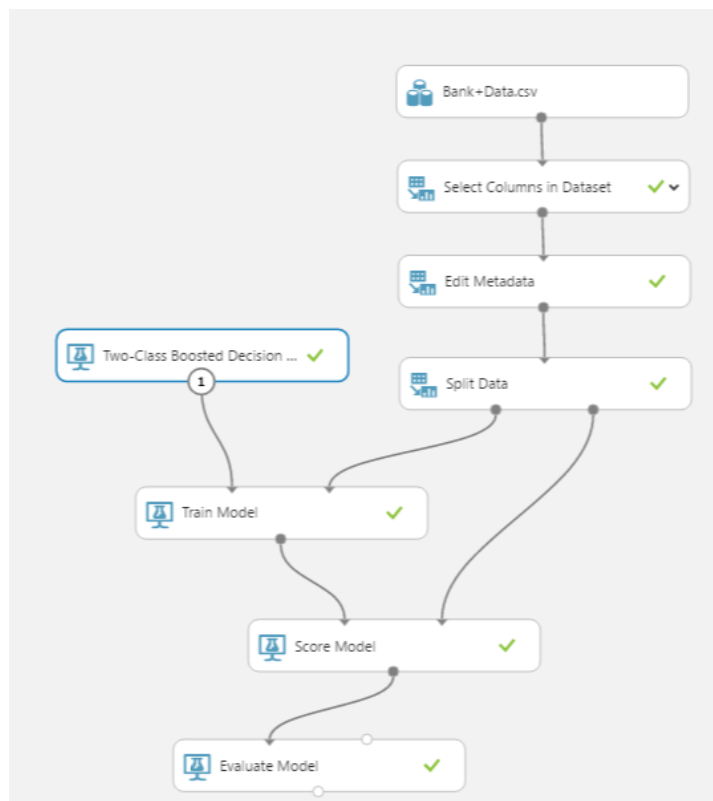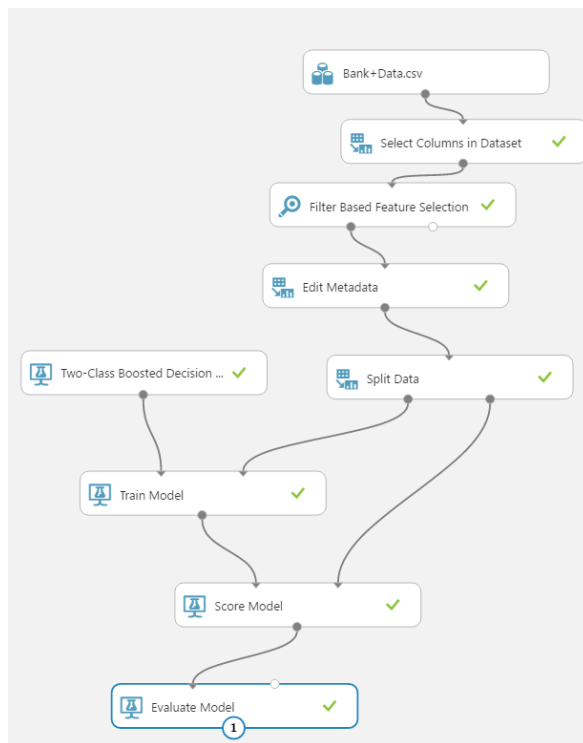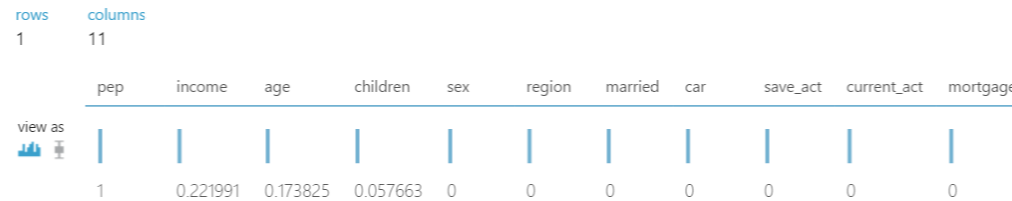
In this problem we will use historical data from past customer responses to build a classification model. We will apply the trained model to a new set of prospects to whom we may want extend an offer for a PEP. Rather than doing a mass marketing campaign to all new prospects, we would like to target those that are likely to respond positively to our offer (according to our classification model).

- First, an Azure Machine Learning model is generated by using all the given attributes to test the accuracy:
- Two-Class Boosted Decision Tree classification model is selected, default parameters are used.
- 70 % of the data is used to train the model and the remaining is to test.
- The model provided 84.4 % accuracy to predicted the customers who would accept or deny the offer by all the attributes.

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
68	15	0.844	0.840	0.5	0.906

False Positive	True Negative	Recall	F1 Score		
13	84	0.819	0.829		

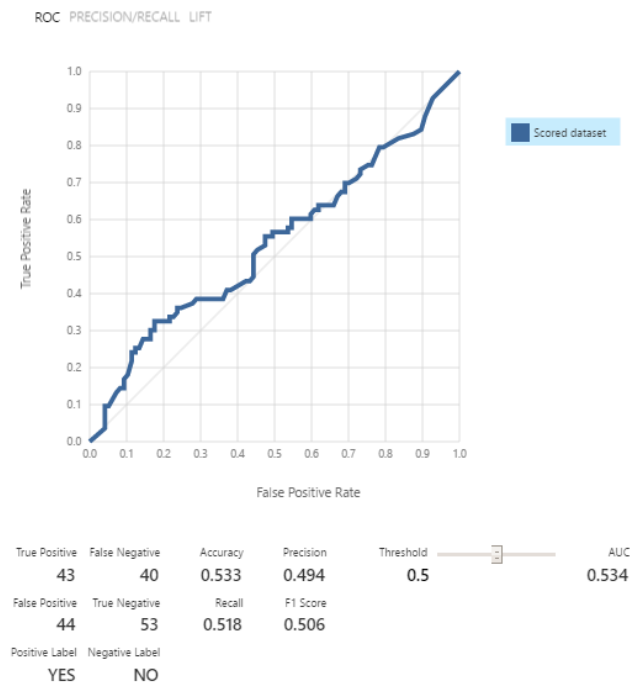| Positive Label	Negative Label
YES	NO

- In order to find out which features can be eliminated, "Filter Based Feature Selection" is used and results are shown below.
- The default "Pearson Correlation" is used.
- Only "income", "age" and "children" features showed higher impact compared to other features
- Then model is re-run by using these three features only, however, the accuracy of the model dropped to 53.3 %!

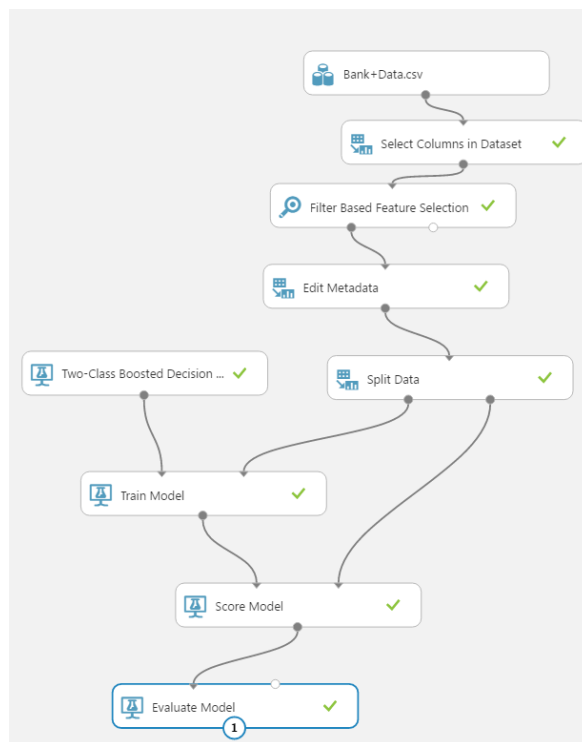Assignment -2 ❯ Filter Based Feature Selection ❯ Features

rows	columns
1	11

	pep	income	age	children	sex	region	married	car	save_act	current_act	mortgage
view as	1	0.221991	0.173825	0.057663	0	0	0	0	0	0	0

Assignment -2 ❯ Evaluate Model ❯ Evaluation results

ROC PRECISION/RECALL LIFT

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
43	40	0.533	0.494	0.5	0.534

False Positive	True Negative	Recall	F1 Score
44	53	0.518	0.506

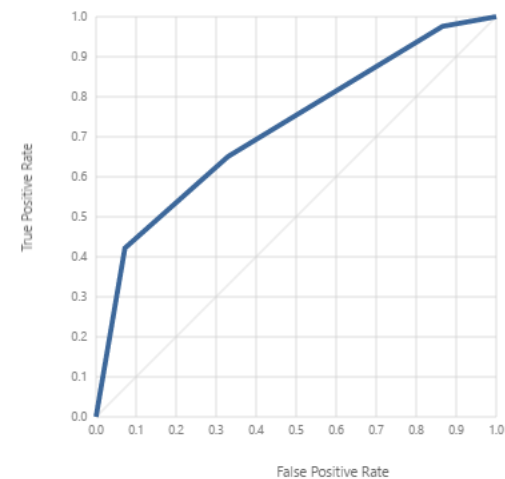| Positive Label	Negative Label
YES	NO

2

- When the "Mutual Information" option is used a different series of features are selected.
- When the model is re-run by using these features only, the accuracy of the model increased to 69.4 % compared to "Pearson Correlation" option. However, the accuracy is still lower than using all the features case.

Assignment -2 › Filter Based Feature Selection › Features

rows   columns
1   11

	pep	children	income	married	age	region	save_act	sex	current_act	mortgage	car
view as											
	1	0.084564	0.037109	0.017765	0.016945	0.003083	0.002619	0.001083	0.000302	0.000284	0.000177

Bank+Data.csv
Select Columns in Dataset ✓
Filter Based Feature Selection ✓
Edit Metadata ✓
Two-Class Boosted Decision ... ✓   Split Data ✓
Train Model ✓
Score Model ✓
Evaluate Model ✓

ROC PRECISION/RECALL LIFT

Scored dataset

True Positive Rate

False Positive Rate

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
35	48	0.694	0.833	0.5	0.722

False Positive	True Negative	Recall	F1 Score		
7	90	0.422	0.560		

Positive Label	Negative Label
YES	NO

3

- Below, comparison of the two attributes against to the PEP offer are shown; "sex" and "car". It can be seen that neither being male or female, nor owning a car has high impact on PEP offer acceptance
- Therefore, these two attributes are removed from the data set.
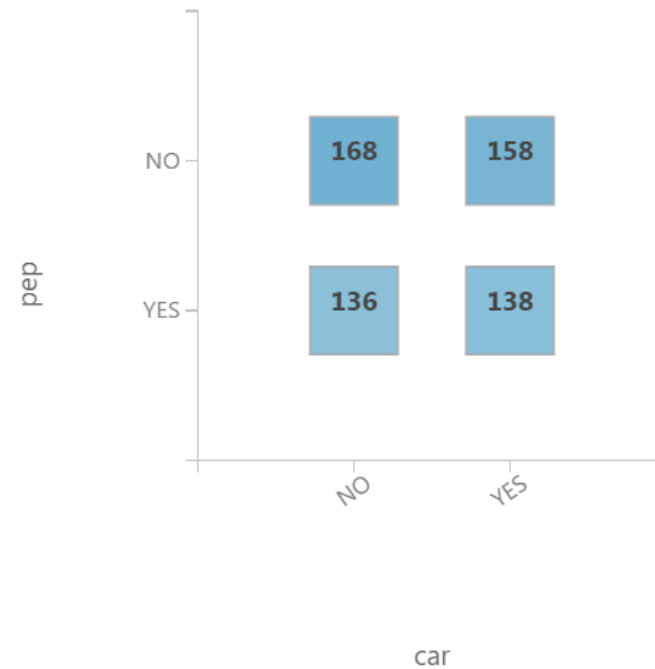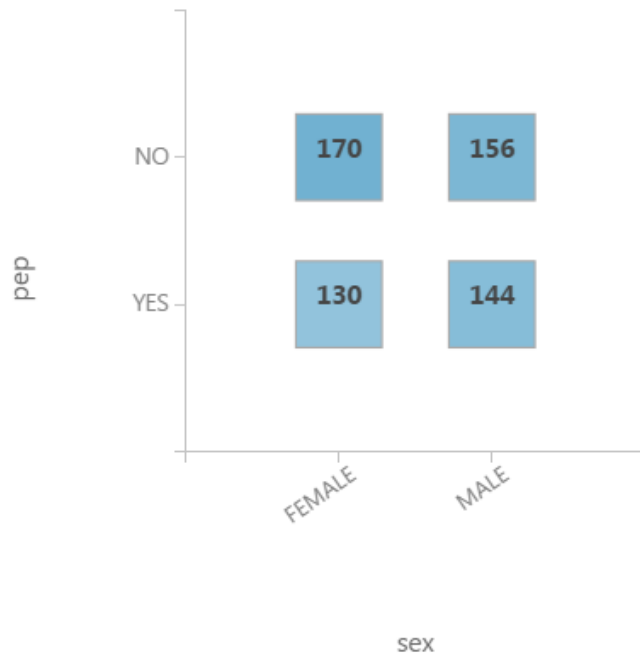
- Unlike previous "Filter Base Feature Selection" cases, removing the "sex" and "car" attributes improved the model accuracy
- As shown below, the total number of false positives were reduced and the model accuracy improved to 88.9 % in prediction of PEP offer acceptance.

Two-Class Boosted Decision...

Create trainer mode
Single Parameter

Maximum number of l...
20

Minimum number of s...
10

Learning rate
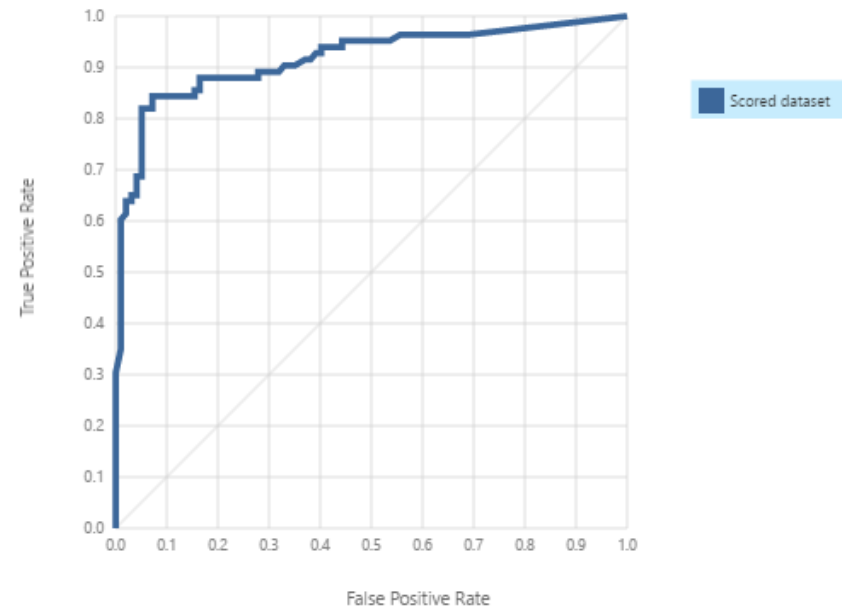0.2

Number of trees const...
100

Random number seed

☑ Allow unknown ca...

START TIME   7/20/20...
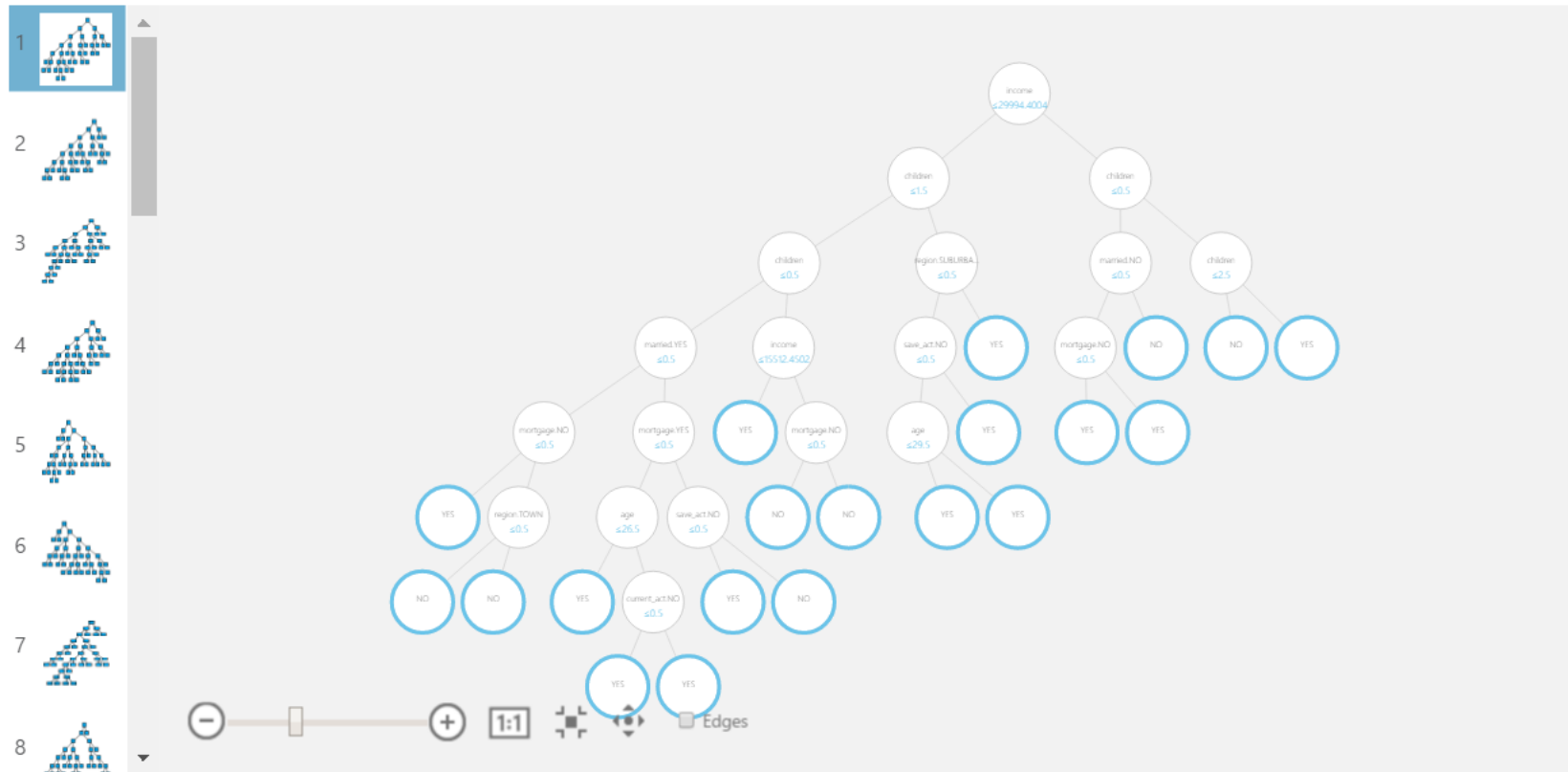END TIME   7/20/20...

ROC   PRECISION/RECALL   LIFT

Scored dataset

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
70	13	0.889	0.909	0.5	0.918

False Positive	True Negative	Recall	F1 Score		
7	90	0.843	0.875		

Positive Label	Negative Label
YES	NO

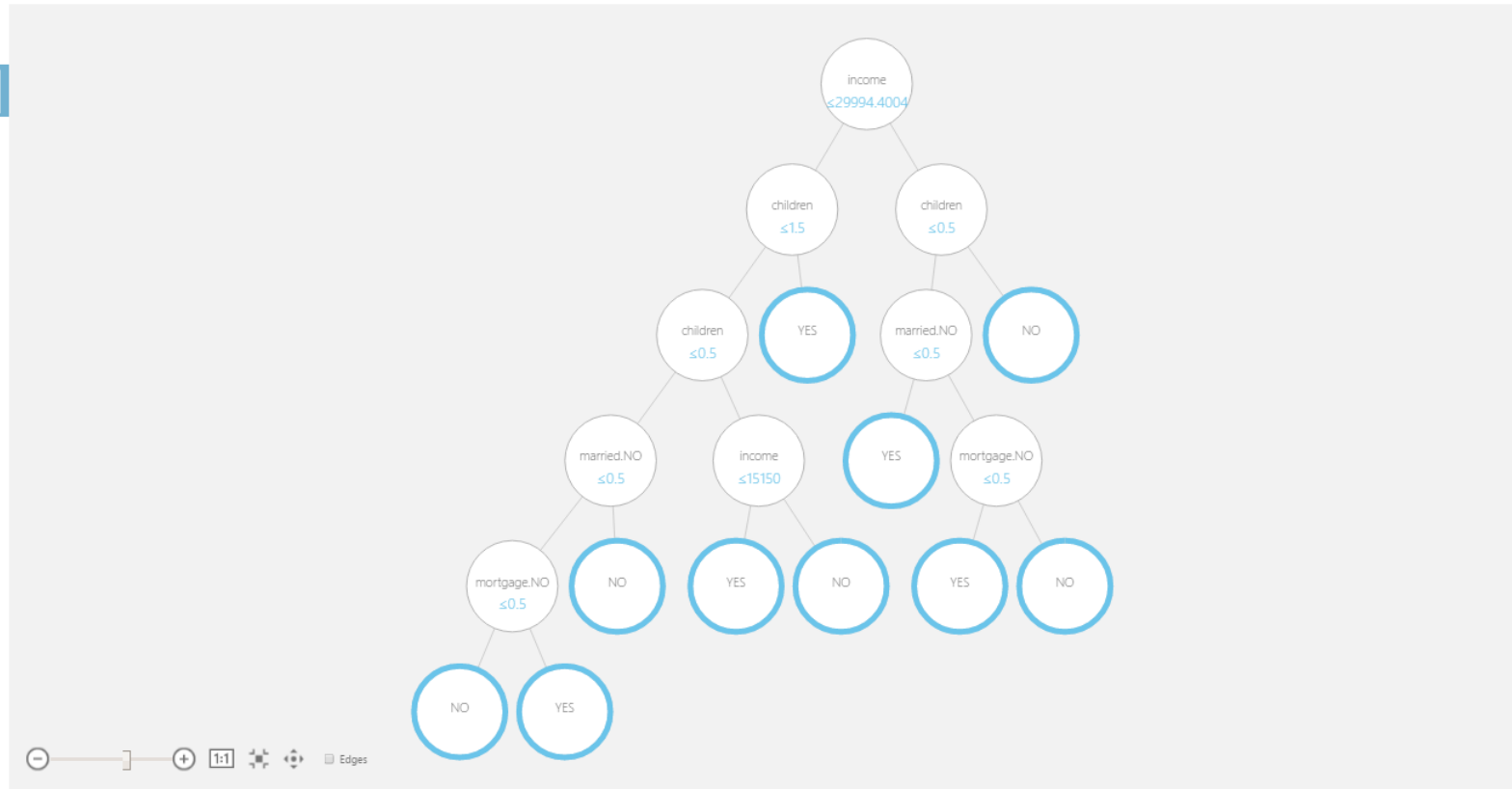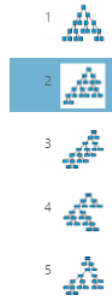- 100 different trees are constructed, an example is shown below :

- Finally the Two-Class Boosted Decision Tree model "maximum number of leves" is reduced from 20 to 10 and "minimum number of training instances" is reduced from 10 to 5 and the total number of trees constructed is limited to 5 so that each tree can be prestened below.
- Based on the structure of the trees, "income" and "children" features are found as the most important fetaures of this model.
- As shown below, the total number of false positives was increased and the model accuracy dropped to 87.8 % .
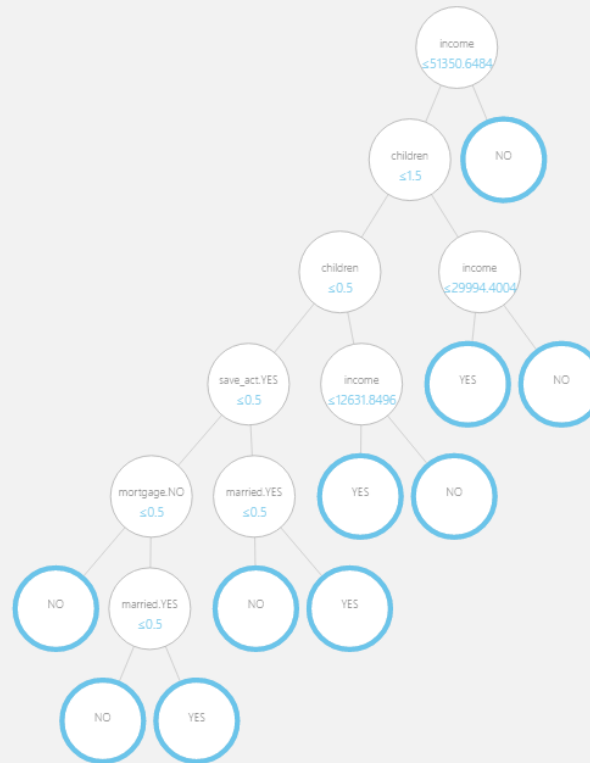
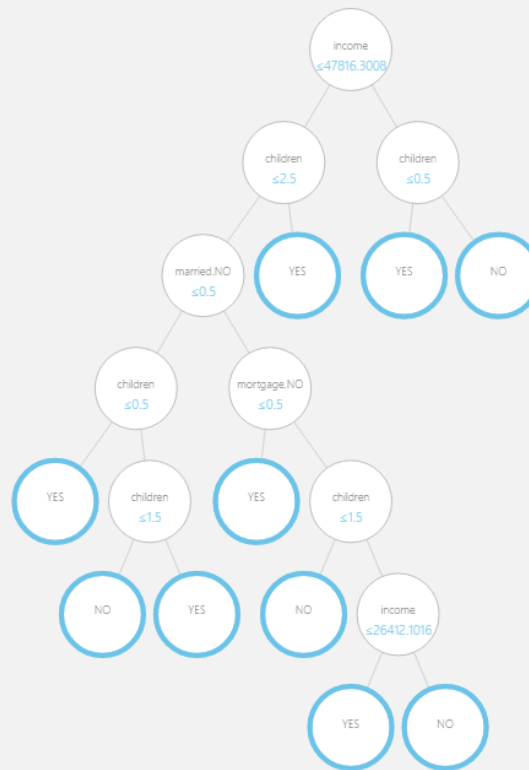Assignment -2 › Evaluate Model › Evaluation results

ROC PRECISION/RECALL LIFT

Properties   Project

◢ Two-Class Boosted Decision ...

Create trainer mode

| Single Parameter ▼ |

Maximum number of le... ≡

| 10 |

Minimum number of sa... ≡

| 5 |

Learning rate ≡

| 0.2 |

Number of trees constr... ≡

| 5 |

Random number seed ≡

| |

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
73	10	0.878	0.859	0.5	0.935

False Positive	True Negative	Recall	F1 Score
12	85	0.880	0.869

Positive Label	Negative Label
YES	NO