

Index

<i>Creating a model to predict Employee's Attrition (Binary Classification task) using Data Science Platforms such as Azure ML Studio, RapidMiner, Knime Analytics)</i>	1
1. Azure Machine Learning Studio	2
Two-Class Logistic Regression.....	3
Two-Class Boosted Decision Tree	4
Two-Class Decision Forest	5
Two-Class Decision Jungle	6
The Two-Class Decision Forest Model has been deployed as a web-service in Azure ML Studio (classic).	6
The experiment shared in the Gallery:.....	8
2. RapidMiner Studio	9
Data Preparation	9
Choosing the models.....	12
Evaluating the models.....	12
The best model in details	13
Most significant fields for the model	14
Optimizing the model parameters	15
The structure of predictive model (Gradient Boosted Trees)	18
3. Knime Analytics Platform	19
Converting the target filed to categorical type	19
SMOTE module to fix the class imbalance	20
Comparing the models.....	22
Logistic Regression	22
Decision Tree	24
Gradient Boosted Tree.....	26
Random Forest	28
The best model.....	30

Creating a model to predict Employee's Attrition (Binary Classification task) using Data Science Platforms such as Azure ML Studio, RapidMiner, Knime Analytics)

The task is to create predictive models using different Data Science platforms and compare the results of prediction choosing the best model for this particular task. I used the dataset from Kaggle.com <https://www.kaggle.com/tejashvi14/employee-future-prediction>

The task is to predict if an employee will leave the company or will keep working in the same company in the next 2 years, the model will be created based on the dataset from Kaggle. The task is a Binary Classification task.

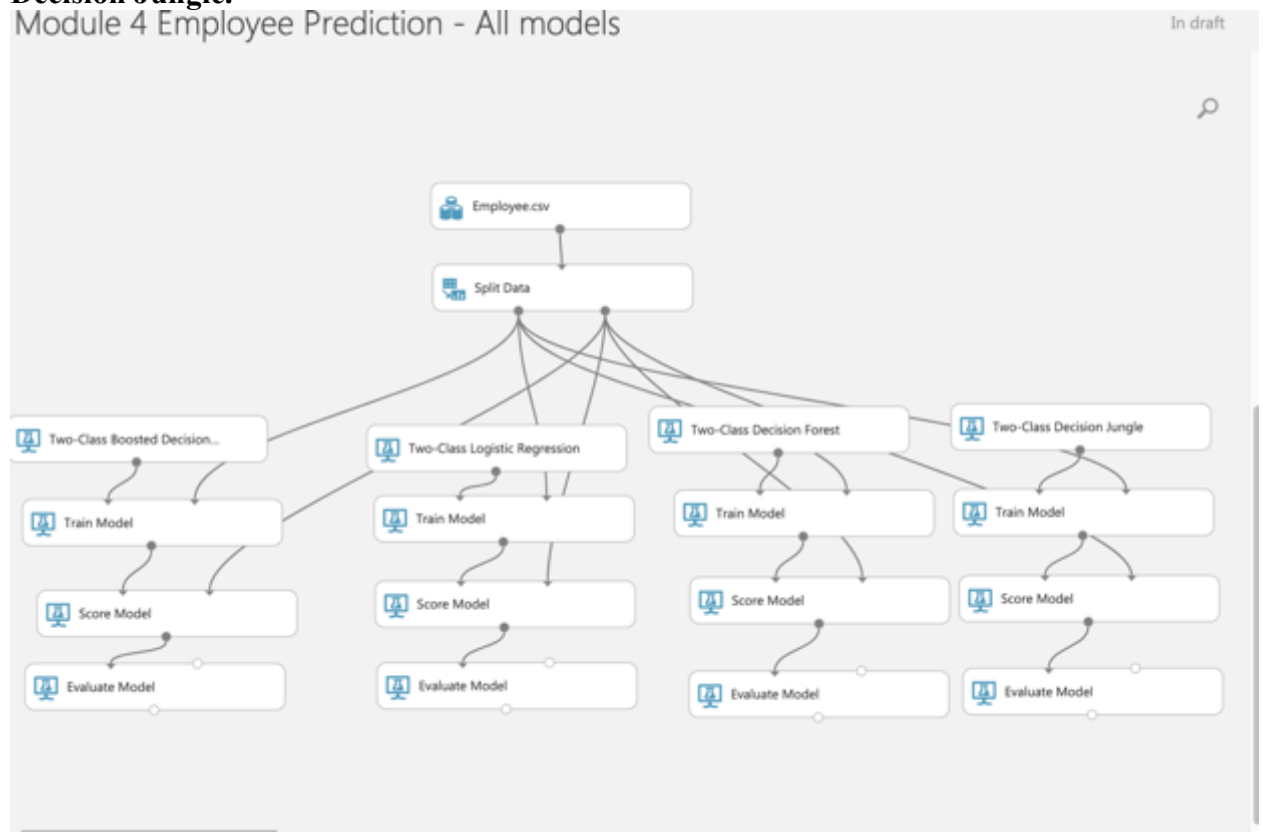
We have the following data for the model creation:

1. Education
2. JoiningYear
3. City
4. PaymentTier
5. Age
6. Gender
7. EverBenched
8. ExperienceInCurrentDomain
9. LeaveOrNot – target field

1. Azure Machine Learning Studio

We will work with Azure ML Studio first.

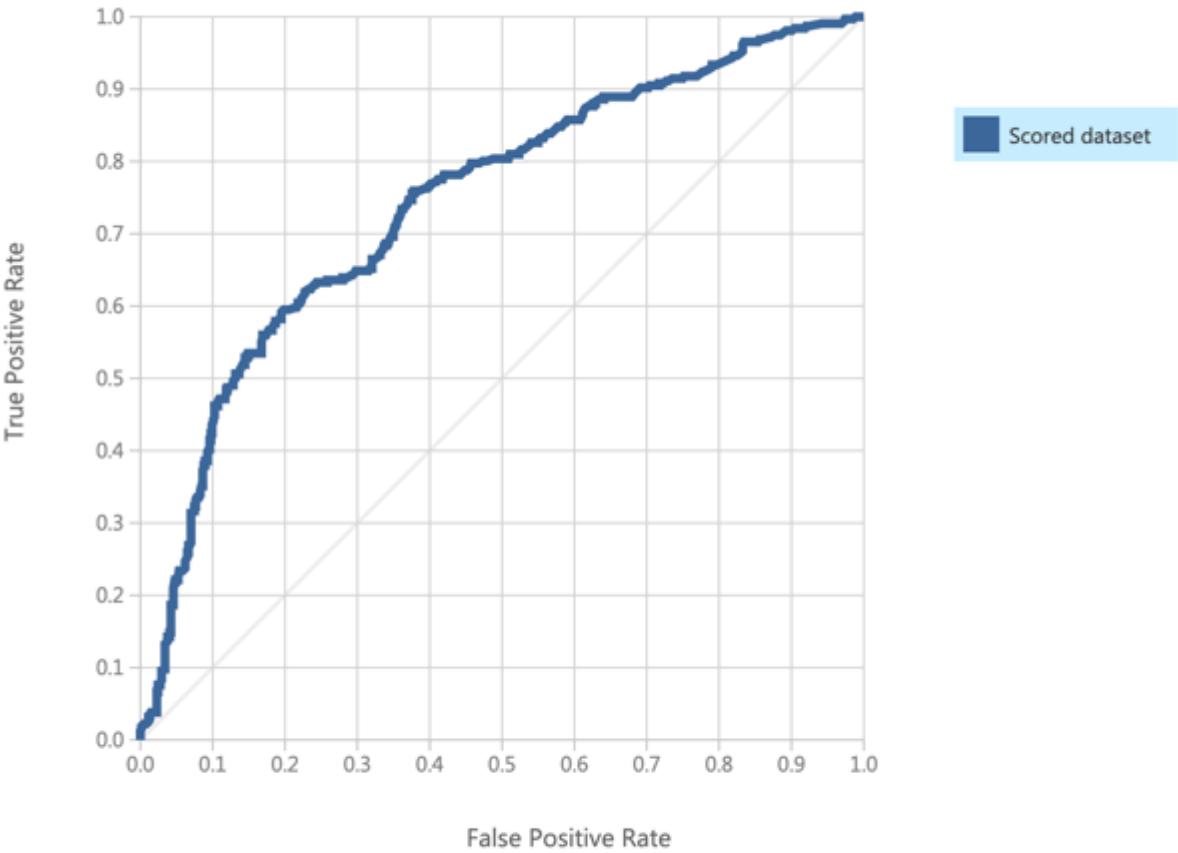
Using Azure ML Studio (classic) we create different models, train the models, and evaluate them using the scoring metrics. The following models will be created: **Two-Class Logistic Regression, Two-Class Boosted Decision Tree, Two-Class Decision Forest, Two-Class Decision Jungle.**



Let's compare the results we received after creating each model.

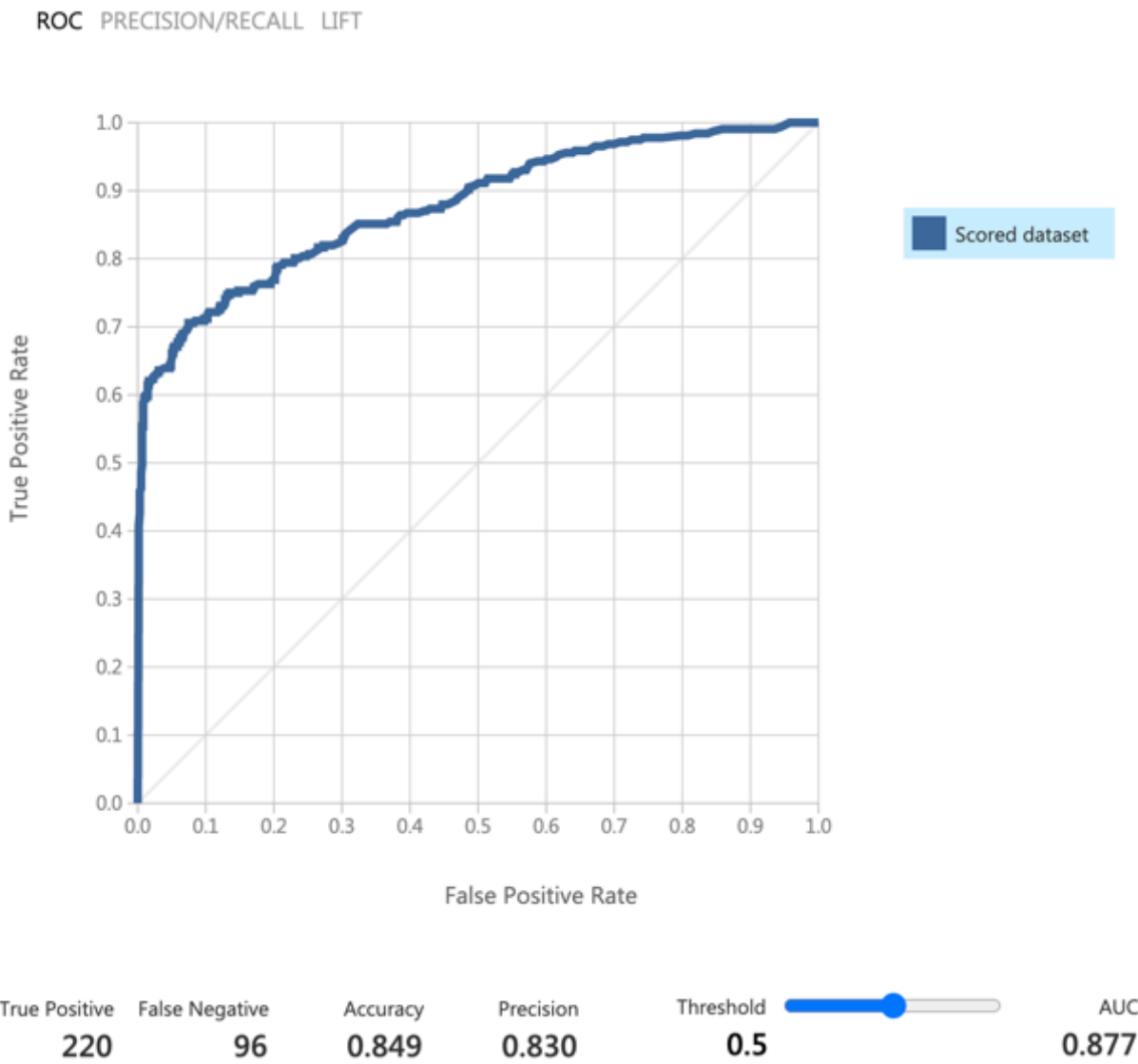
Two-Class Logistic Regression

ROC PRECISION/RECALL LIFT

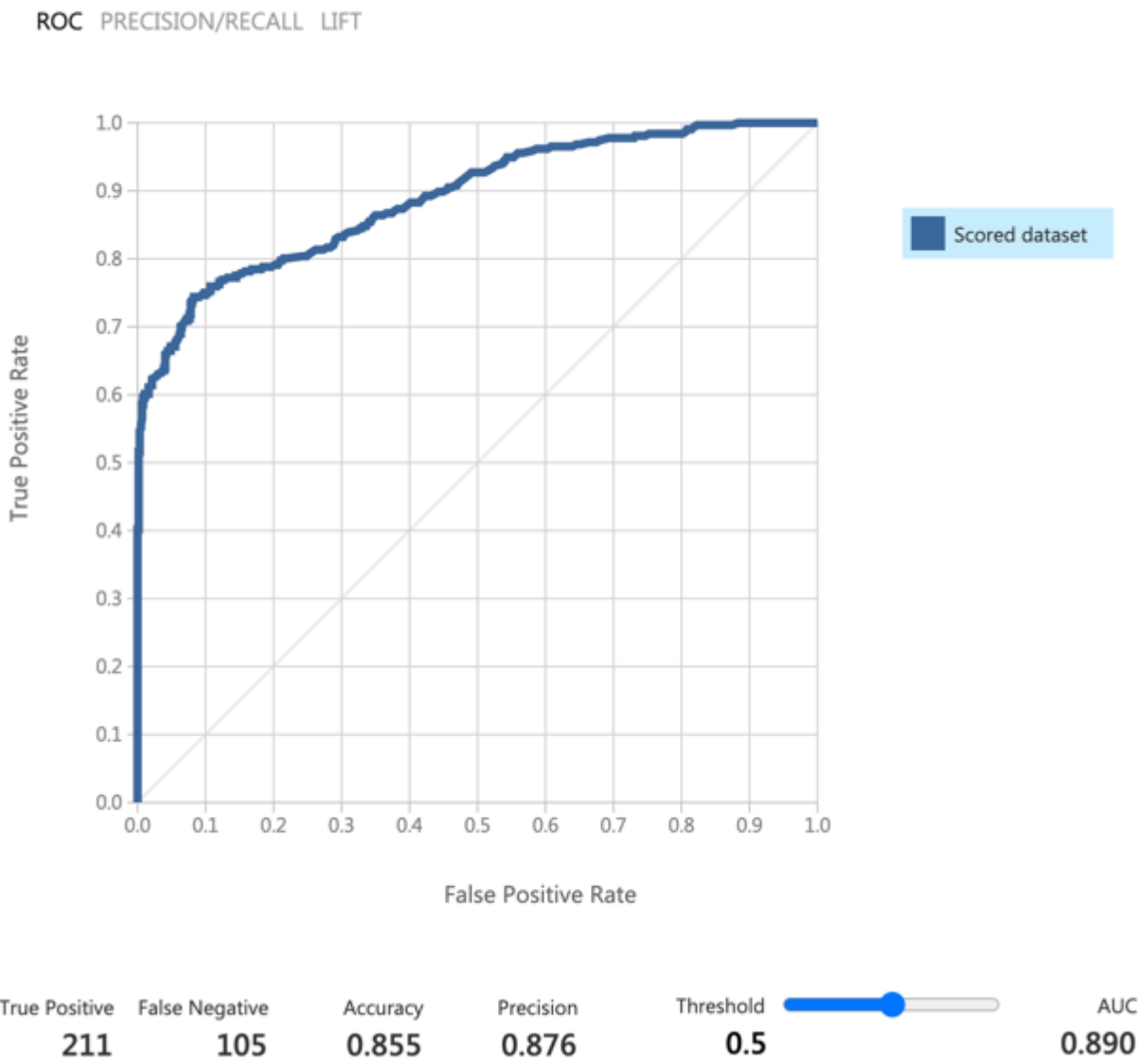


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
149	167	0.748	0.687	0.5	0.743

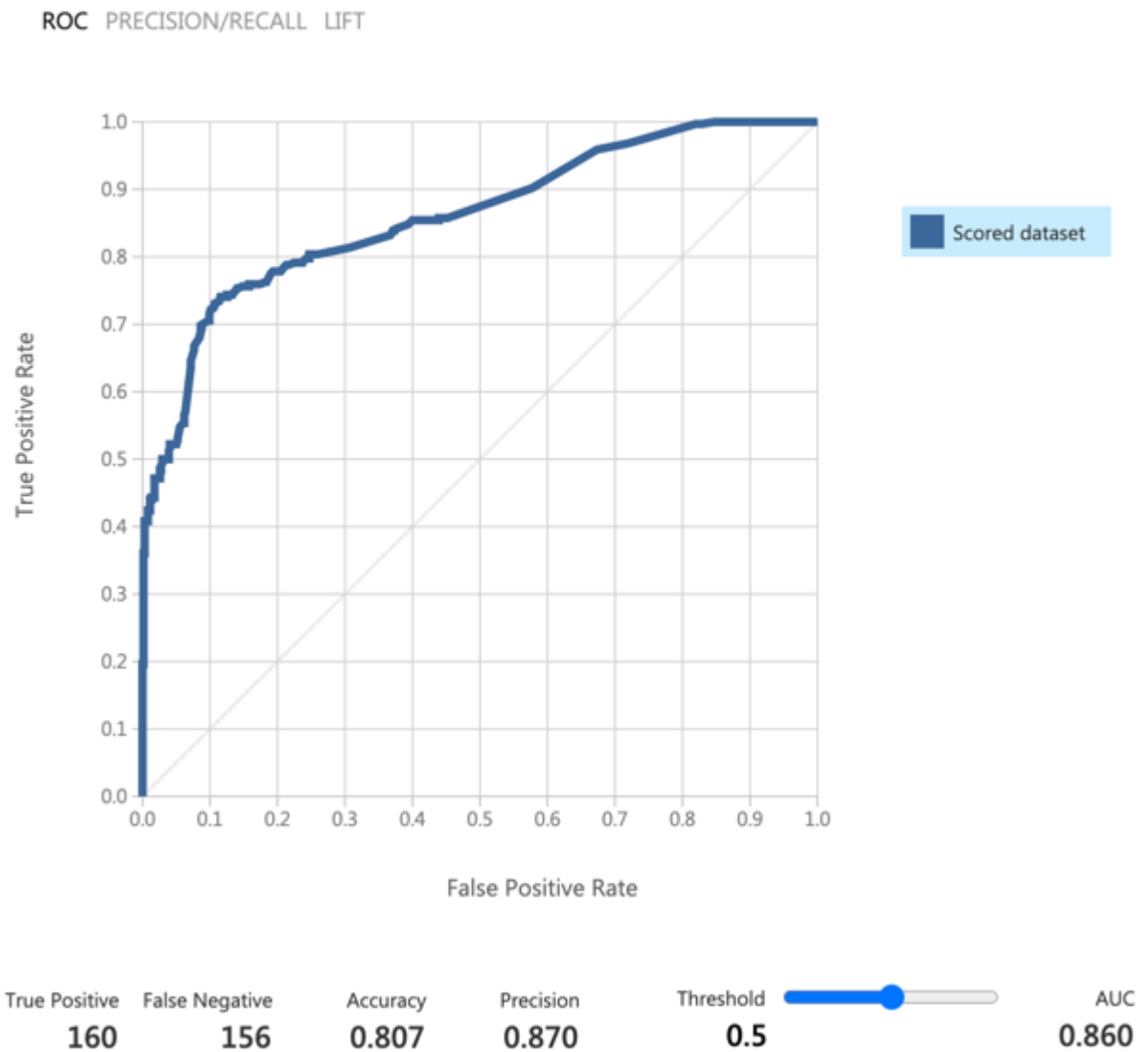
Two-Class Boosted Decision Tree



Two-Class Decision Forest



Two-Class Decision Jungle

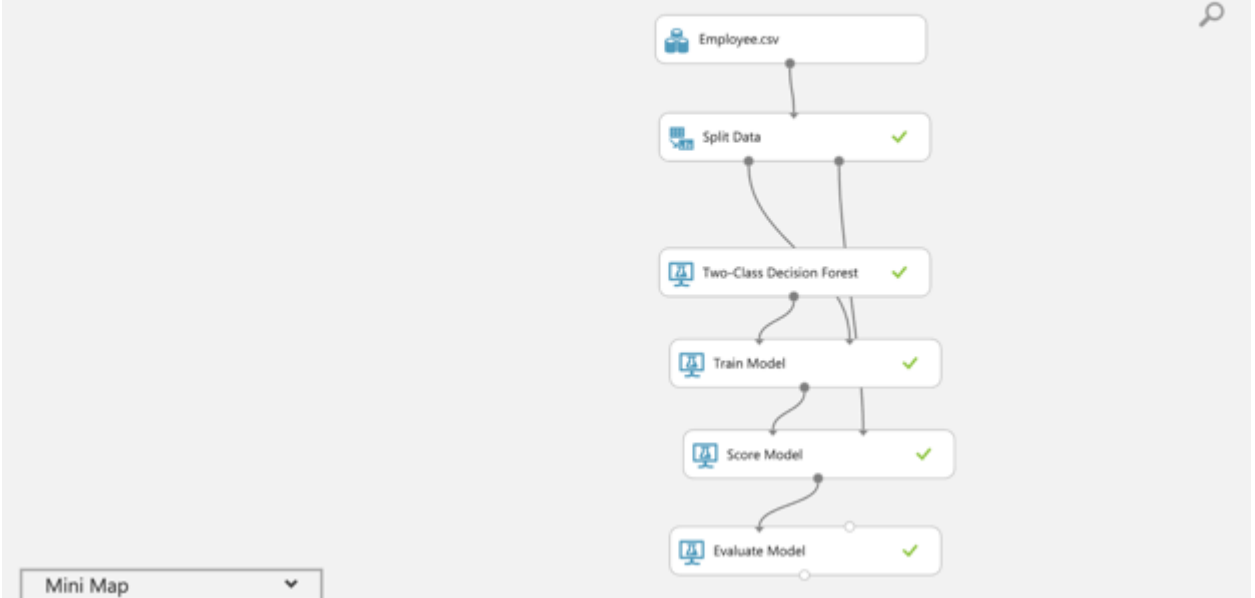


The best model is **Two-Class Decision Forest** with **AUC of 0.89**.

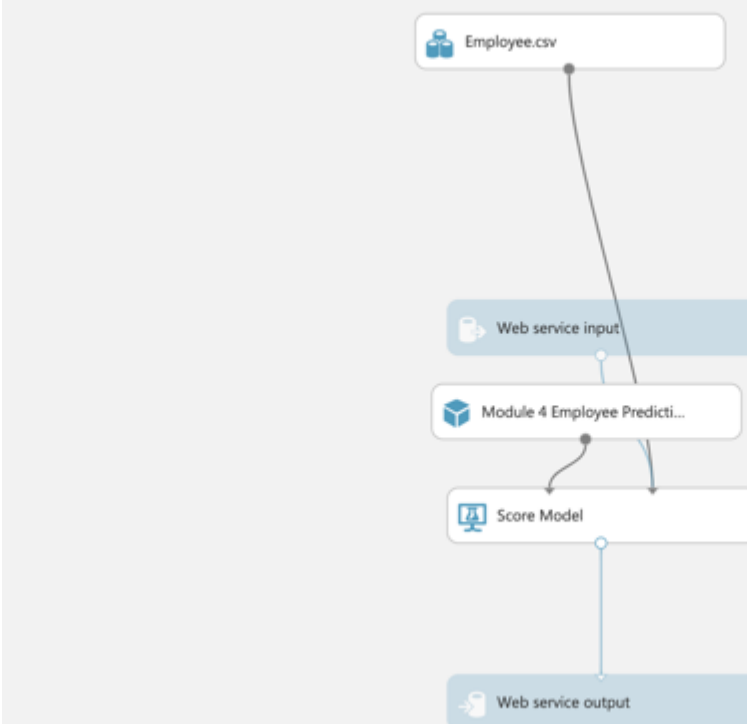
The Two-Class Decision Forest Model has been deployed as a web-service in Azure ML Studio (classic).

Module 4 Employee Prediction - Best Model

Finished running



Module 4 Employee Prediction - Best Model [Predictive Exp.]



Microsoft Machine Learning Studio (classic)

module 4 employee prediction - best model [predictive exp.]

DASHBOARD CONFIGURATION

General **New Web Services Experience** *preview*

Published experiment
View snapshot View latest

Description
No description provided for this web service.

API key
XCHg18CV850H64k38dpM6G6woT7gdX05v9pHOTC8EwMubtoT13HPw3b/EaSqCuG2FV5nRfdrndEQBA==

Default Endpoint

API HELP PAGE	TEST	APPS	LAST UPDATED
REQUEST/RESPONSE	Test <i>Test preview</i>	Excel 2013 or later Excel 2010 or earlier workbook	3/14/2022 2:42:22 PM
BATCH DISCUSSION	<i>Test preview</i>	Excel 2013 or later workbook	3/14/2022 2:42:22 PM

input1

Education

Masters

JoiningYear

2016

City

Pune

PaymentTier

2

Age

24

Gender

Male

EverBenched

No

ExperienceInCurrentDomain

3

LeaveOnNot

1

Test Request-Response

output1

Education

Masters

JoiningYear

2016

City

Pune

PaymentTier

2

Age

24

Gender

Male

EverBenched

No

ExperienceInCurrentDomain

3

LeaveOnNot

1

Scored Labels

0

Scored Probabilities

0.485070385946345

The experiment shared in the Gallery

<https://gallery.cortanaintelligence.com/Experiment/Module-4-Employee-Prediction-Best-Model-Predictive-Exp>

EXPERIMENT

Module 4 Employee Prediction - Best Model [Predictive Exp.]


 Eulypena Eucarpoua Hooostaaa • January 17, 2022

 Be the first to like.

Summary

Description



Open in Studio (classic)

+ Add to Collection

1 view

1 download

2. RapidMiner Studio

Data Preparation

The dataset for training the models and evaluating them is imported into RapidMiner Studio. We use Turbo Prep and Auto Model technologies for data preparation and building the models.

<new process> - RapidMiner Studio Educational 9.10.001 @ MacBook-Pro-Ekaterina.local

Views: Design Results Turbo Prep Auto Model Deployments

Result History ExampleSet (//Local Repository/data/Employee)

Open in Turbo Prep Auto Model Filter (4,653 / 4,653 examples): all

Row No.	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	Experiencel...	LeaveOrNot
1	Bachelors	2017	Bangalore	3	34	Male	No	0	0
2	Bachelors	2013	Pune	1	28	Female	No	3	1
3	Bachelors	2014	New Delhi	3	38	Female	No	2	0
4	Masters	2016	Bangalore	3	27	Male	No	5	1
5	Masters	2017	Pune	3	24	Male	Yes	2	1
6	Bachelors	2016	Bangalore	3	22	Male	No	0	0
7	Bachelors	2015	New Delhi	3	38	Male	No	0	0
8	Bachelors	2016	Bangalore	3	34	Female	No	2	1
9	Bachelors	2016	Pune	3	23	Male	No	1	0
10	Masters	2017	New Delhi	2	37	Male	No	2	0
11	Masters	2012	Bangalore	3	27	Male	No	5	1
12	Bachelors	2016	Pune	3	34	Male	No	3	0
13	Bachelors	2018	Pune	3	32	Male	Yes	5	1
14	Bachelors	2016	Bangalore	3	39	Male	No	2	0
15	Bachelors	2012	Bangalore	3	37	Male	No	4	0
16	Bachelors	2017	Bangalore	1	29	Male	No	3	0
17	Bachelors	2014	Bangalore	3	34	Female	No	2	0
18	Bachelors	2014	Pune	3	34	Male	No	4	0
19	Bachelors	2015	Pune	2	30	Female	No	0	1
20	Bachelors	2016	New Delhi	2	22	Female	No	0	1
21	Bachelors	2012	Bangalore	3	37	Male	No	0	0
22	Masters	2017	New Delhi	2	28	Male	No	4	0
23	Bachelors	2017	New Delhi	2	36	Male	No	3	0

ExampleSet (4,653 examples, 0 special attributes, 9 regular attributes)

<new process> – RapidMiner Studio Educational 9.10.001 @ MacBook-Pro-Ekaterina.local

Views: Design Results Turbo Prep Auto Model Deployments

Result History ExampleSet (//Local Repository/data/Employee)

Filter (9 / 9 attributes): Search for Attribute.

Name	Type	Missing	Statistics	Values
Education	Nominal	0	Least PHD (179) Most Bachelors (3601)	Bachelors (3601), Masters (873), ...[1 more]
JoiningYear	Integer	0	Min 2012 Max 2018	Average 2015.063
City	Nominal	0	Least New Delhi (1157) Most Bangalore (2228)	Bangalore (2228), Pune (1268), ...[1 more]
PaymentTier	Integer	0	Min 1 Max 3	Average 2.698
Age	Integer	0	Min 22 Max 41	Average 29.393
Gender	Nominal	0	Least Female (1875) Most Male (2778)	Male (2778), Female (1875)
EverBenched	Nominal	0	Least Yes (478) Most No (4175)	No (4175), Yes (478)
ExperienceInCurrentDomain	Integer	0	Min 0 Max 7	Average 2.906
LeaveOrNot	Integer	0	Min 0 Max 1	Average 0.344

Showing attributes 1 – 9 Examples: 4,653 Special Attributes: 0 Regular Attributes: 9

<new process> – RapidMiner Studio Educational 9.10.001 @ MacBook-Pro-Ekaterina.local

Views: Design Results Turbo Prep Auto Model Deployments Find data, operators, etc. All Studio

Turbo Prep Transform

1 column selected

RENAME CHANGE TYPE REMOVE COPY FILTER RANGE SAMPLE SORT REPLACE SPLIT

Employee

Select columns to transform (hold Shift for selecting a range of columns; Ctrl for ide-selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the ...

COMMIT TRANSFORMATION CANCEL

Commit all changes to this data set - all transformations will be computed on the full data set and you will go back to the main view.

UNDO SHOW HISTORY

Education Category	JoiningYear Number	City Category	PaymentTier Number	Age Number	Gender Category	EverBenched Category	ExperienceInCurr... Number	LeaveOrNot Category
Bachelors	2017	Bangalore	3	34	Male	No	0	0
Bachelors	2013	Pune	1	28	Female	No	3	1
Bachelors	2014	New Delhi	3	38	Female	No	2	0
Masters	2016	Bangalore	3	27	Male	No	5	1
Masters	2017	Pune	3	24	Male	Yes	2	1
Bachelors	2016	Bangalore	3	22	Male	No	0	0
Bachelors	2015	New Delhi	3	38	Male	No	0	0
Bachelors	2016	Bangalore	3	34	Female	No	2	1
Bachelors	2016	Pune	3	23	Male	No	1	0
Masters	2017	New Delhi	2	37	Male	No	2	0
Masters	2012	Bangalore	3	27	Male	No	5	1
Bachelors	2016	Pune	3	34	Male	No	3	0
Bachelors	2018	Pune	3	32	Male	Yes	5	1
Bachelors	2016	Bangalore	3	39	Male	No	2	0
Bachelors	2012	Bangalore	3	37	Male	No	4	0
Bachelors	2017	Bangalore	1	29	Male	No	3	0
Bachelors	2014	Bangalore	3	34	Female	No	2	0
Bachelors	2014	Pune	1	14	Male	No	4	0

4,653 rows - 9 columns (5 nominal, 4 numerical)

Using Auto Model, we tune the setting for model creation and select the output field.

Auto Model

Load Data | Select Task | Prepare Target | Select Inputs | Model Types | Results

RESTART BACK NEXT

Predict
Want to predict the values of a column?

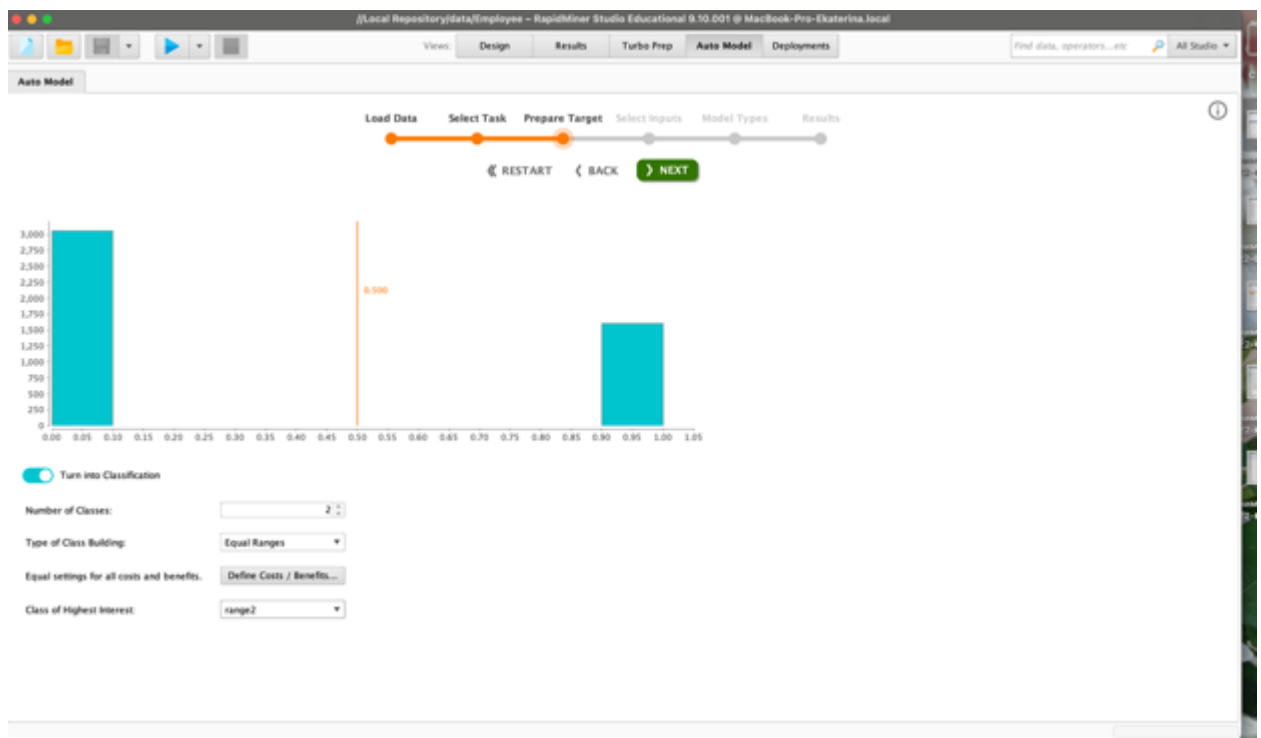
Clusters
Want to identify groups in your data?

Outliers
Want to detect outliers in your data?

Education Category	Joining Year	City Category	Payment Tier Number	Age Number	Gender Category	Ever Benched Category	Experience in Current...	Leave Or Not Number
Bachelors	2017	Bangalore	3	34	Male	No	0	0
Bachelors	2013	Pune	1	28	Female	No	3	1
Bachelors	2014	New Delhi	3	38	Female	No	2	0
Masters	2016	Bangalore	3	27	Male	No	5	1
Masters	2017	Pune	3	24	Male	Yes	2	1
Bachelors	2016	Bangalore	3	22	Male	No	0	0
Bachelors	2015	New Delhi	3	38	Male	No	0	0
Bachelors	2016	Bangalore	3	34	Female	No	2	1
Bachelors	2016	Pune	3	23	Male	No	1	0
Masters	2017	New Delhi	2	37	Male	No	2	0
Masters	2012	Bangalore	3	27	Male	No	5	1
Bachelors	2016	Pune	3	34	Male	No	3	0
Bachelors	2018	Pune	3	32	Male	Yes	5	1
Bachelors	2016	Bangalore	3	39	Male	No	2	0
Bachelors	2012	Bangalore	3	37	Male	No	4	0
Bachelors	2017	Bangalore	1	29	Male	No	3	0

4,653 rows - 9 columns (4 nominal, 5 numerical)

We evaluate the dataset based on the target and see that we have a class imbalance.



We evaluate the input data and choose the fields that will be used to train the model. All fields in the provided dataset need to be included for the model creation.

Workflow: Load Data → Select Task → Prepare Target → **Select Inputs** → Model Types → Results

Selected: 8 / Total: 8

Buttons: ☒ Select All ☒ Deselect All

Selected	Status	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
<input checked="" type="checkbox"/>	●		Education	0.65%	0.06%	77.39%	0.00%	3.75%
<input checked="" type="checkbox"/>	●		JoiningYear	3.30%	0.15%	23.81%	0.00%	0.00%
<input checked="" type="checkbox"/>	●		City	0.59%	0.06%	47.88%	0.00%	11.70%
<input checked="" type="checkbox"/>	●		PaymentTier	3.91%	0.06%	75.05%	0.00%	0.00%
<input checked="" type="checkbox"/>	●		Age	0.26%	0.43%	13.86%	0.00%	0.00%
<input checked="" type="checkbox"/>	●		Gender	4.87%	0.04%	59.70%	0.00%	2.15%
<input checked="" type="checkbox"/>	●		EverBenchd	0.62%	0.04%	89.73%	0.00%	0.95%
<input checked="" type="checkbox"/>	●		ExperienceCurrentDomain	0.09%	0.17%	23.36%	0.00%	0.00%

Choosing the models

We need to select what models will be trained, so we only keep those we want to use.

Workflow: Load Data → Select Task → Prepare Target → Select Inputs → **Model Types** → Results

Buttons: ☒ Select All ☒ Deselect All

Execution

Execute on: Local Computer (this machine)

Queue: No queues available

Select Folder for Storing Results

The results of this run will be stored in the folder selected below. We recommend to use an empty folder in the selected AI Hub repository.

- Local Repository Local
- Temporary Repository Local

Models

- ☐ Naive Bayes
- ☒ Generalized Linear Model
 - ☒ Use Regularization ☐ Calculate p-Values
- ☒ Logistic Regression
- ☒ Fast Large Margin
- ☒ Automatically Optimize
- ☐ Deep Learning
- ☒ Decision Tree
 - ☒ Automatically Optimize Maximal Depth: 20
- ☒ Random Forest
 - ☒ Automatically Optimize Number of Trees: 20 Maximal Depth: 20
- ☒ Gradient Boosted Trees
 - ☒ Automatically Optimize Number of Trees: 20 Maximal Depth: 20 Learning Rate: 0.01
- ☐ Support Vector Machine
 - ☒ Automatically Optimize

Data Preparation

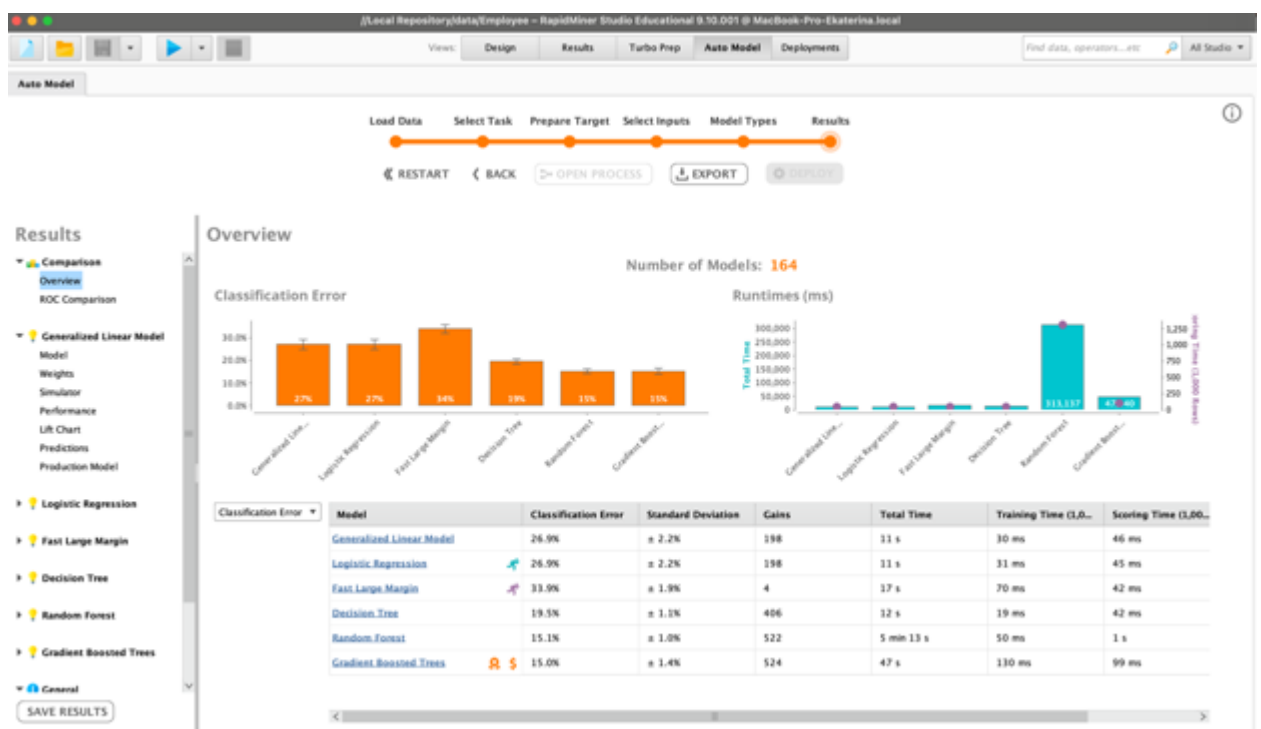
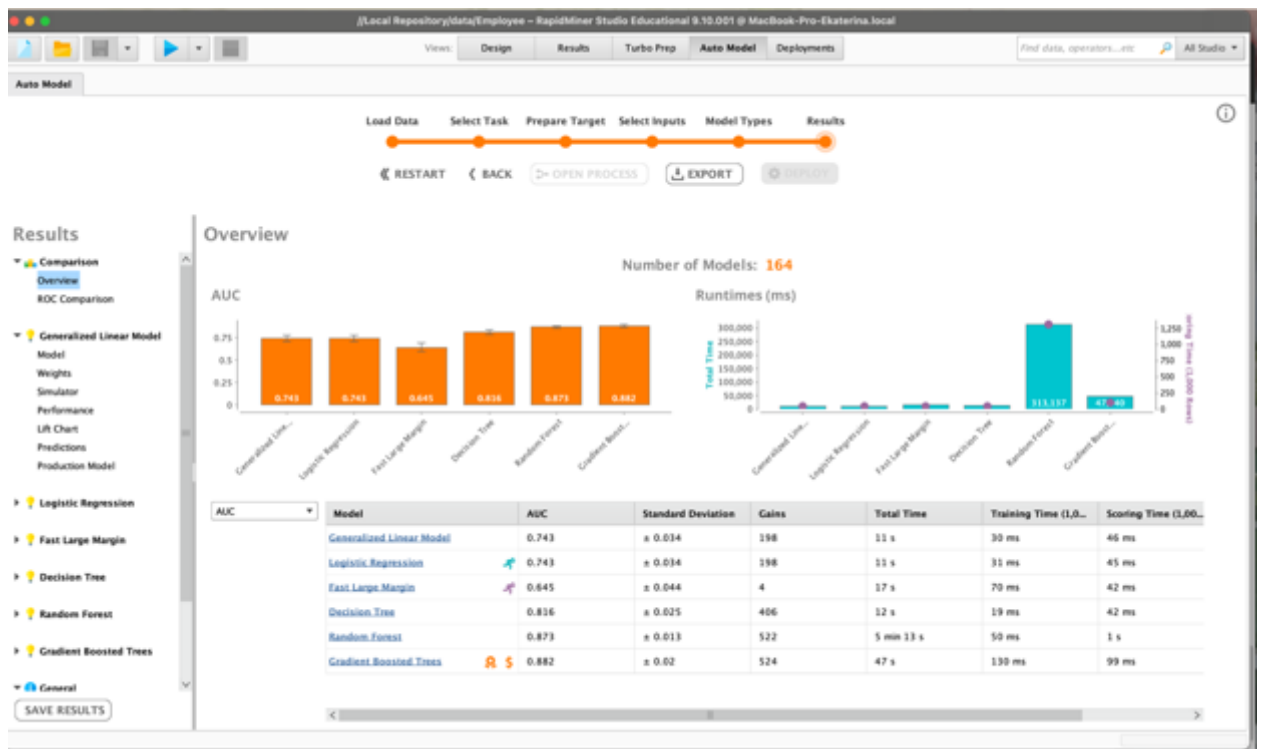
- ☒ Remove Columns with Too Many Values
 - Maximum Number of Values: 50
- ☐ Extract Date Information
- ☐ Extract Text Information
 - Select Text Columns List
 - Number of Extracted Features: 1,000
- ☐ Automatic Feature Selection
 - Additional Minutes (Maximum): 60
 - Final Feature Set should be: Accurate
- ☐ Automatic Feature Generation
 - Function Complexity can be: Medium

Column Analysis

- ☒ Correlations between Columns
- ☒ Importance of Columns
- ☒ Explain Predictions

Evaluating the models

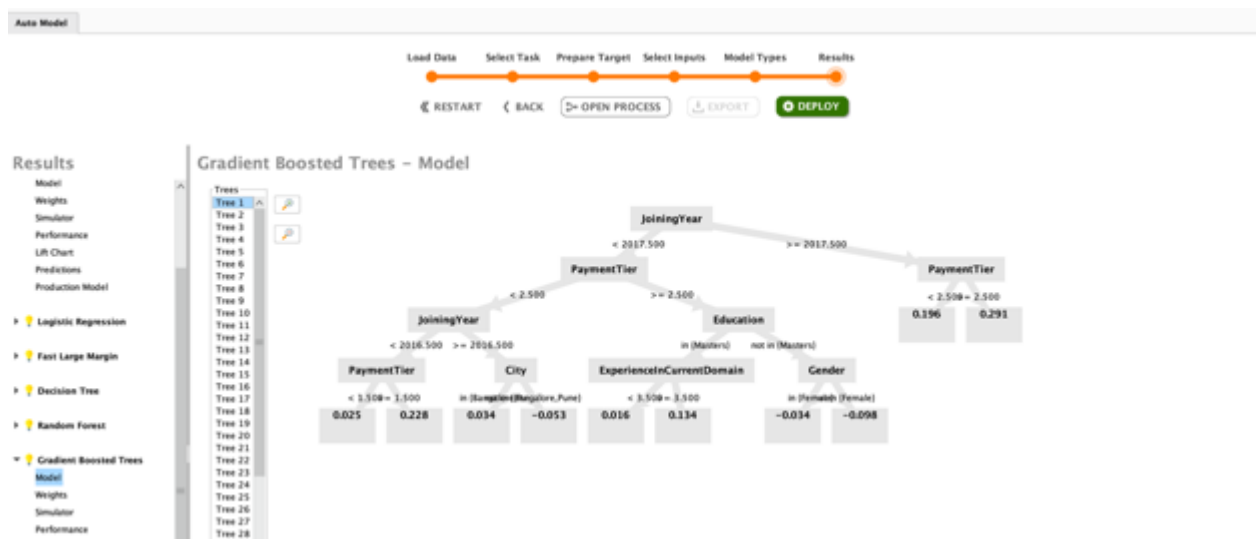
The next step is to evaluate the results of models we trained. We will select the best model based on AUC and **Classification error** parameters. The best model is **Gradient Boosted Trees**.



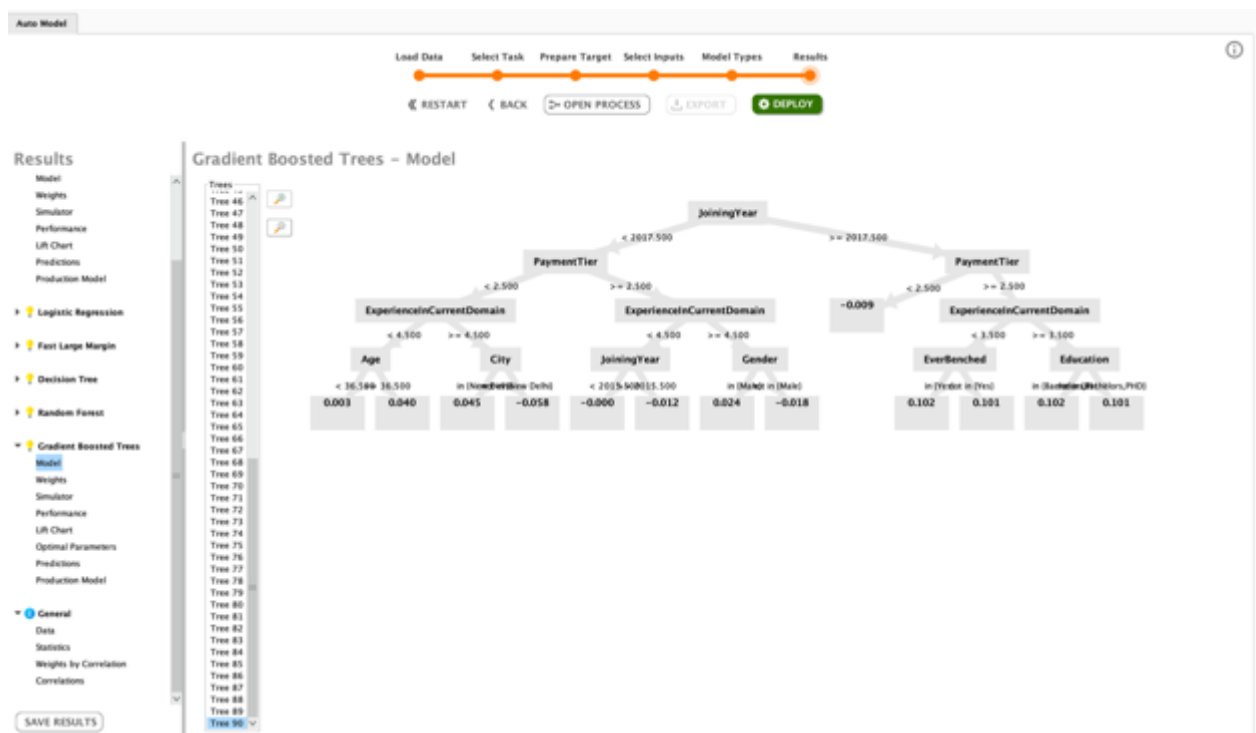
The best model in details

Let's have a closer look at the best model we received - Gradient Boosted Trees. The model has 90 Solution Trees.

The first Solution Tree has the following structure:

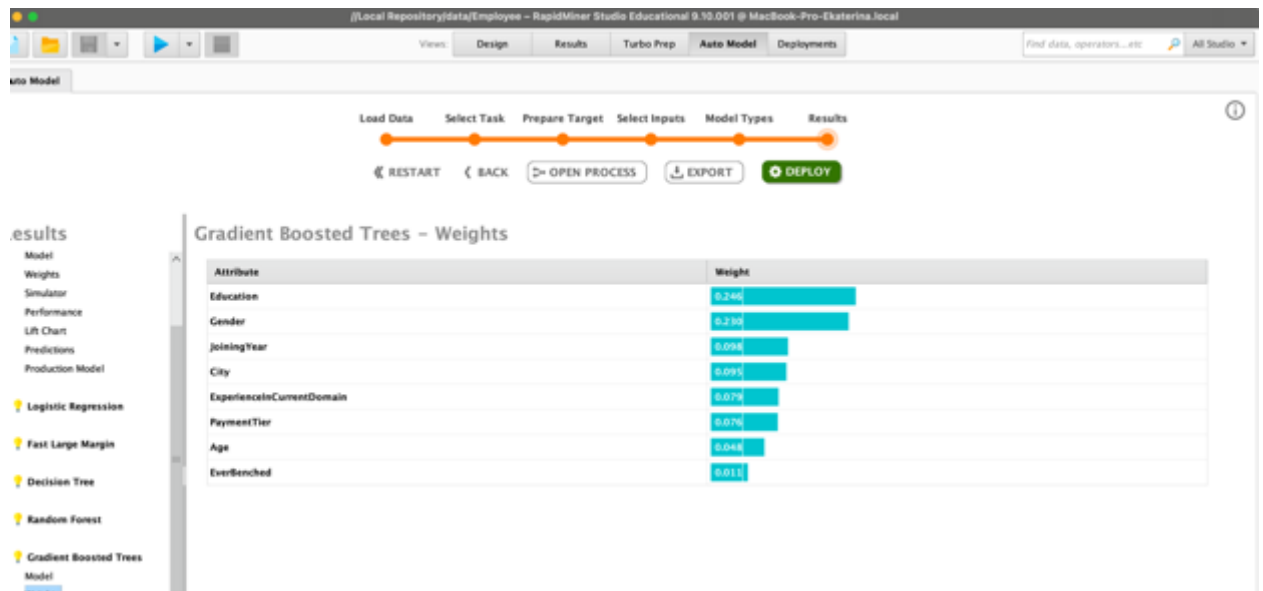


The best Solution Tree has the structure below:



Most significant fields for the model

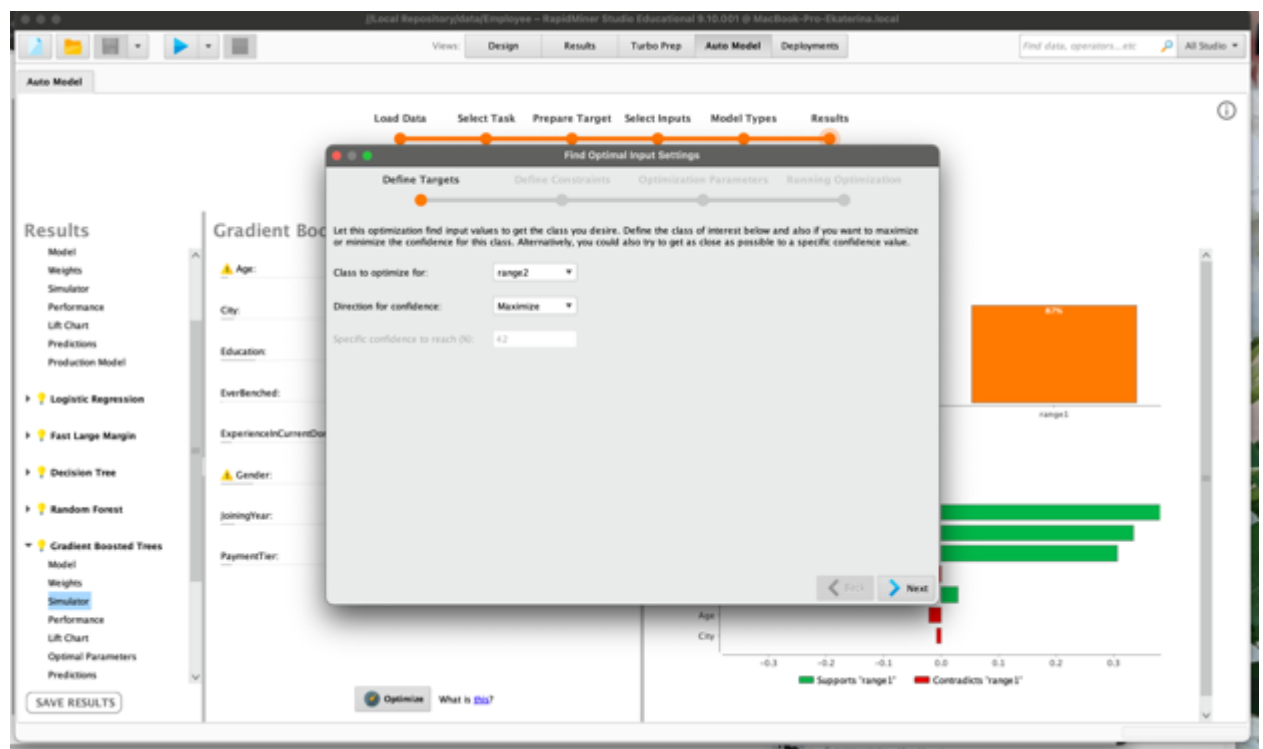
Let's see what fields had the most significant impact on the solution when creating the model.

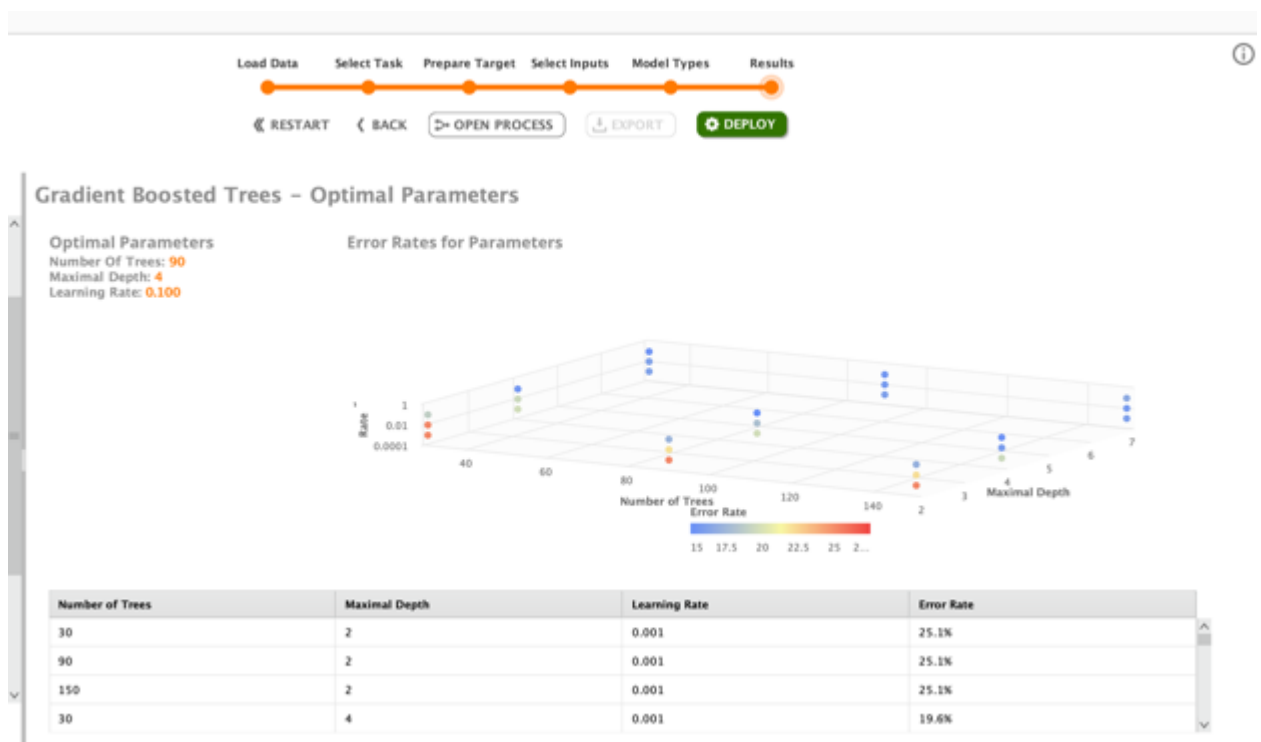
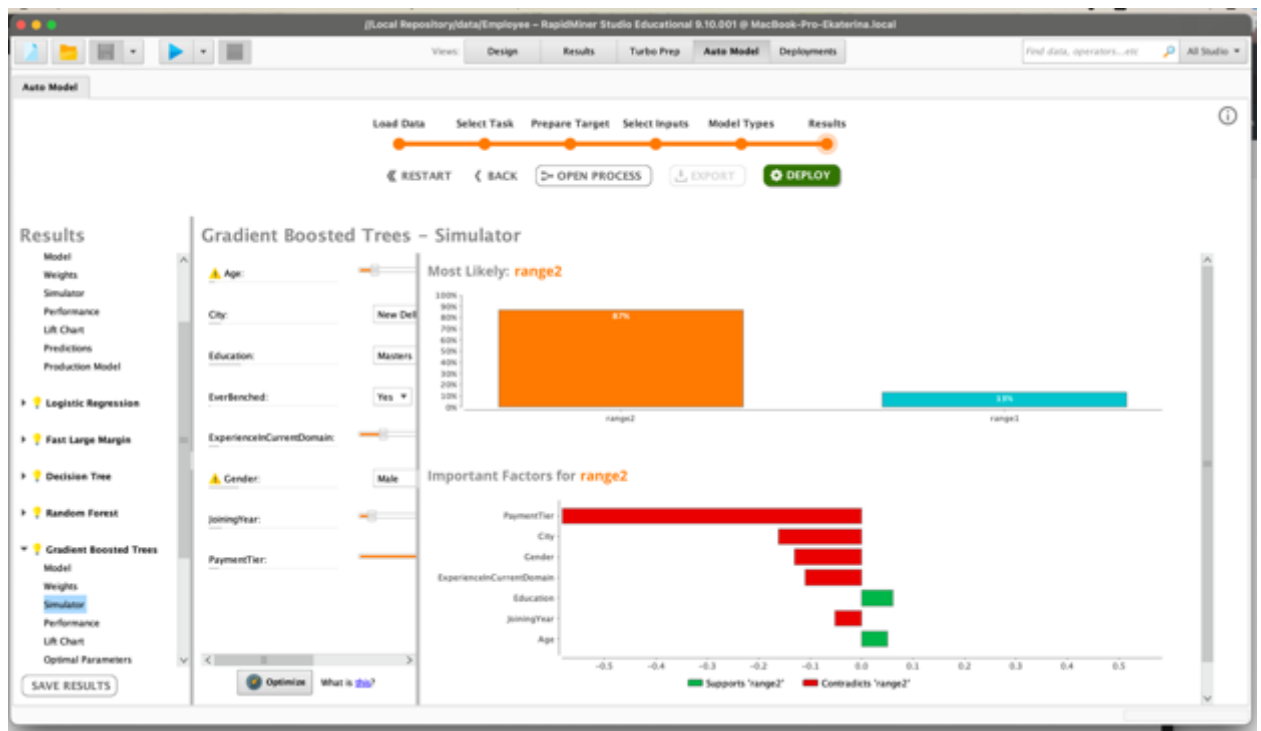


The highest weight in this model is shown for the following fields: **Education, Gender, JoiningYear**.

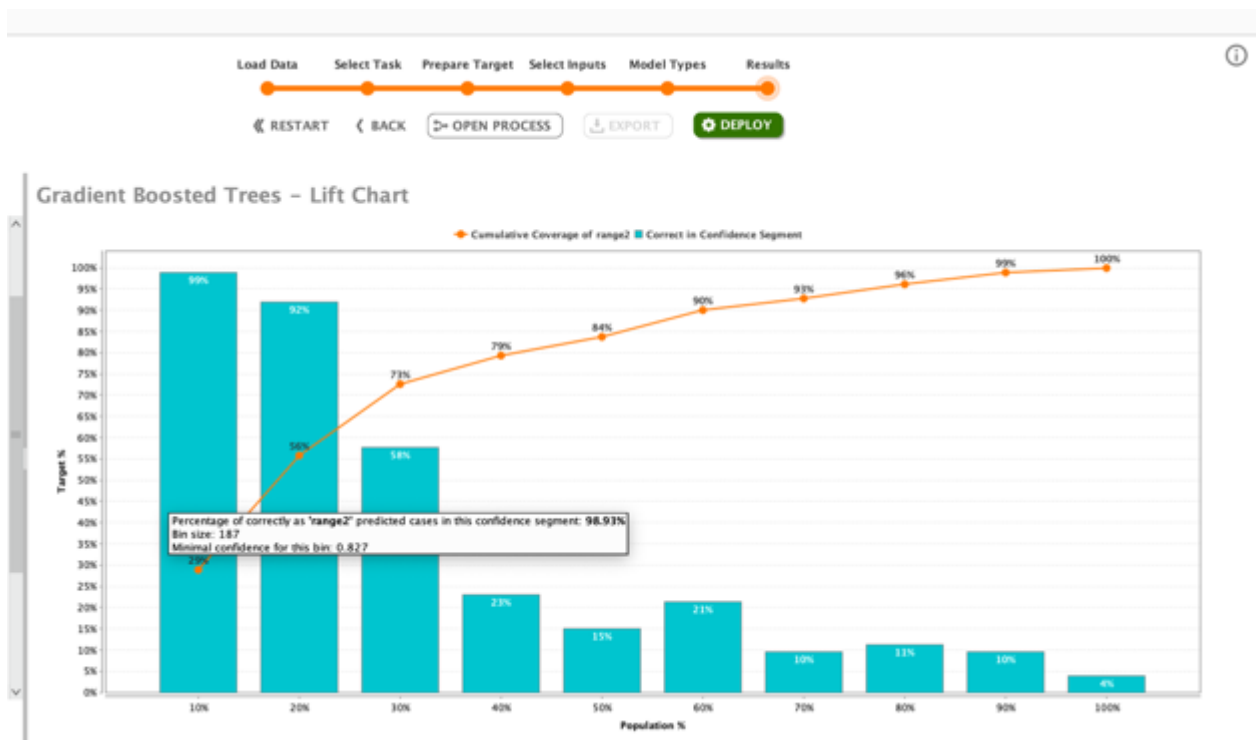
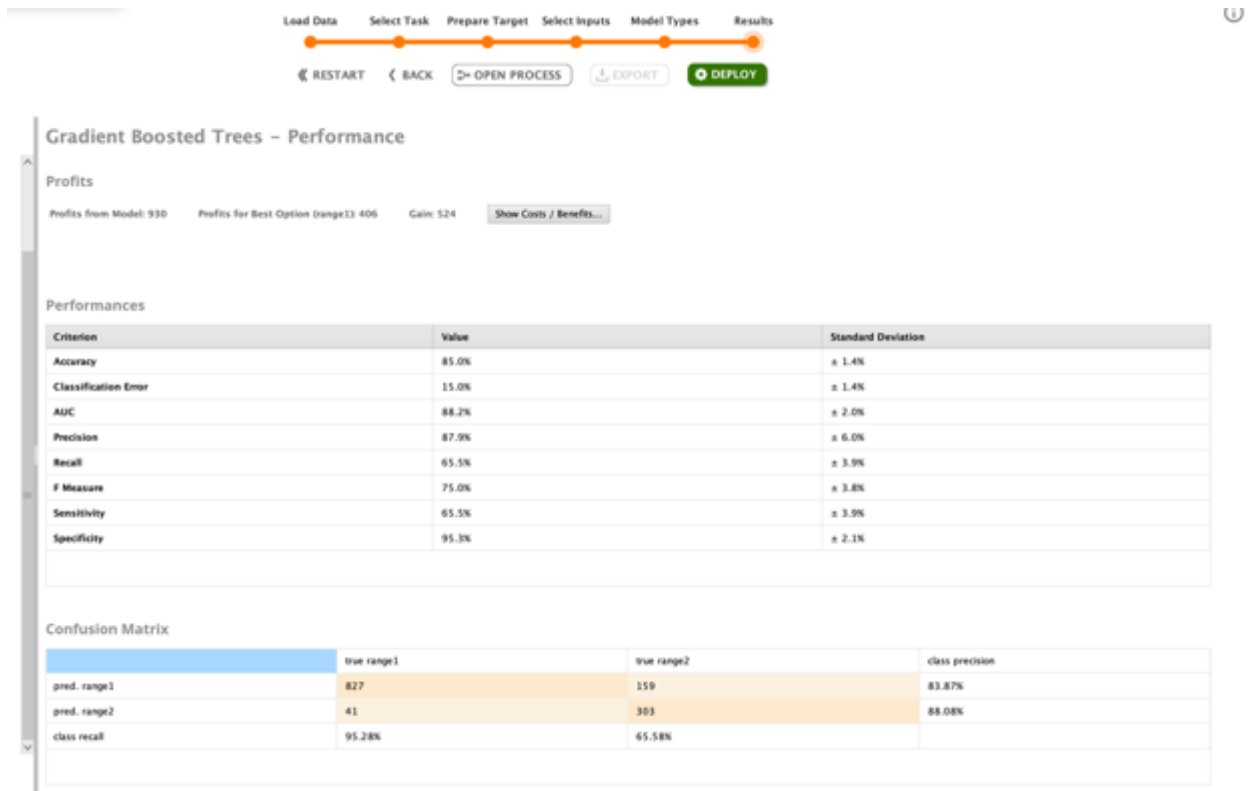
It means that **Education, Gender, JoiningYear** are the parameters that have the most impact on if the employee will leave the company or not.

Optimizing the model parameters





Let's have a look at the rest of the metrics, they all are quite good.



Now let's compare the results of the prediction provided by the model and the results of the target field that we have in provided dataset.

Load Data Select Task Prepare Target Select Inputs Model Types Results

RESTART BACK OPEN PROCESS EXPORT DEPLOY

Gradient Boosted Trees – Predictions

Row No.	LeaveOrNot	prediction...	confidence...	confidence...	cost	Education	City	Gender	EverBnched	JoiningYear	PaymentTier	Age	Experience...
1	range2	range2	0.447	0.553	0.105	Masters	Pune	Male	Yes	2017	3	24	2
2	range1	range1	0.896	0.104	0.792	Bachelors	New Delhi	Male	No	2015	3	38	0
3	range1	range1	0.864	0.136	0.727	Masters	New Delhi	Male	No	2017	2	37	2
4	range2	range2	0.191	0.809	0.618	Masters	Bangalore	Male	No	2012	3	27	5
5	range1	range1	0.866	0.134	0.733	Bachelors	Bangalore	Male	No	2016	3	39	2
6	range1	range1	0.868	0.132	0.737	Bachelors	Bangalore	Male	Yes	2015	3	27	5
7	range1	range1	0.882	0.118	0.763	Bachelors	Bangalore	Male	No	2017	3	29	4
8	range1	range1	0.834	0.166	0.669	Bachelors	Bangalore	Male	No	2015	3	23	1
9	range2	range2	0.176	0.824	0.648	Bachelors	Pune	Female	No	2013	2	31	2
10	range1	range1	0.842	0.158	0.683	Bachelors	Bangalore	Male	No	2014	3	23	1
11	range1	range1	0.880	0.120	0.760	Bachelors	Bangalore	Male	No	2016	3	40	5
12	range2	range2	0.163	0.837	0.673	Bachelors	New Delhi	Female	No	2018	2	34	0
13	range1	range1	0.907	0.093	0.814	Bachelors	Pune	Male	Yes	2014	3	30	4
14	range1	range2	0.311	0.689	0.377	Masters	New Delhi	Male	No	2017	2	23	1
15	range1	range1	0.795	0.205	0.590	Bachelors	Bangalore	Male	No	2014	3	36	0
16	range1	range1	0.700	0.300	0.400	Bachelors	Bangalore	Female	No	2013	3	30	1
17	range2	range2	0.145	0.855	0.711	Bachelors	Pune	Female	No	2015	2	26	4
18	range1	range1	0.841	0.159	0.682	Bachelors	Bangalore	Female	Yes	2014	3	31	5
19	range1	range1	0.887	0.113	0.775	Bachelors	New Delhi	Female	Yes	2017	3	31	5
20	range1	range1	0.921	0.079	0.842	PHD	New Delhi	Male	No	2013	3	28	2
21	range1	range1	0.891	0.109	0.783	Bachelors	Bangalore	Male	No	2016	3	38	2
22	range1	range1	0.912	0.088	0.824	Masters	Pune	Female	No	2014	3	39	2
23	range1	range1	0.726	0.274	0.451	Masters	New Delhi	Male	No	2013	3	29	3
24	range1	range1	0.866	0.134	0.732	Bachelors	Bangalore	Male	No	2015	3	30	5
25	range2	range2	0.203	0.797	0.594	PHD	Bangalore	Male	No	2013	2	25	1

We can export the results of prediction provided by the model if it is needed.

Load Data Select Task Prepare Target Select Inputs Model Types Results

Export Data 'Employee Temp 66942 Predictions'

Select Format Select Location Writing Done

Turbo Prep
Open the data in Turbo Prep

Repository
Store the data in a repository

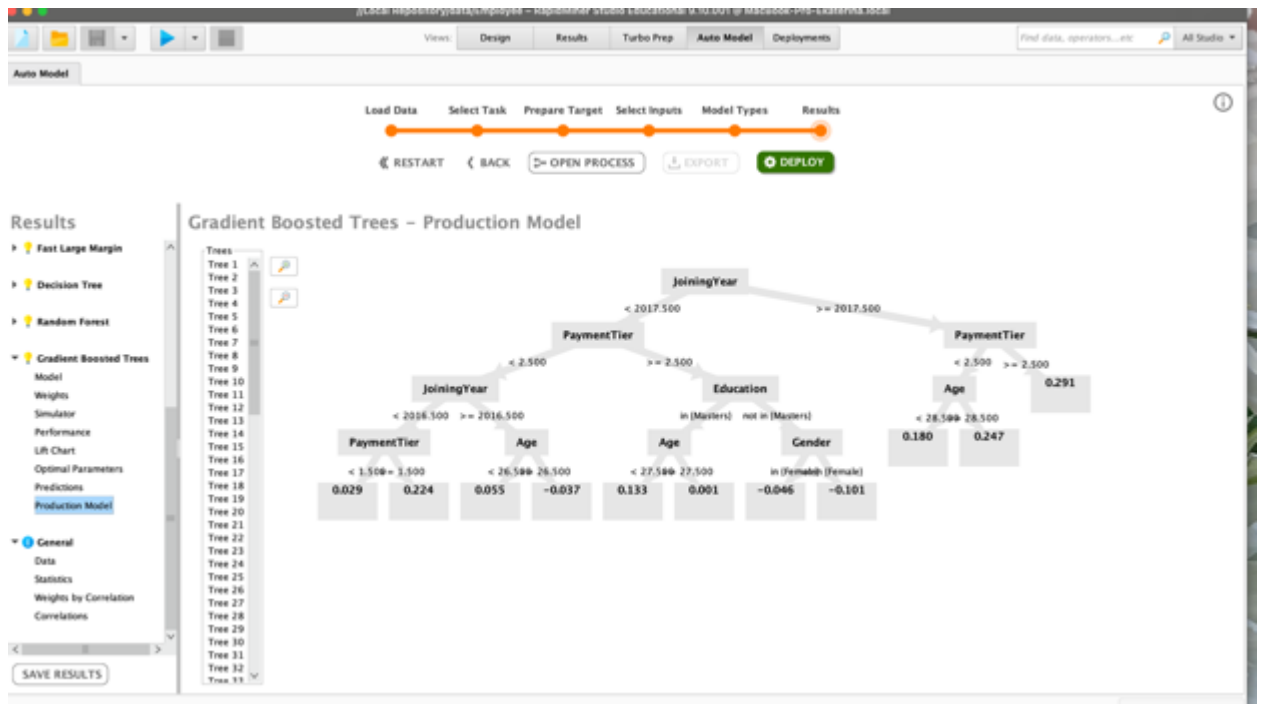
Qlik
Export the data for Qlik (qlik)

Excel
Export the data as Excel file (.xlsx)

CSV
Export the data as CSV file (.csv)

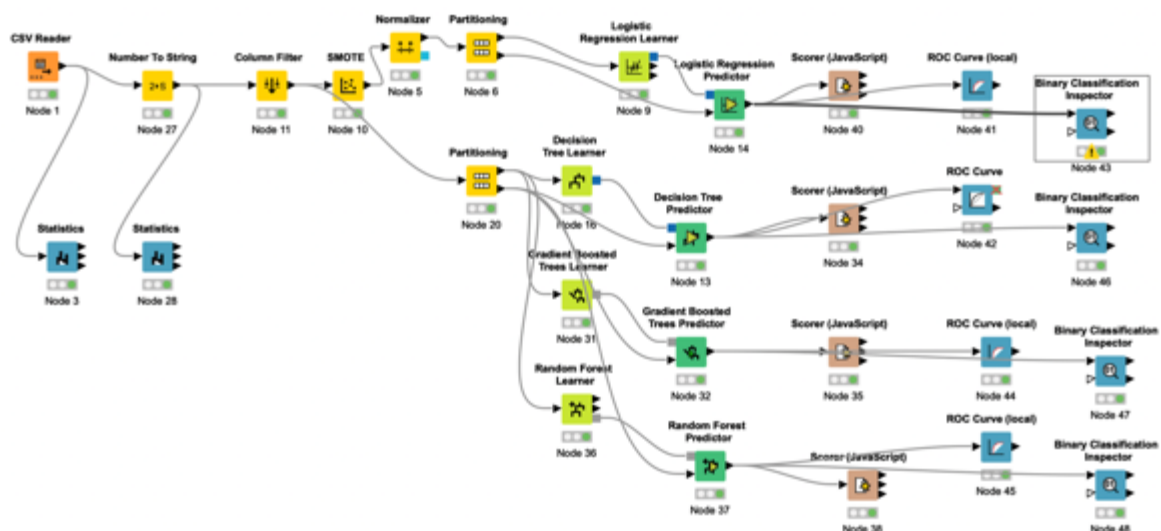
BACK NEXT

The structure of predictive model (Gradient Boosted Trees)



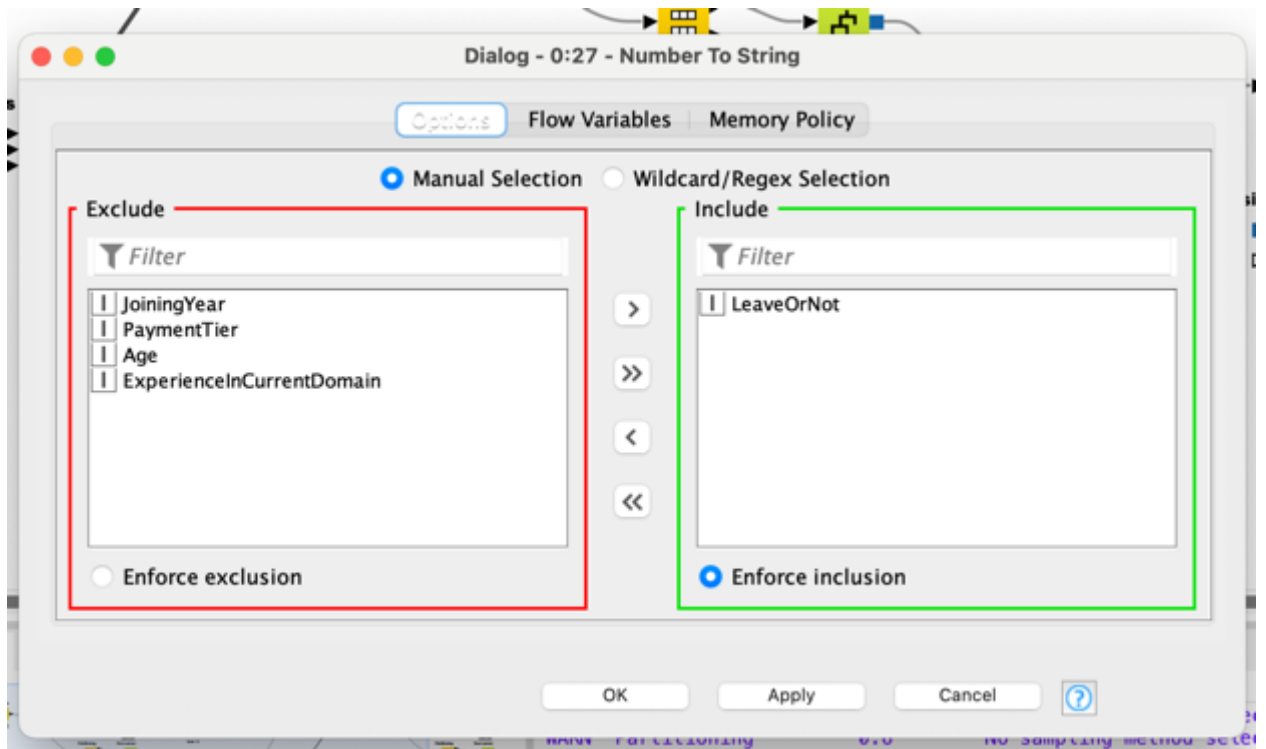
3. Knime Analytics Platform

Let's use **Knime Analytics Platform** and create and train 4 ML models: **Logistic Regression**, **Gradient Boosted Trees**, **Decision Tree**, **Random Forest**.

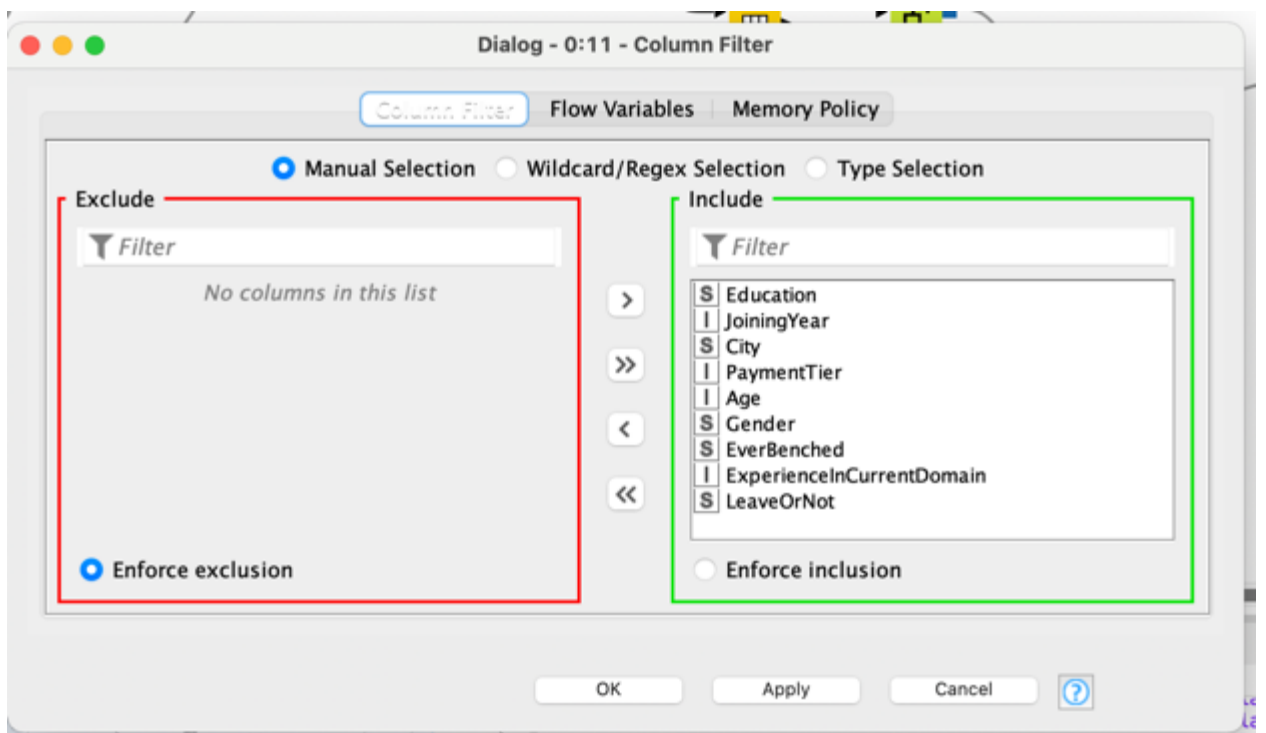


Converting the target field to categorical type

The target field is numerical now, but in order to use it as a target field for classification task, we need to convert it to categorical type.



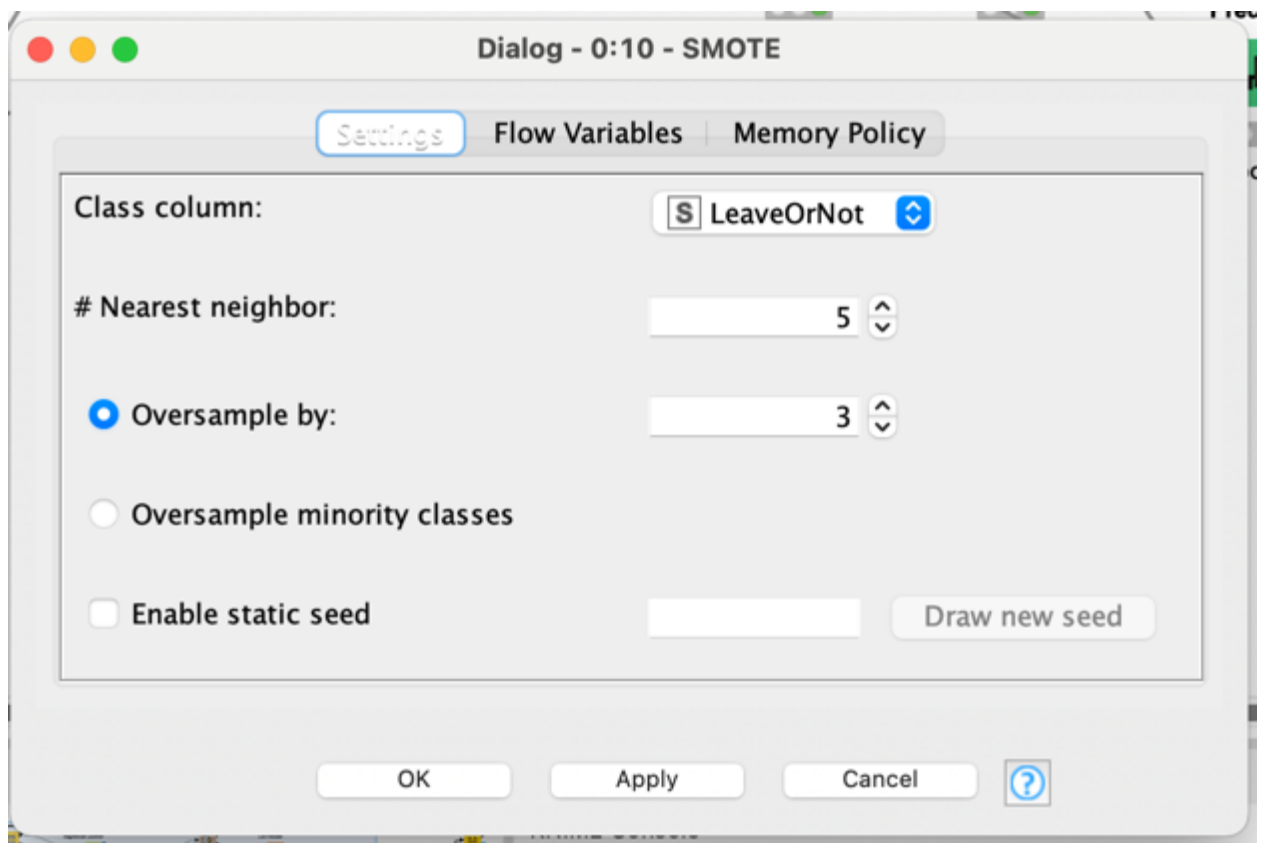
The dataset is clean, there are no outliers, missed values, all fields will be used for creating the model, no need to drop any fields.



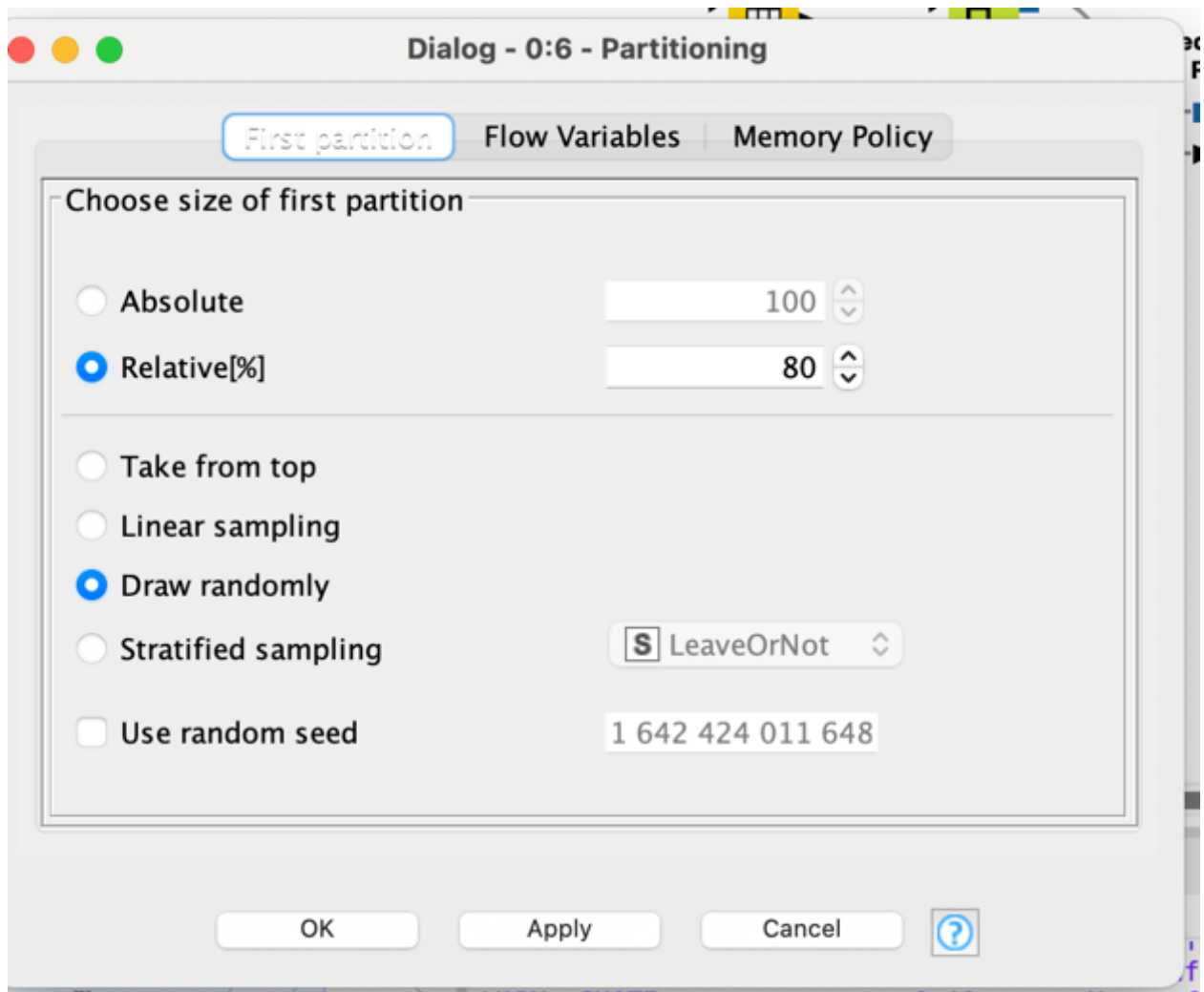
SMOTE module to fix the class imbalance

As we already know, there is a class imbalance in this dataset, that means that there is a severe skew in the class distribution and this class imbalance may affect the model that will be created. If we train the model on the imbalanced data, the predictions would be skewed as well.

To alleviate the imbalance, we will use SMOTE module to artificially increase the data to balance the dataset.



Let's split the dataset into the training (80%) and test sets (20%).



Comparing the models
Logistic Regression

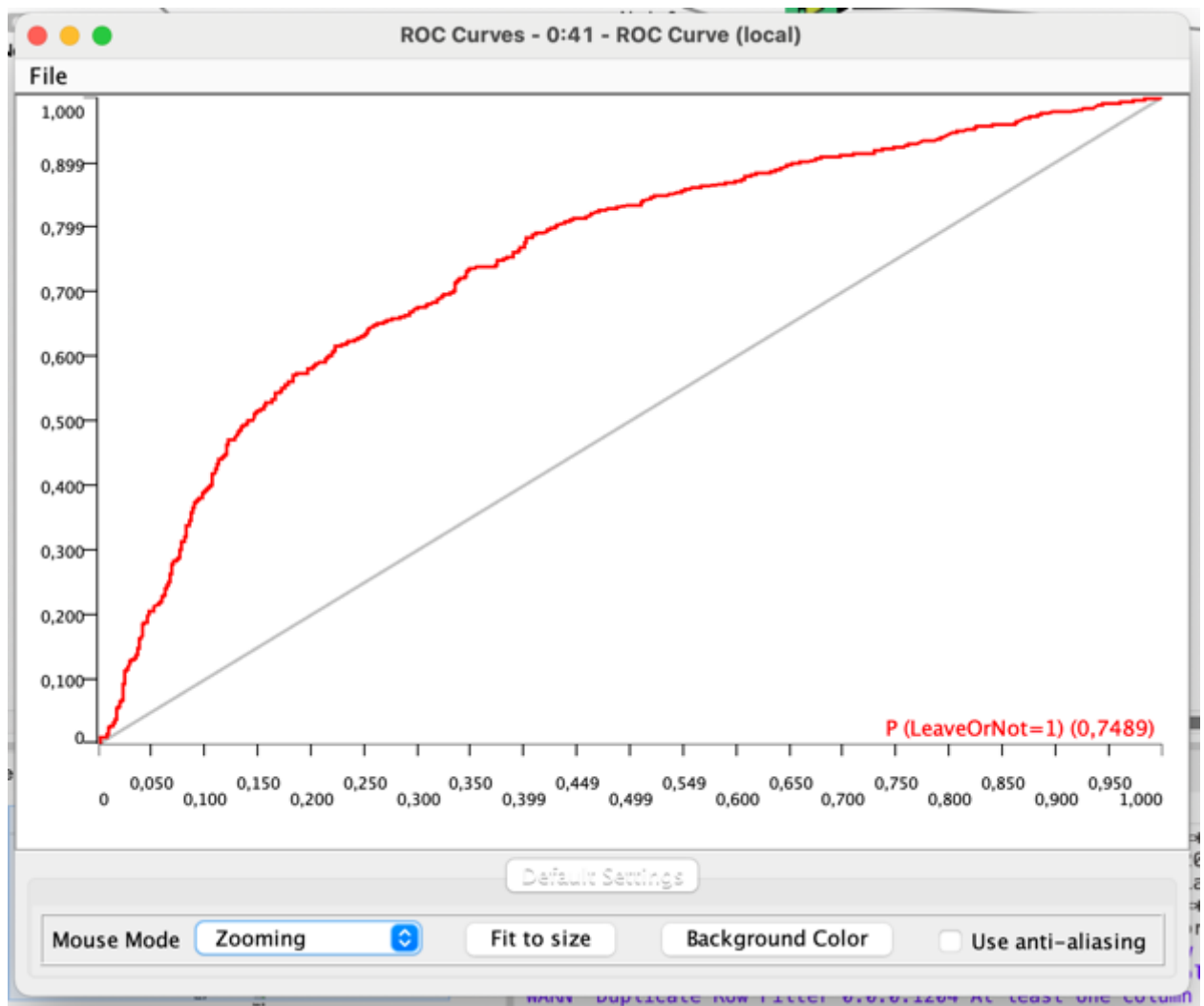
Confusion Matrix

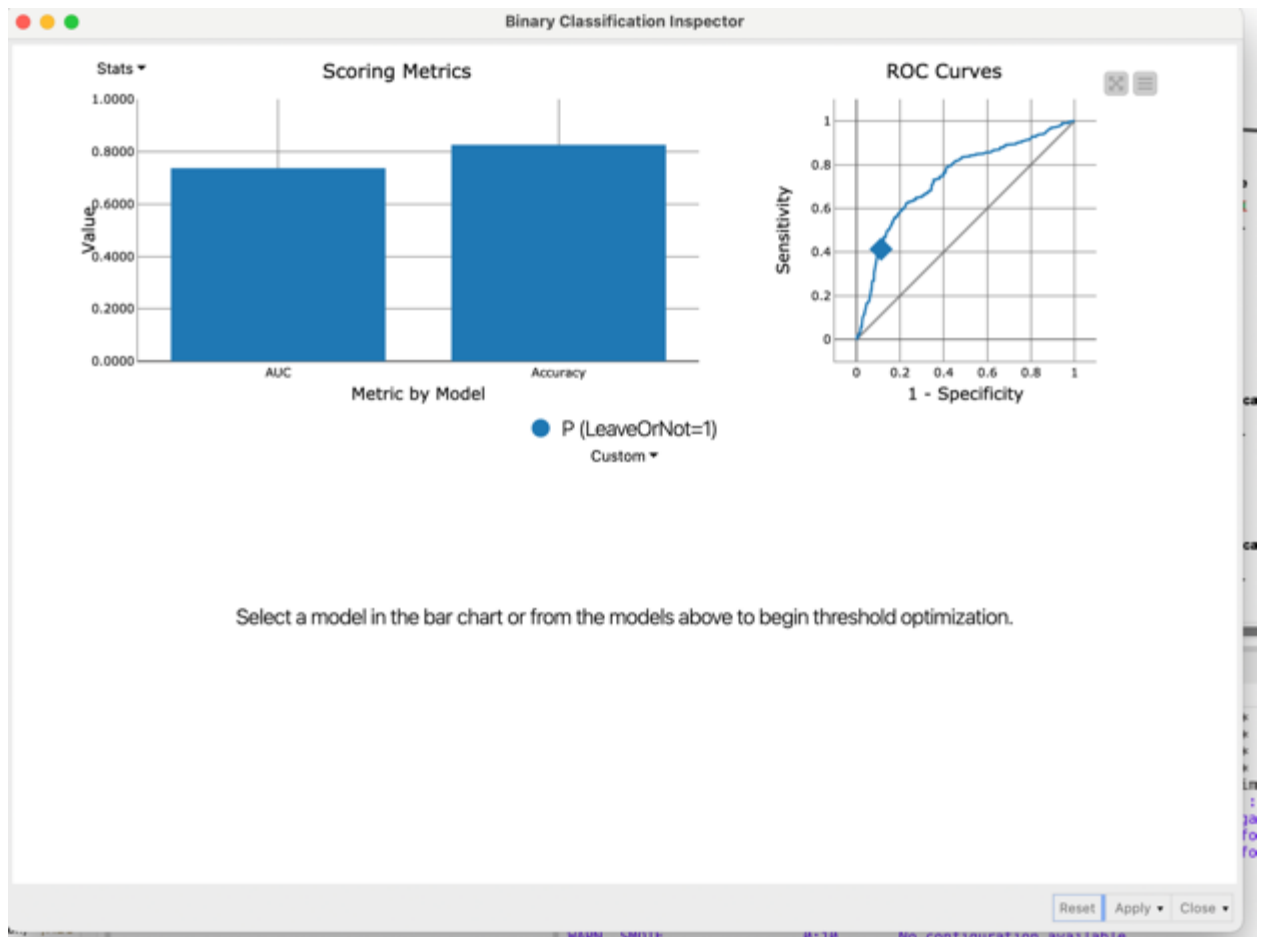
Scorer View
Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	2209	268	89.18%
1 (Actual)	724	522	41.89%
	75.32%	66.08%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
73.35%	26.65%	0.342	2731	992





Decision Tree

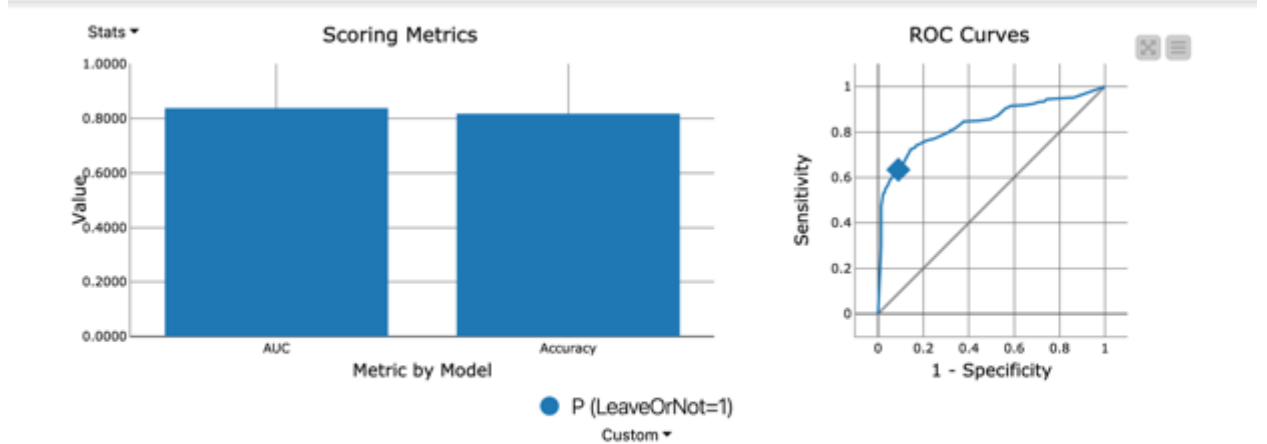
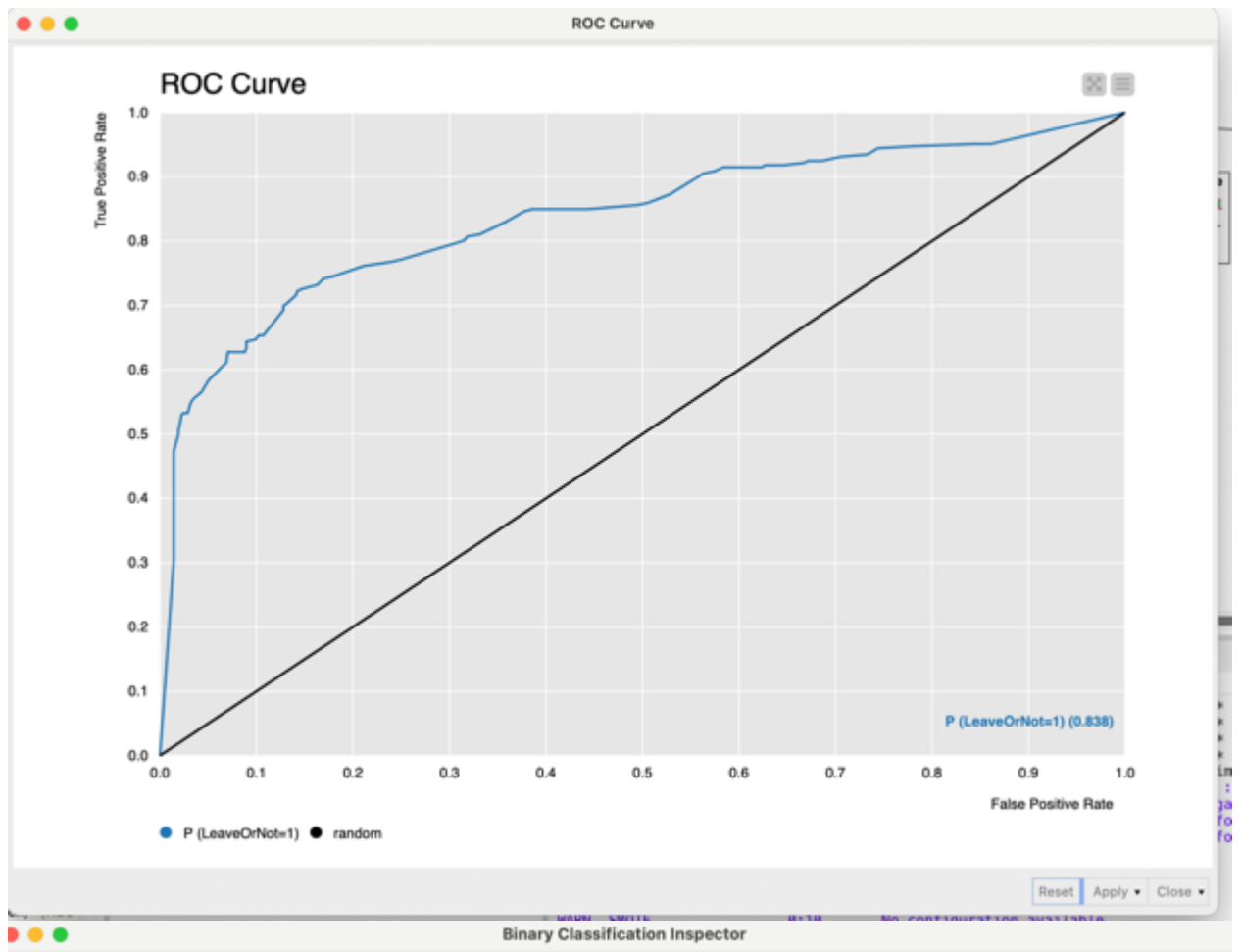
Scorer View

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	570	55	91.20%
1 (Actual)	114	192	62.75%
	83.33%	77.73%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
81.85%	18.15%	0.567	762	169



Gradient Boosted Tree

Scorer View

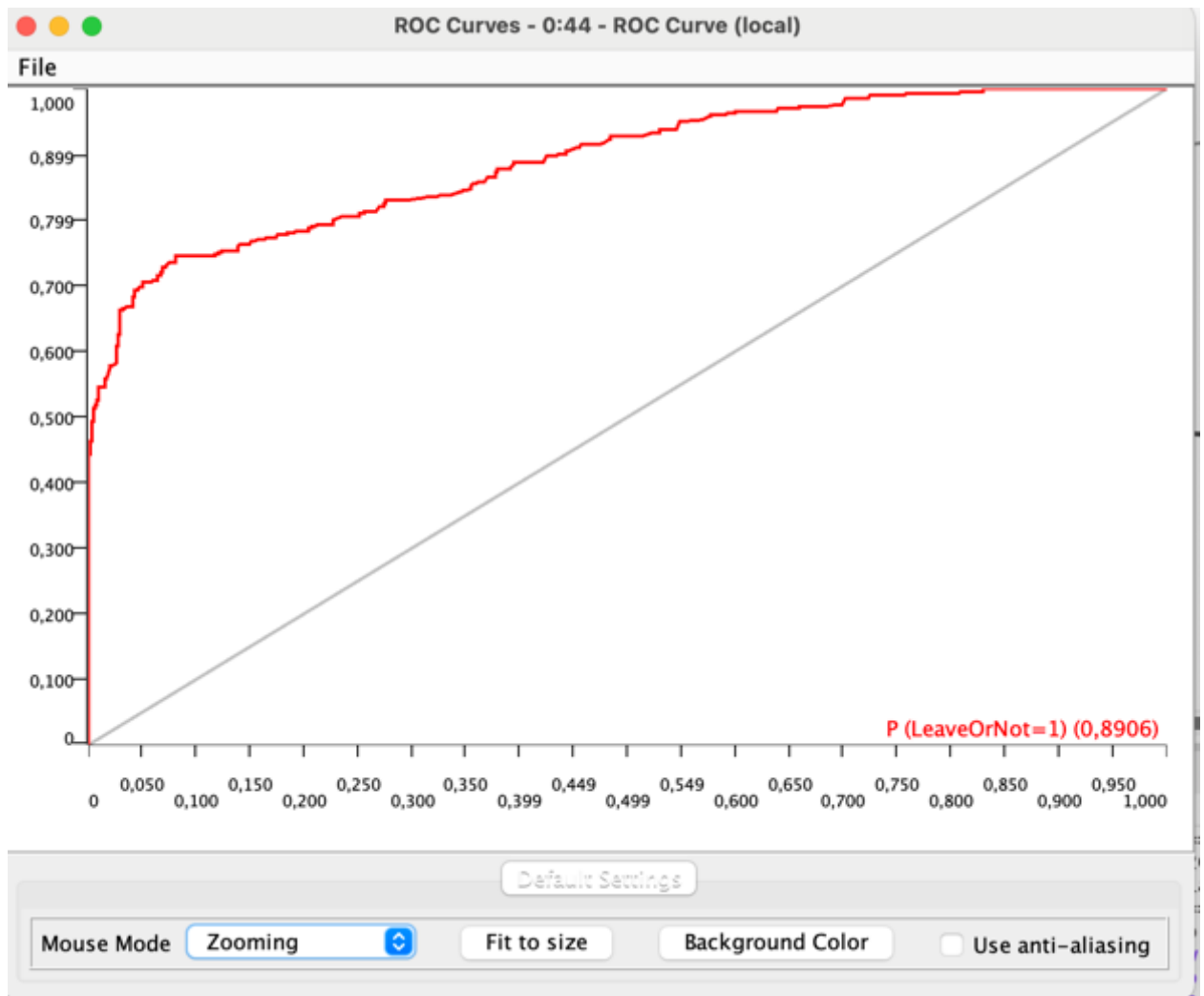
Confusion Matrix

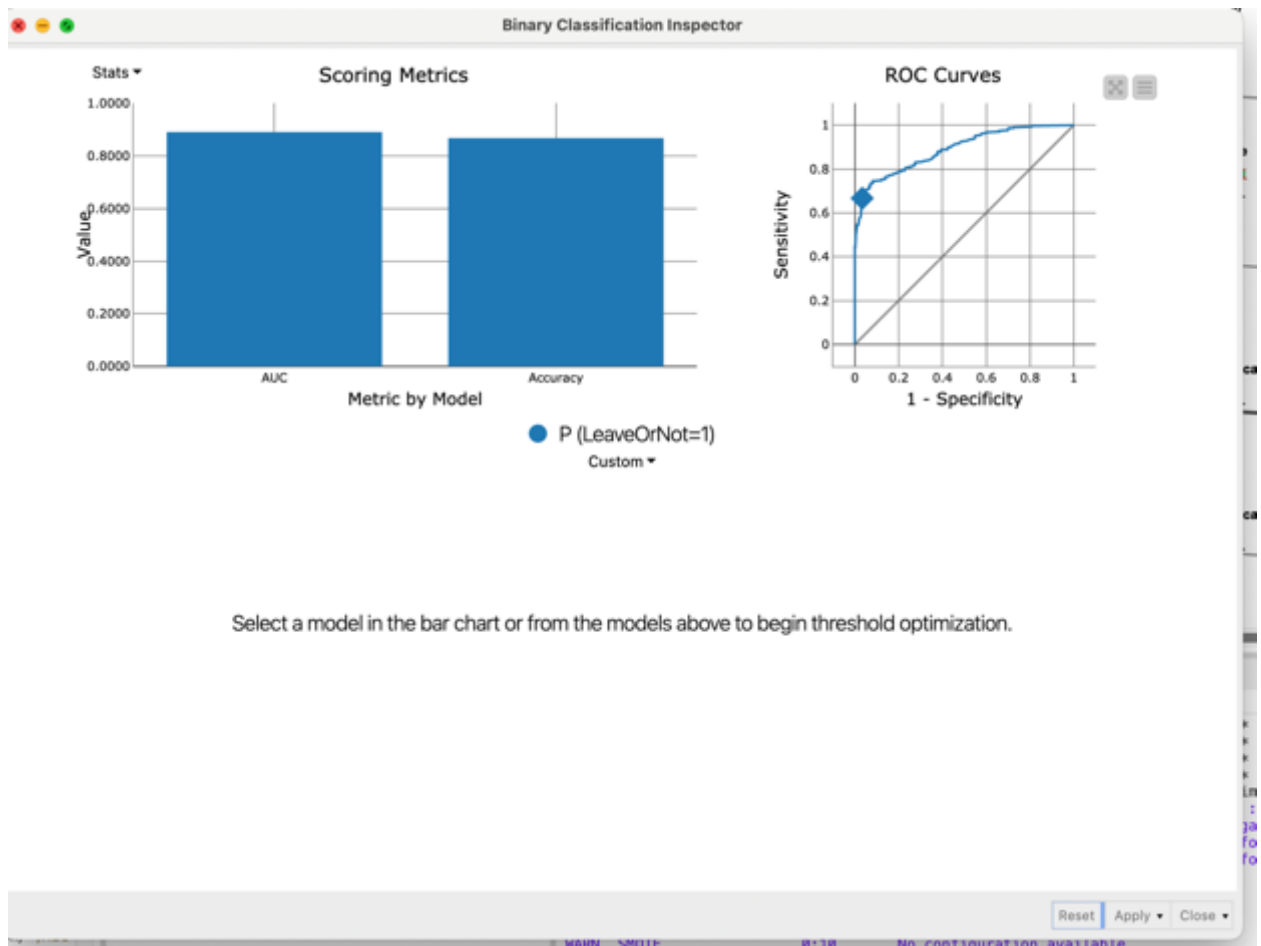


	0 (Predicted)	1 (Predicted)	
0 (Actual)	605	20	96.80%
1 (Actual)	102	204	66.67%
	85.57%	91.07%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
86.90%	13.10%	0.681	809	122





Random Forest

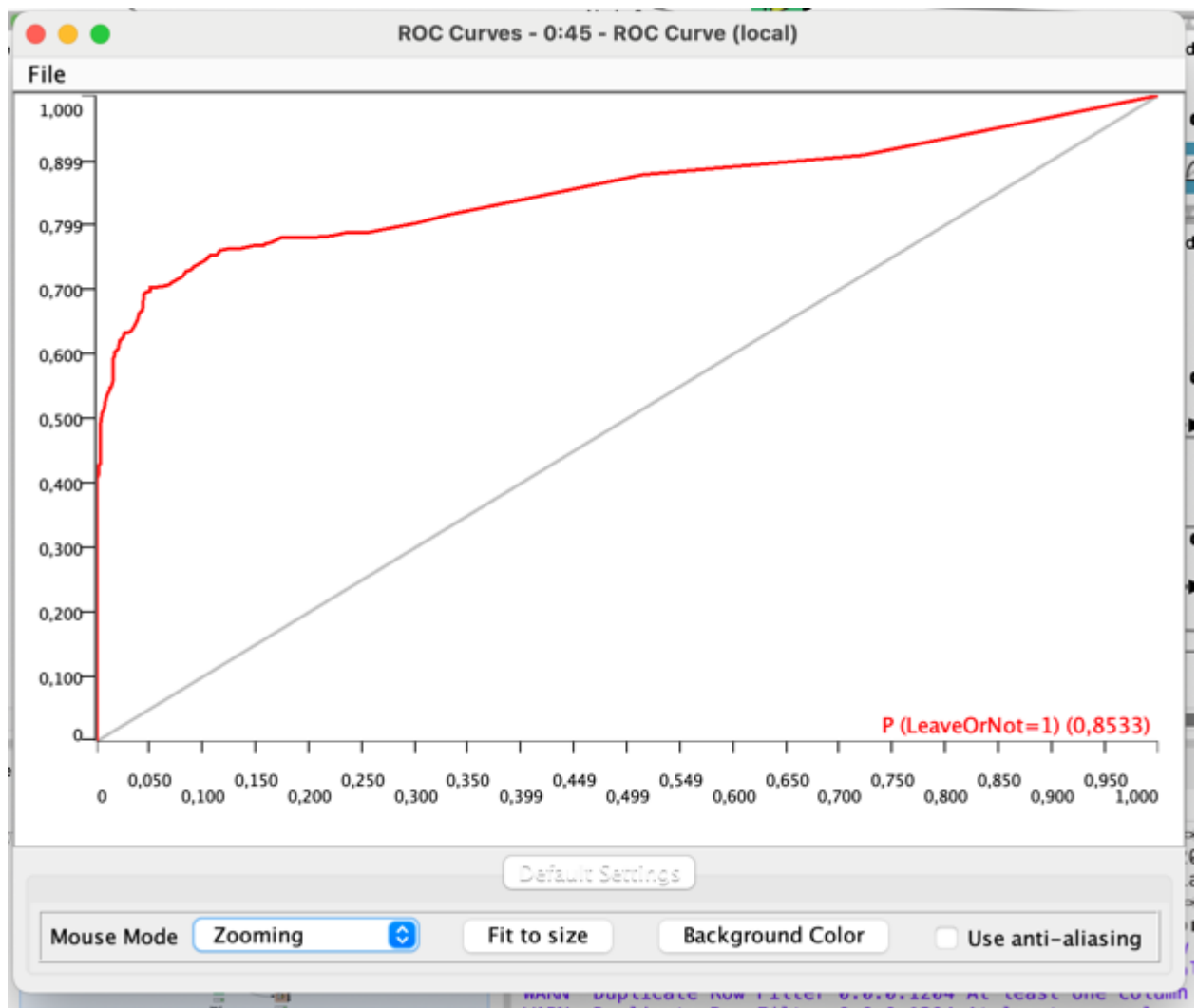
Scorer View

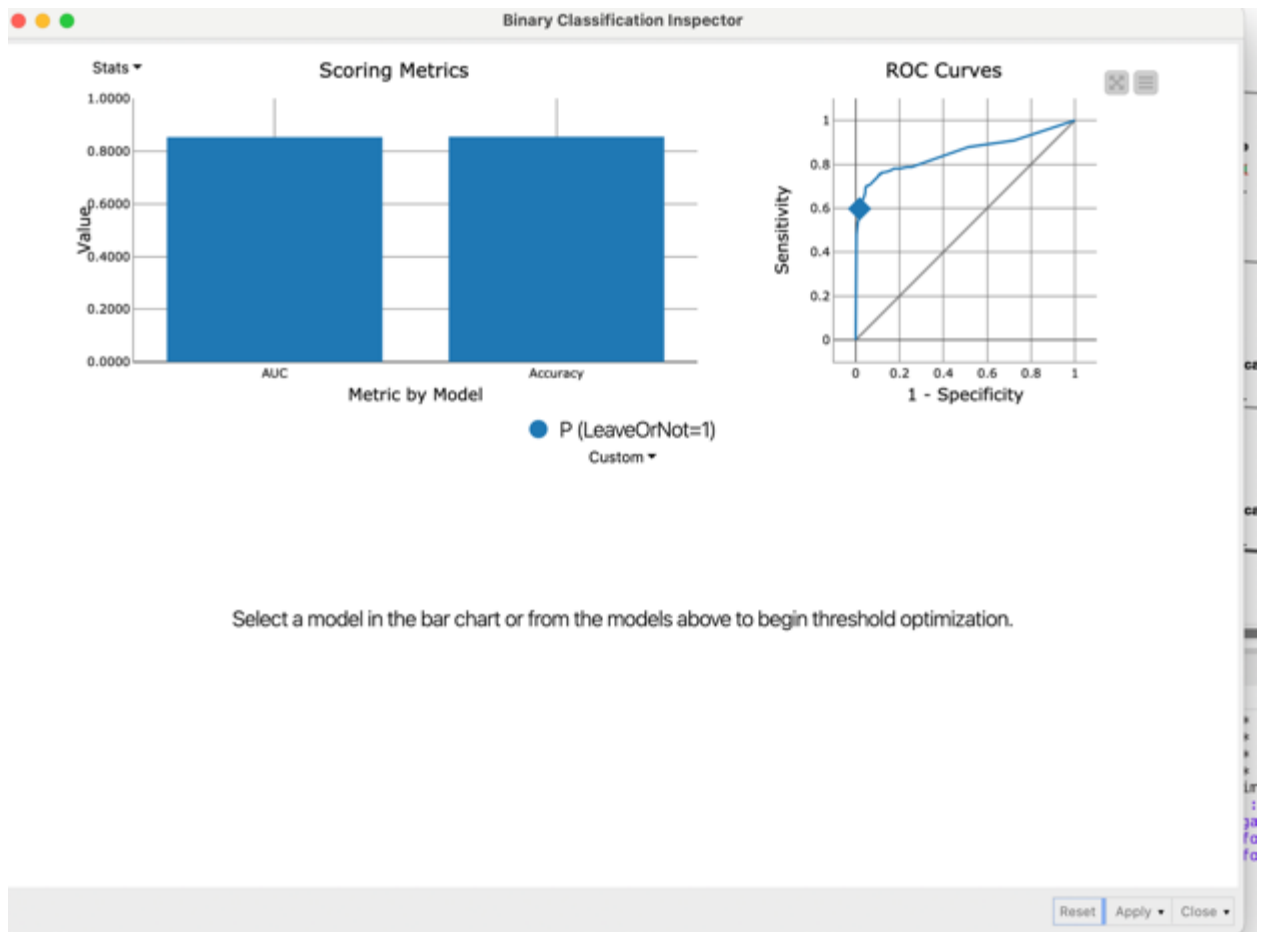
Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	615	10	98.40%
1 (Actual)	124	182	59.48%
	83.22%	94.79%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
85.61%	14.39%	0.640	797	134





The best model

The best model is **Gradient Boosted Tree** with AUC of **0.8906**.