Rapid Miner

Bank Customers Churn – Classification Using Ensemble Model

Outline

01
Business Understanding

02
Data Understanding

03
Data Preparation

04
Modeling

05
Evaluation

06
deployment

CRISP-DM

which stands for Cross-Industry Standard **Process for Data Mining**, is a structured methodology for guiding data mining projects. It consists of six key phases

01 – Business Understanding

As we know, it is much **more expensive to sign** in a new client than **keeping an existing one.**

It is advantageous for banks to know what leads a client towards the decision to leave the company.

Churn prevention allows companies to **develop loyalty programs** and retention campaigns to **keep as many customers** as possible.

02 – Data Understanding

Features

Active member
of bank or not

Duration of usage
of credit card

Row No.	CustomerId	Exited	RowNumber	Surname	CreditScore	Geography	Gender	Age	Tenure
1	15634602	1	1	Hargrave	619	France	Female	42	2
2	15647311	0	2	Hill	608	Spain	Female	41	1

Set id **Set Label**

Credit Score of
Customer

02 – Data Understanding

Features

Balance in credit card

Having credit card

Estimated salary of customer

Satisfaction for complaint resolution

Points Earned

Balance	NumOfProdu...	HasCrCard	IsActiveMem...	EstimatedSal...	Complain	Satisfaction S...	Card Type	Point Earned
0	1	1	1	101348.880	1	2	DIAMOND	464
83807.860	1	0	1	112542.580	1	3	DIAMOND	456

Number of product

Is customer active

Any concern or complaint

Type of card

03 – Data Preparation

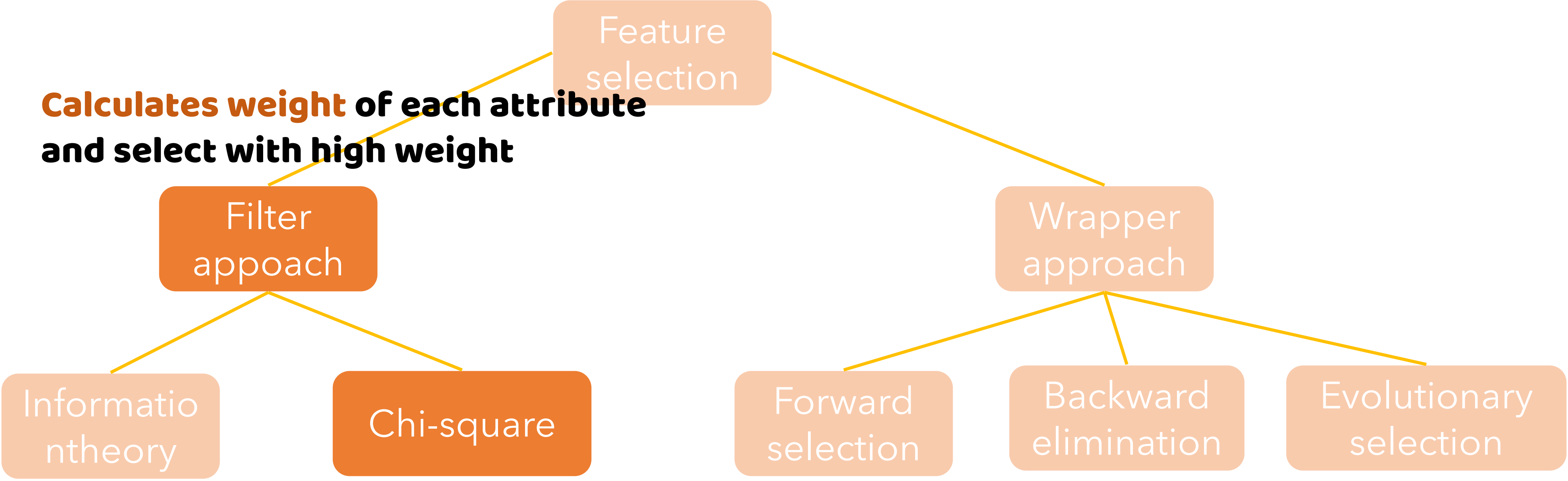
Feature Selection

Feature selection is the process of **selecting subset of features** (or attributes) that **contribute most to the prediction variable** for use in model construction

03 – Data Preparation

Feature Selection

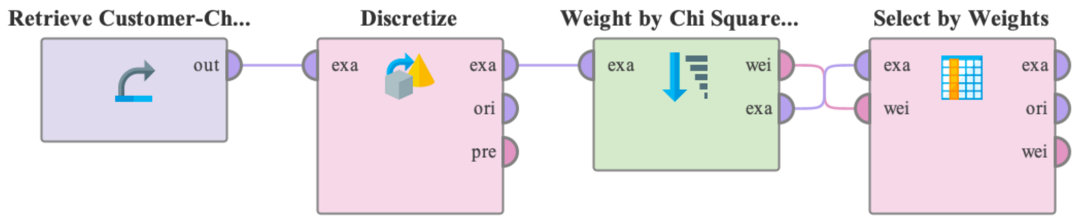
Calculates weight of each attribute and select with high weight



Can only be used with nominal attribute and caculate using the following equation

03 – Data Preparation

Feature Selection



optional

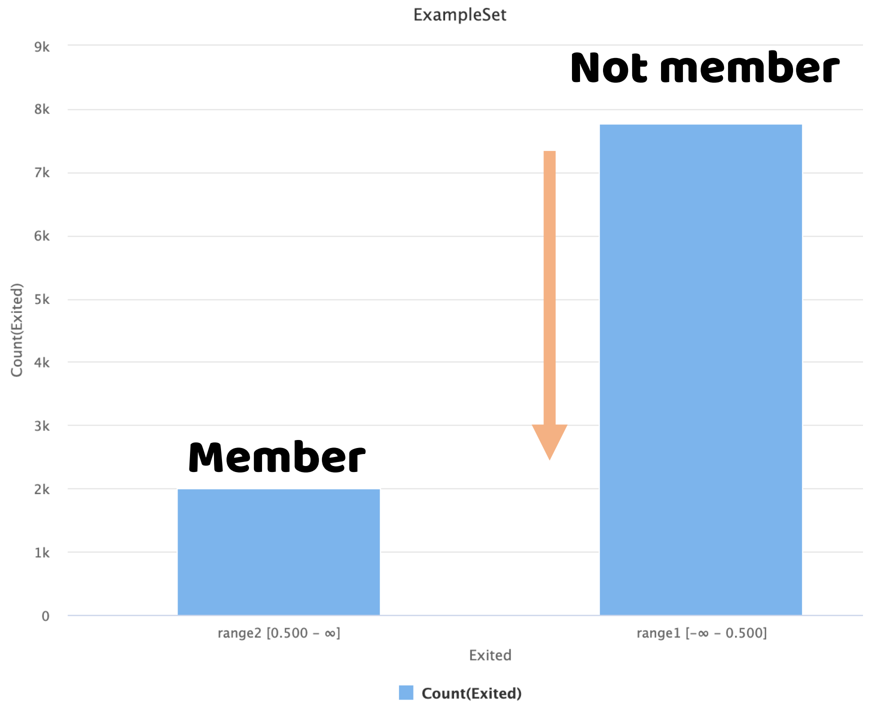
attribute	weight ↓
Complain	9683.357
Surname	2756.877
NumOfPr...	1492.983
Age	1379.422

Row No.	CustomerId	Exited	Surname	Age	NumOfProdu...	Complain
1	15634602	range2 [0.500 - ∞]	Hargrave	42	1	1
2	15647311	range1 [-∞ - 0.500]	Hill	41	1	1
3	15619304	range2 [0.500 - ∞]	Onio	42	3	1
4	15701354	range1 [-∞ - 0.500]	Boni	39	2	0
5	15737888	range1 [-∞ - 0.500]	Mitchell	43	1	0
6	15574012	range2 [0.500 - ∞]	Chu	44	2	1
7	15592531	range1 [-∞ - 0.500]	Bartlett	50	2	0
8	15656148	range2 [0.500 - ∞]	Obinna	29	4	1
9	15792365	range1 [-∞ - 0.500]	He	44	2	0
10	15592389	range1 [-∞ - 0.500]	H?	27	1	0
11	15767821	range1 [-∞ - 0.500]	Bearce	31	2	0

Features include Surname, Age, NumofProducts, Complain

03 – Data Preparation

Imbalance Classification



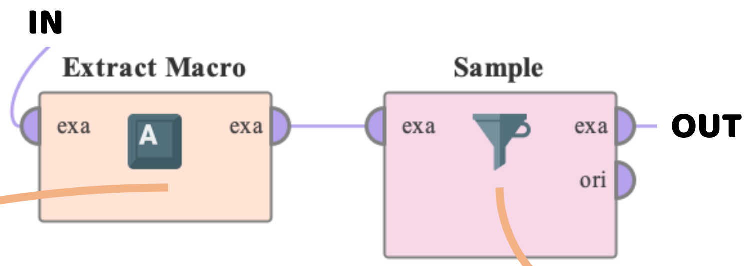
Dealing by **Re-sampling approach** – Fix data

Use **Under sampling** by random majority class to decrease

Bank Customer Churn - Classification

03 – Data Preparation

Imbalance Classification



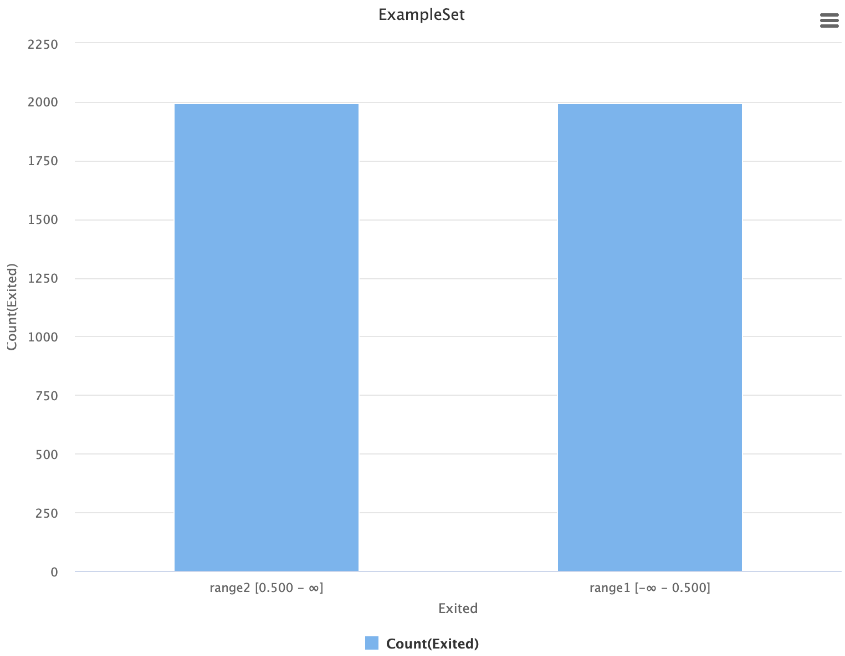
Extract Macro

macro	Excited
macro type	statistics
statistics	count
attribute name	Exited
attribute value	range2 [0.500 - ∞]

class	size
range1 [-∞ - 0.500]	%{Excited}
range2 [0.500 - ∞]	%{Excited}

03 – Data Preparation

Imbalance Classification

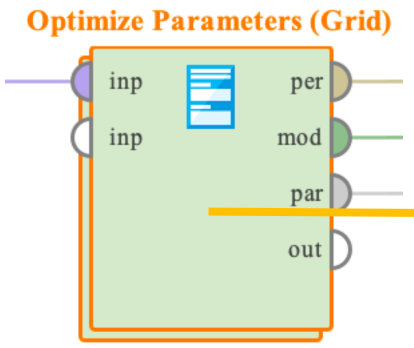


After dealing with Imbalance

Bank Customer Churn - Classification

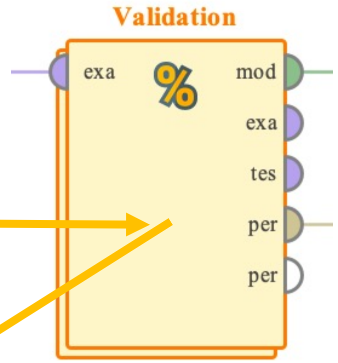
04 – Modeling

Ensemble



1

Optimize for select subprocess



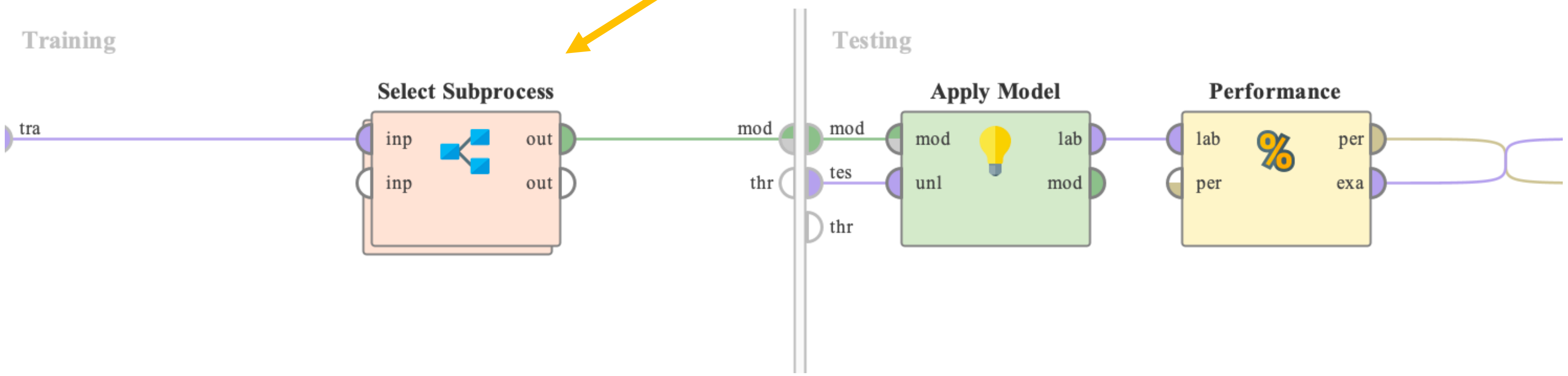
A cross-validation evaluating a decision tree model.

2

Use Cross-validation to split data
Number of fold = 10

3

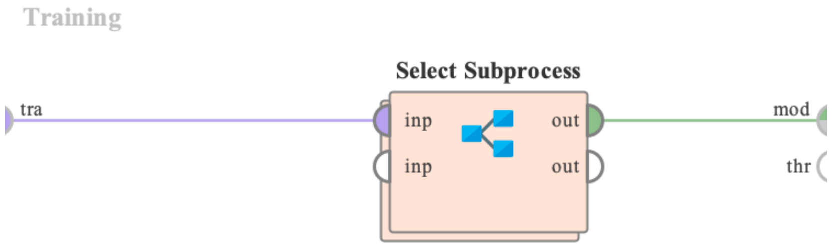
Training with multiple model



Bank Customer Churn - Classification

04 – Modeling

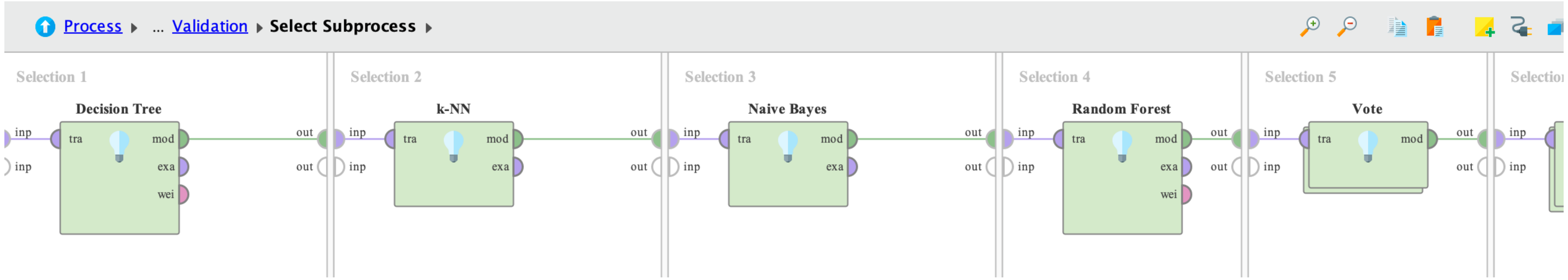
Ensemble



4

In select subprocess we have 8 subprocess including

1. **Decision Tree**
2. **K-NN**
3. **Naïve Bayes**
4. **Random Forest**
5. **Vote**
6. **Bagging**
7. **Adaboost**
8. **Stacking**



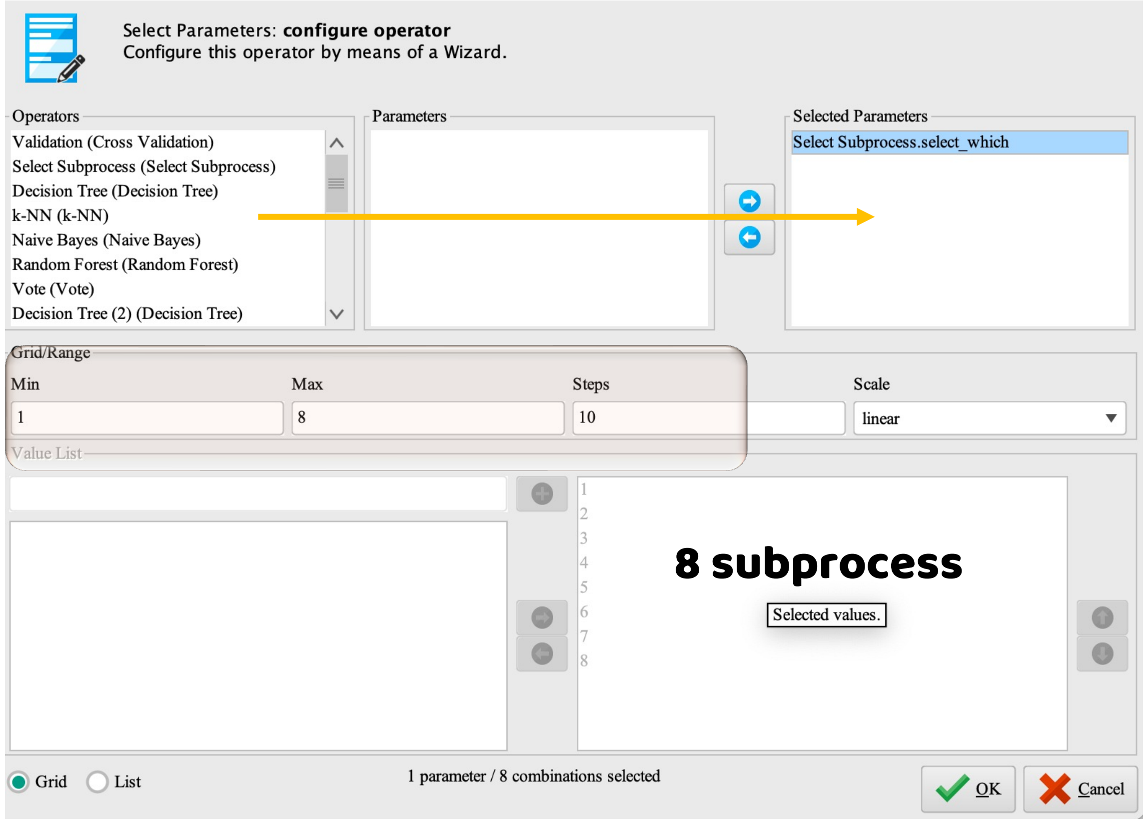
04 – Modeling

Ensemble

5

Optimize parameter by select subprocess

Run Model



Bank Customer Churn - Classification

04 – Modeling

Ensemble

6 **Select by Accuracy**

Optimize Parameters (Grid) (8 rows, 3 columns)

iteration	Select Subprocess.select_which	accuracy ↓
4	4	0.998
5	5	0.998
1	1	0.998
3	3	0.998
6	6	0.998
7	7	0.998
8	8	0.997
2	2	0.991

Random Forest is the most accuracy that program recommended

05 – Evaluation

Random Forest

accuracy: 99.80% +/- 0.20% (micro average: 99.80%)

	true range1 [-∞ - 0.500]	true range2 [0.500 - ∞]	class precision
pred. range1 [-∞ - 0.500]	1990	3	99.85%
pred. range2 [0.500 - ∞]	5	1992	99.75%
class recall	99.75%	99.85%	

precision: 99.75% +/- 0.35% (micro average: 99.75%) (positive class: range2 [0.500 - ∞])

recall: 99.85% +/- 0.24% (micro average: 99.85%) (positive class: range2 [0.500 - ∞])

Accuracy = 99.8 %
Precision = 99.75 %
Recall = 99.85 %

F1- score = 99.79 %

that mean the model performs well

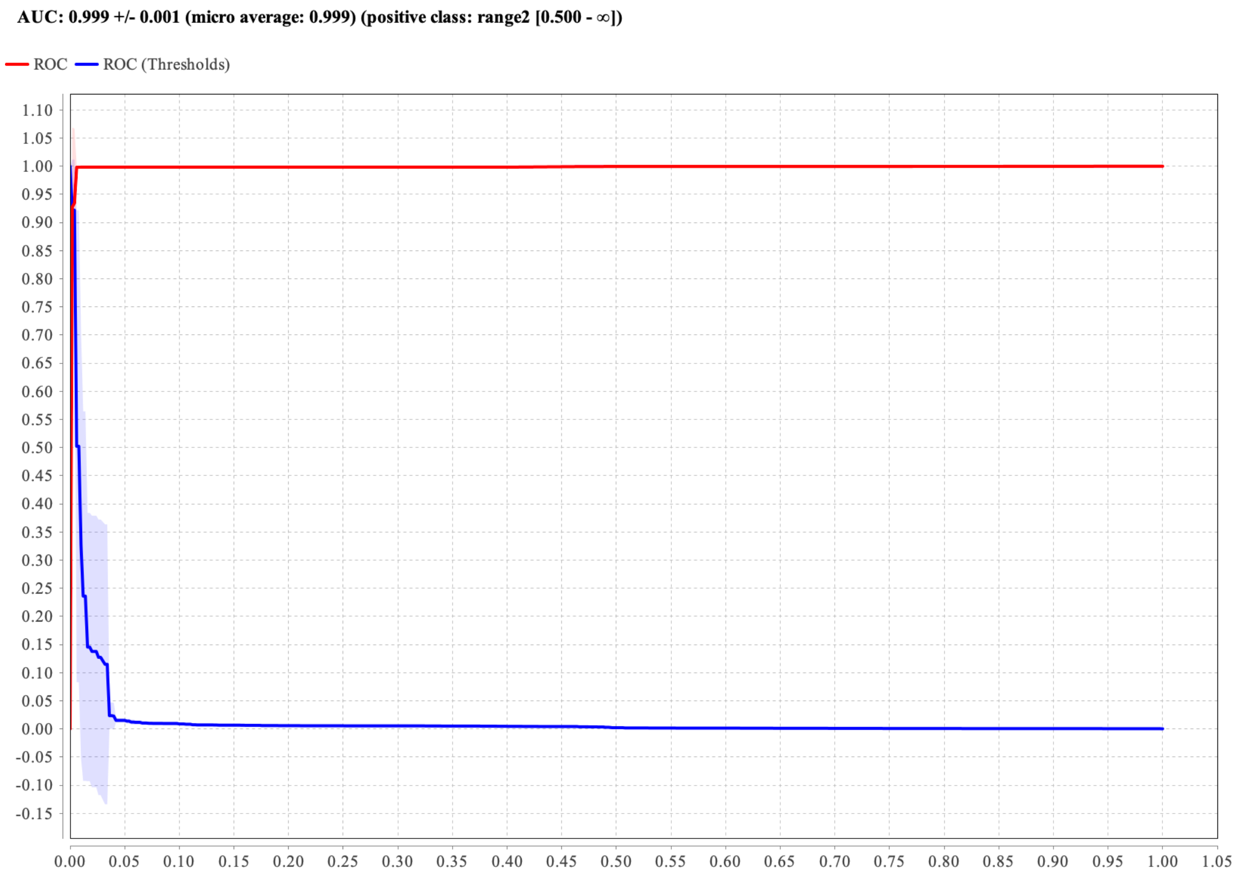
Bank Customer Churn - Classification

05 – Evaluation

Random Forest

The AUC value is between 0-1

That mean the model can classify accurately



06 – Deployment

After saving the model, we can **connect it with new data** to **predict customer churn** (tendency to leave the bank). This will allow us to **identify trends and build strategies to retain** these customers.