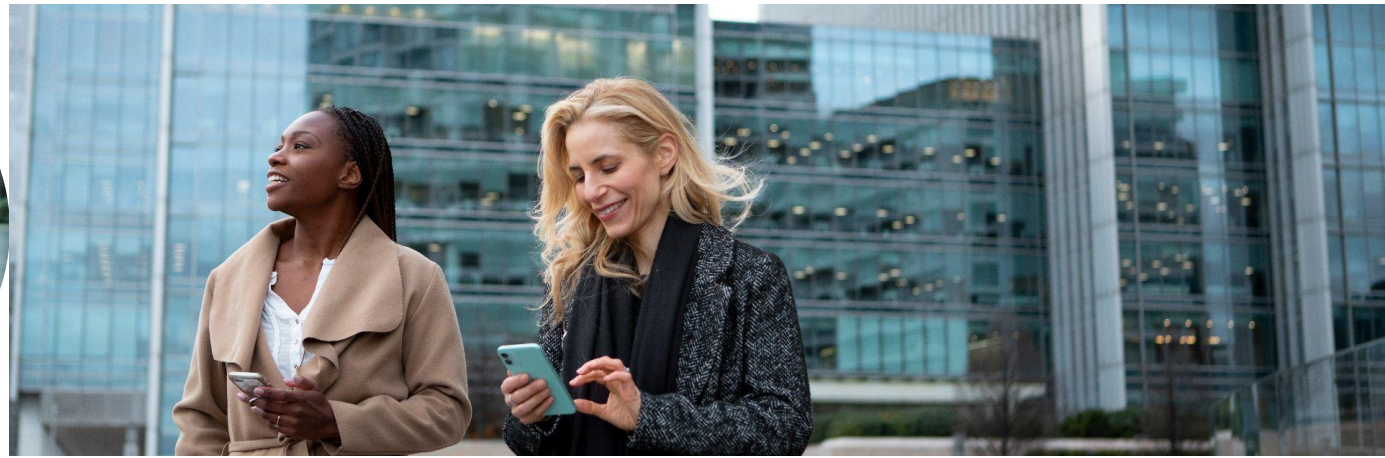
01 Data Understanding



Dataset (Raw Data)

Database : hotel_app

Table : sales_data

Row : 100,000

Attribute : 8

Row No.	store_id	customer_id	product_id	product_ca...	date	amount	single_price	transacti
1	Store 01	Customer 1508	53642	Toys	Apr 1, 2007	3	90.246	1
2	Store 15	Customer 169	90945	Movies	Feb 15, 2005	2	60.586	2
3	Store 12	Customer 124	18548	Movies	Sep 26, 2007	5	96.613	3
4	Store 05	Customer 1988	85359	Books	May 7, 2005	5	16.963	4
5	Store 01	Customer 475	80069	Clothing	Jan 6, 2008	5	65.215	5
6	Store 11	Customer 761	55848	Sports	Jun 3, 2006	3	56.475	6
7	Store 10	Customer 741	11762	Health	Sep 19, 2006	3	26.873	7

ExampleSet (100,000 examples,0 special attributes,8 regular attributes)

Condition

สมมุติว่านักศึกษาอยู่ในทีม Data Science ที่จะต้องไปทำ Proof Of Concept (POC) ให้กับบริษัทแห่งหนึ่งที่ขายสินค้าแบบออนไลน์ (online) ซึ่งบริษัทนี้ต้องการทราบว่าถ้าจะแนะนำสินค้าประเภท Electronics ควรจะแนะนำให้กับใครบ้างถึงจะมีโอกาสซื้อสินค้านั้น โดยมีแนวทางการทำงานดังนี้

1. จากข้อมูลในตาราง sales_data ในฐานข้อมูล นักศึกษาต้องสร้าง training data โดยพิจารณาเป็นรายข้อมูลแถวให้เป็น customer แต่ละคน และคอลัมน์มีตามเงื่อนไขดังนี้

1.1 ข้อมูลการซื้อสินค้าที่เป็นประเภท Electronics ที่เป็น label คำตอบให้ใช้ของเดือน November 2008 เท่านั้น (1-30 November 2008)

1.2 สำหรับแอตทริบิวต์ทั่วไปใช้ข้อมูลการซื้อสินค้าประเภทต่างๆ ของแต่ละ customer ย้อนหลังไปเป็นจำนวน 6 เดือน (May 2008 - October 2008)

1.3 เชื่อมโยงข้อมูลในส่วนที่ (1) และส่วนที่ (2) เข้าด้วยกันโดยใช้ customer_id และสร้างแอตทริบิวต์ Response ขึ้นมาเป็นลาเบลโดยที่ ถ้าลูกค้าคนใดไม่มีการซื้อสินค้าประเภท Electronics ในเดือน November 2008 จะให้เป็นค่า "NO" ถ้ามีการซื้อจะให้เป็นค่า "YES"

Dataset (Label Data)

Database : hotel_app

Table : sales_data

Row : 274

Attribute : 8

Row No.	store_id	customer_id	product_id	product_ca...	date	amount	single_price	transacti
1	Store 05	Customer 741	22633	Electronics	Nov 13, 2008	4	35.021	58
2	Store 02	Customer 1304	75848	Electronics	Nov 1, 2008	8	47.909	414
3	Store 14	Customer 1690	45170	Electronics	Nov 28, 2008	8	13.623	1305
4	Store 09	Customer 1006	90389	Electronics	Nov 25, 2008	3	30.670	1424
5	Store 06	Customer 456	84849	Electronics	Nov 28, 2008	4	25.481	1444
6	Store 11	Customer 1277	83022	Electronics	Nov 11, 2008	8	87.499	2183
7	Store 13	Customer 1997	81021	Electronics	Nov 22, 2008	3	83.981	3378

ExampleSet (274 examples,0 special attributes,8 regular attributes)

Dataset (History Data)

Database : hotel_app

Table : sales_data

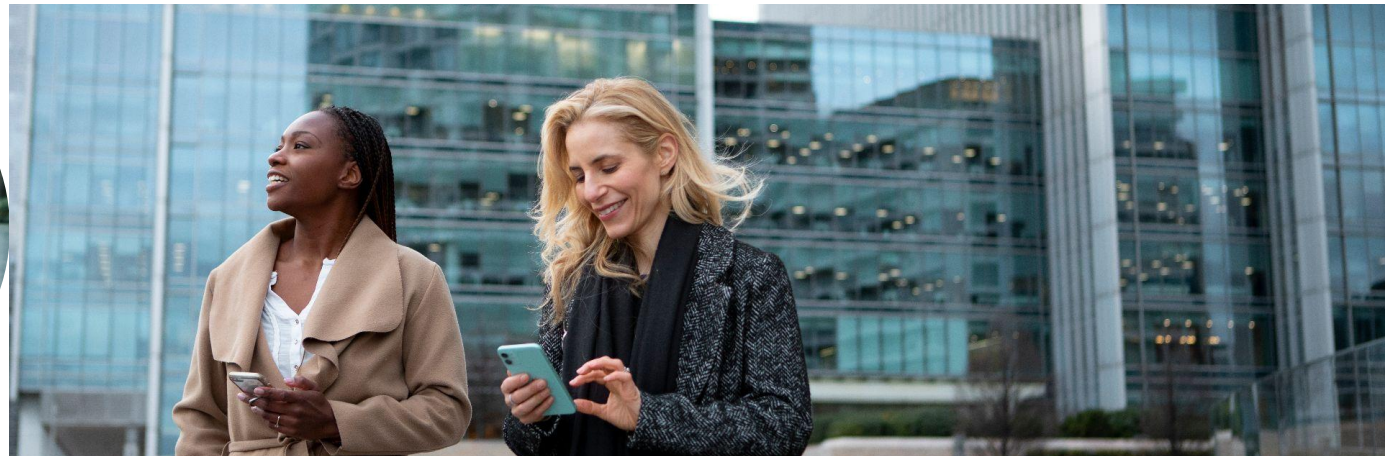
Row : 13,253

Attribute : 8

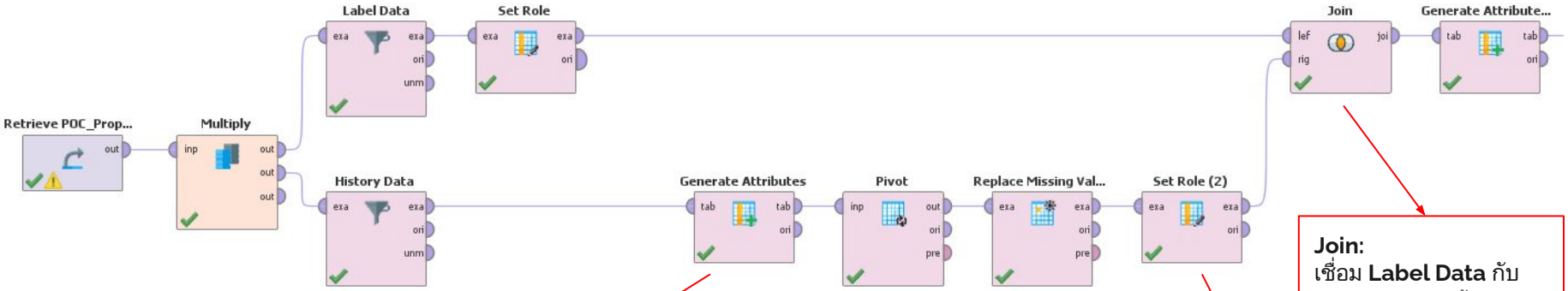
Row No.	store_id	customer_id	product_id	product_ca...	date	amount	single_price	transacti
1	Store 14	Customer 1572	58636	Health	Jun 5, 2008	5	81.849	11
2	Store 08	Customer 1384	20905	Electronics	May 25, 2008	8	89.178	13
3	Store 05	Customer 1832	37283	Books	Jun 17, 2008	1	18.647	41
4	Store 08	Customer 1609	11433	Toys	Jun 10, 2008	9	68.673	49
5	Store 10	Customer 553	78007	Electronics	May 26, 2008	6	12.772	50
6	Store 02	Customer 217	36959	Movies	Jul 17, 2008	6	67.188	53
7	Store 08	Customer 814	70502	Home/Garden	Sep 6, 2008	4	56.285	57

ExampleSet (13,253 examples,0 special attributes,8 regular attributes)

02 Data Preparation



Process



Set Role : customer_id

Response:
if(product_category=="Electronics", "YES", "NO")

Join:
เชื่อม **Label Data** กับ **History Data** ด้วย customer_id

Generate Attributes :
total_price = amount*single_price

Pivot :
Group by = customer_id
Aggregation
- total_price (average)
- transaction_id (count)

Replace Missing Value ใน customer_id ให้เป็น 0

Set Role : customer_id

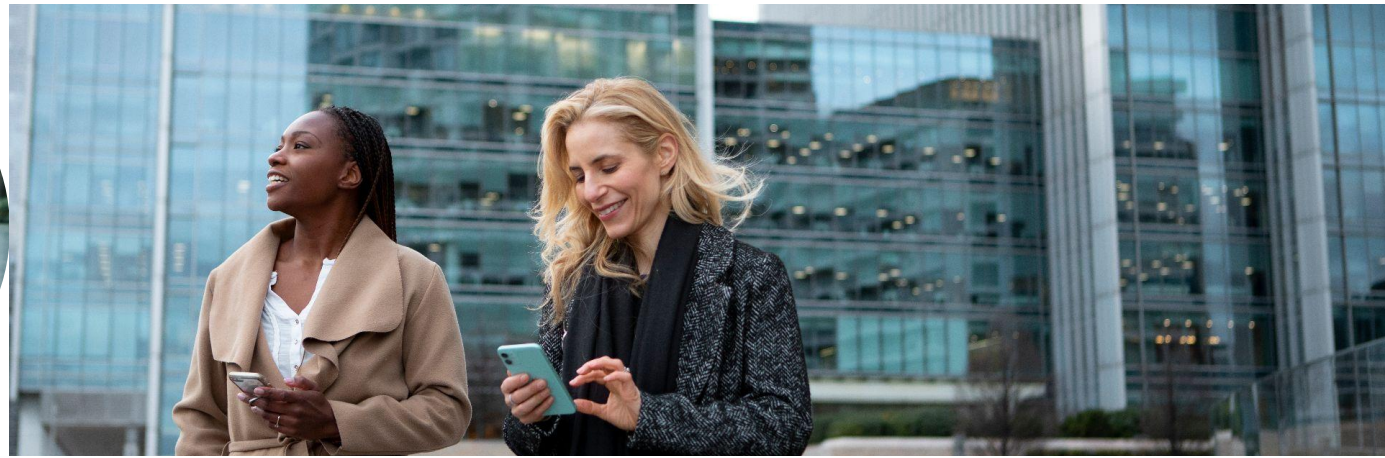
Output

- store_id
- customer_id
- product_id
- product_category
- date
- amount
- single_price
- transaction_id
- average(total_price)_Books
- average(total_price)_Clothing
- average(total_price)_Electronics
- average(total_price)_Health
- average(total_price)_Home/Garden
- average(total_price)_Movies
- average(total_price)_Sports
- average(total_price)_Toys
- count(transaction_id)_Books
- count(transaction_id)_Clothing
- count(transaction_id)_Electronics
- count(transaction_id)_Health
- count(transaction_id)_Home/Garden
- count(transaction_id)_Movies
- count(transaction_id)_Sports
- count(transaction_id)_Toys
- Response

Row No.	customer_id	store_id	product_id	product_ca...	date	amount	single_price	transaction...	average(to...	average(to...
1	Customer 1572	?	?	?	?	?	?	?	280.032	639.294
2	Customer 1384	?	?	?	?	?	?	?	94.217	499.063
3	Customer 1832	Store 02	15183	Electronics	Nov 3, 2008	5	64.455	90393	279.351	460.370
4	Customer 1609	?	?	?	?	?	?	?	73.897	0
5	Customer 553	?	?	?	?	?	?	?	0	0
6	Customer 217	Store 06	82758	Electronics	Nov 28, 2008	9	20.964	40691	0	240.343
7	Customer 814	?	?	?	?	?	?	?	0	0
8	Customer 1387	Store 15	24625	Electronics	Nov 22, 2008	3	42.892	57358	247.882	66.653
9	Customer 877	?	?	?	?	?	?	?	0	0
10	Customer 392	?	?	?	?	?	?	?	103.622	289.451
11	Customer 123	?	?	?	?	?	?	?	102.325	0
12	Customer 1553	?	?	?	?	?	?	?	0	606.369
13	Customer 1914	Store 12	51069	Electronics	Nov 20, 2008	2	64.565	61531	0	0

ExampleSet (2,012 examples,1 special attribute,24 regular attributes)

03 Feature Engineering



Process

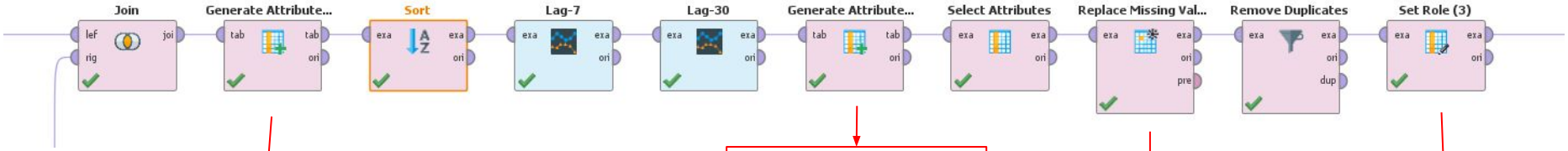
Ascending:
- customer_id
- date

date: lag = 7
Anyone buy Electronics last 7 time periods or not

date: lag = 30
Anyone buy Electronics last 30 time periods or not

Exclude attributes: ['amount', 'date', 'date-7', 'date-30', 'product_category', 'product_id', 'single_price', 'store_id', 'transaction_id']

Remove duplicate: all



Response:
if(product_category=="Electronics", "YES", "NO")

lasted-7: number of days between [date-7] and 2008-11-01

lasted-30: number of days between [date-30] and 2008-11-01

Replace Missing Value with zero (-infinity date):
['date-7', 'date-30', 'lasted-7', 'lasted-30']

Response: label

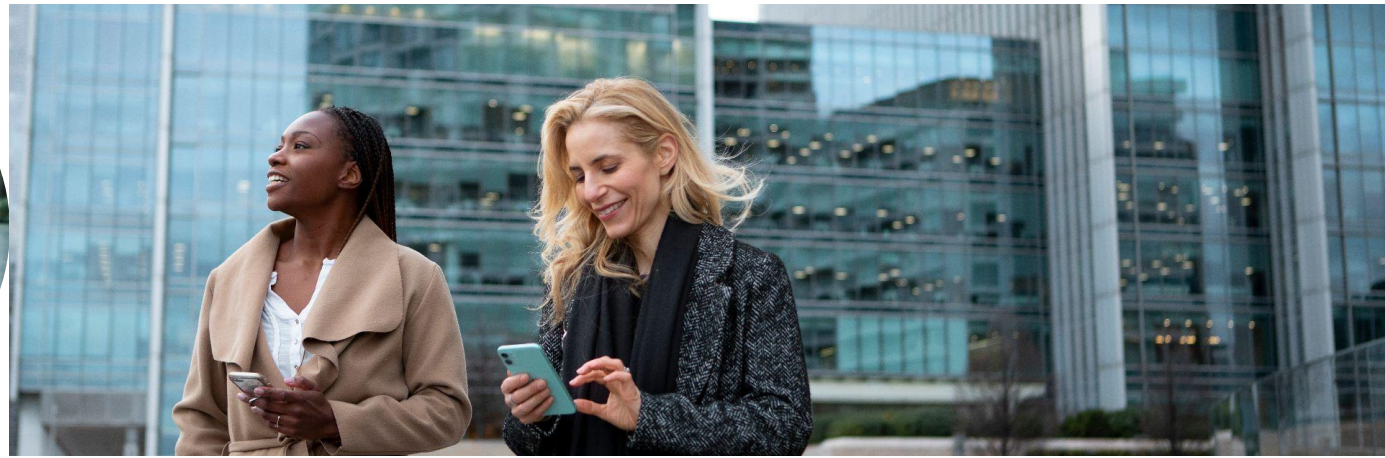
Output

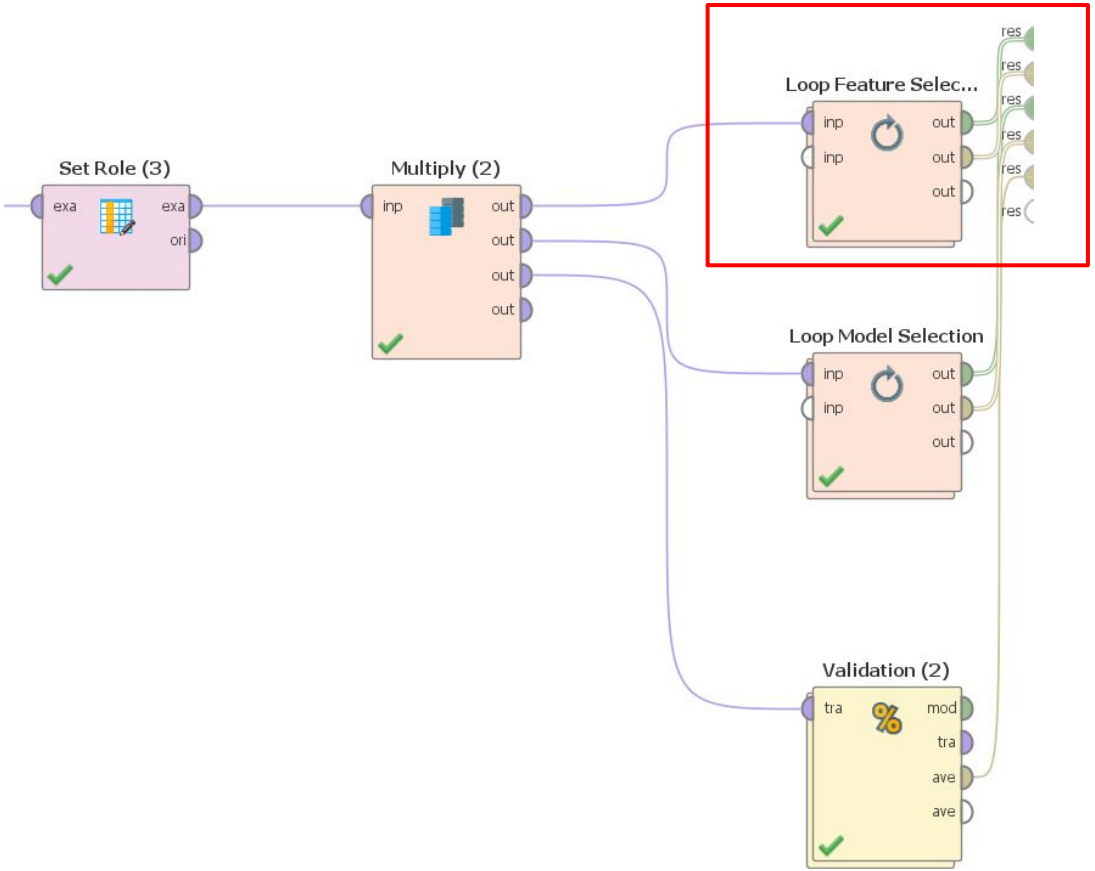
lasted-7
lasted-30
customer_id
average(total_price)_Books
average(total_price)_Clothing
average(total_price)_Electronics
average(total_price)_Health
average(total_price)_Home/Garden
average(total_price)_Movies
average(total_price)_Sports
average(total_price)_Toys
count(transaction_id)_Books
count(transaction_id)_Clothing
count(transaction_id)_Electronics
count(transaction_id)_Health
count(transaction_id)_Home/Garden
count(transaction_id)_Movies
count(transaction_id)_Sports
count(transaction_id)_Toys
Response

Row No.	customer_id	Response	lasted-7	lasted-30	average(to...	average(to...	average(to...	average(to...	average(to...	average(to...
1	Customer 1	NO	0	0	244.606	0	56.317	324.944	0	94.177
2	Customer 10	NO	0	0	154.374	0	0	172.149	0	242.558
3	Customer 100	YES	0	0	282.802	63.964	240.990	160.413	0	0
4	Customer 1000	NO	0	0	302.463	219.856	82.762	74.109	0	0
5	Customer 1001	NO	0	0	0	535.965	0	0	592.319	169.956
6	Customer 1002	NO	0	0	0	0	278.806	0	82.466	425.199
7	Customer 1003	YES	0	0	218.603	0	0	0	100.696	98.401
8	Customer 1004	NO	0	0	0	689.251	325.127	0	0	0
9	Customer 1005	NO	0	0	168.362	0	262.858	452.096	128.882	0
10	Customer 1006	YES	21	0	360.743	193.497	0	0	63.057	0
11	Customer 1007	NO	0	0	0	0	255.601	321.823	0	0
12	Customer 1008	NO	0	0	0	623.192	285.564	485.816	0	0
13	Customer 1009	NO	0	0	219.270	0	443.307	91.129	203.874	523.457

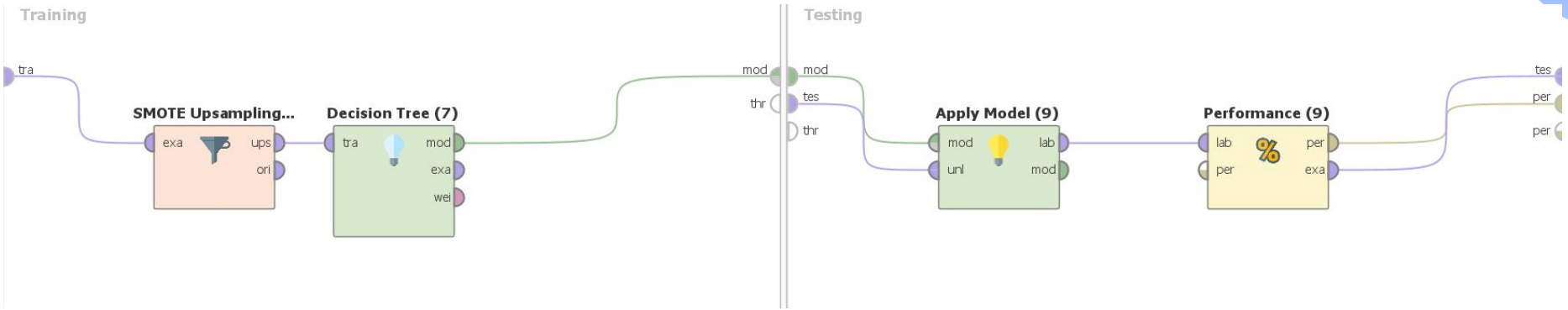
ExampleSet (2,006 examples,2 special attributes,18 regular attributes)

04 Feature Selection




Set Role (3)
exa
exa
ori
Multiply (2)
inp
out
out
out
out
Loop Feature Selec...
inp
inp
out
out
out
res
res
res
res
res
res
Loop Model Selection
inp
inp
out
out
out
Validation (2)
tra
mod
tra
ave
ave

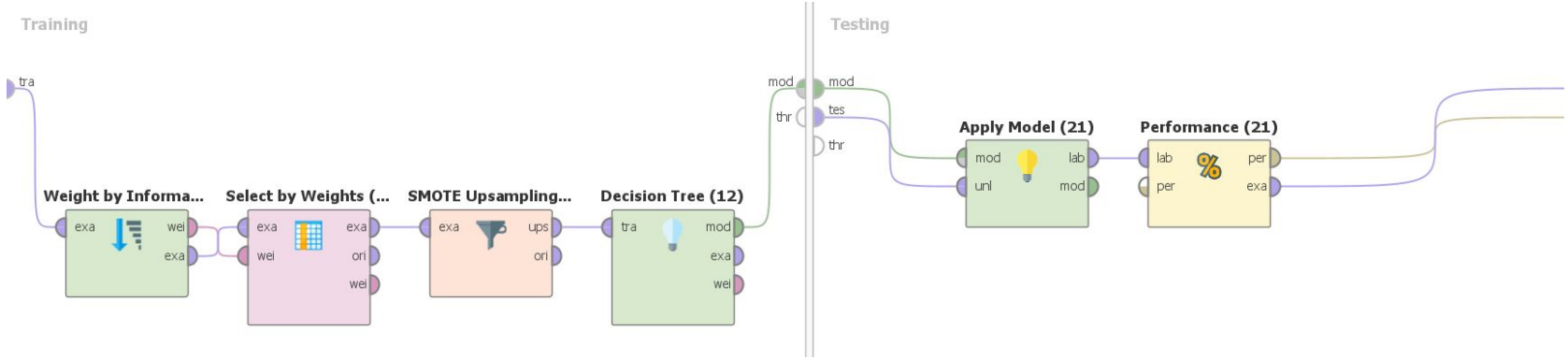
Feature Selection



accuracy: 18.94% +/- 1.77% (micro average: 18.94%)

	true NO	true YES	class precision
pred. NO	121	9	93.08%
pred. YES	1617	259	13.81%
class recall	6.96%	96.64%	

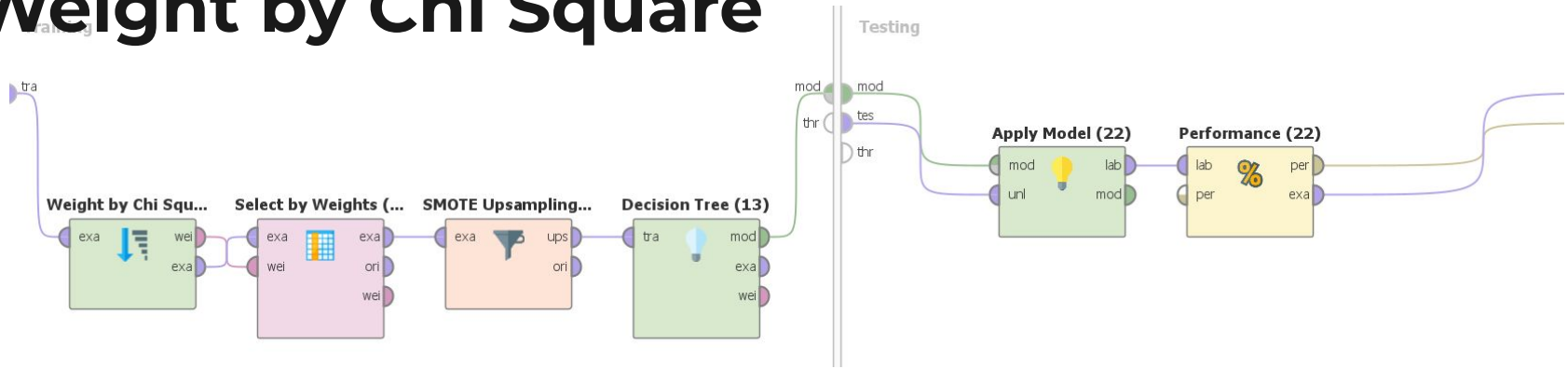
Weight by Information Gain



accuracy: 18.94% +/- 1.77% (micro average: 18.94%)

	true NO	true YES	class precision
pred. NO	121	9	93.08%
pred. YES	1617	259	13.81%
class recall	6.96%	96.64%	

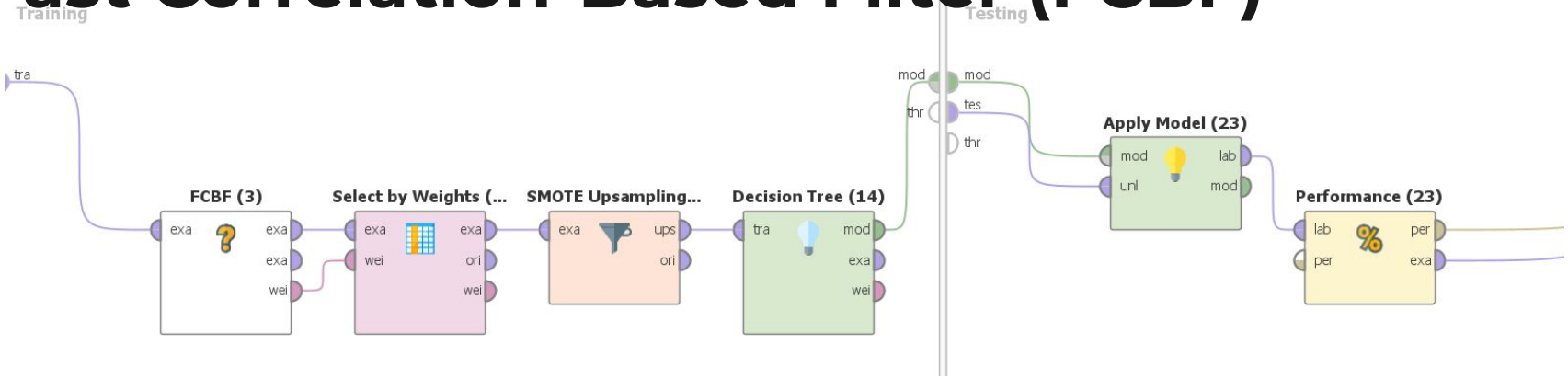
Weight by Chi Square



accuracy: 18.94% +/- 1.77% (micro average: 18.94%)

	true NO	true YES	class precision
pred. NO	121	9	93.08%
pred. YES	1617	259	13.81%
class recall	6.96%	96.64%	

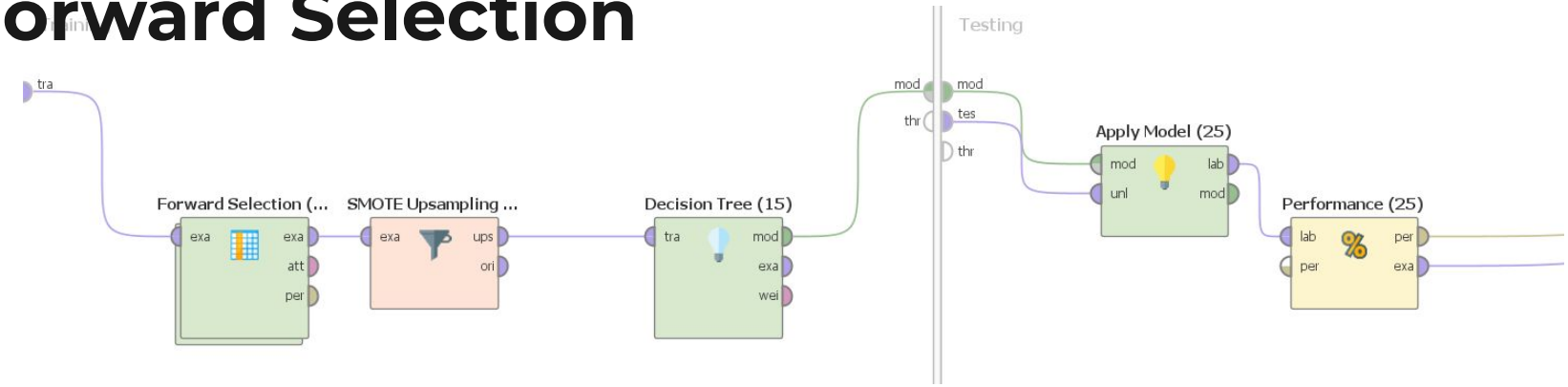
Fast Correlation-Based Filter (FCBF)



accuracy: 18.94% +/- 1.77% (micro average: 18.94%)

	true NO	true YES	class precision
pred. NO	121	9	93.08%
pred. YES	1617	259	13.81%
class recall	6.96%	96.64%	

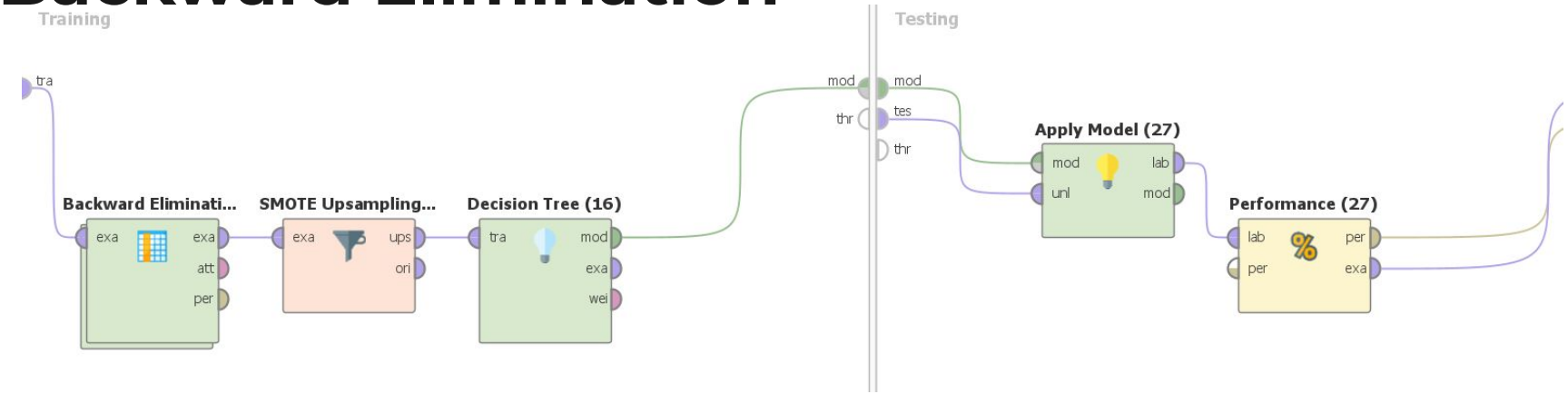
Forward Selection



accuracy: 84.05% +/- 1.61% (micro average: 84.05%)

	true NO	true YES	class precision
pred. NO	1675	257	86.70%
pred. YES	63	11	14.86%
class recall	96.38%	4.10%	

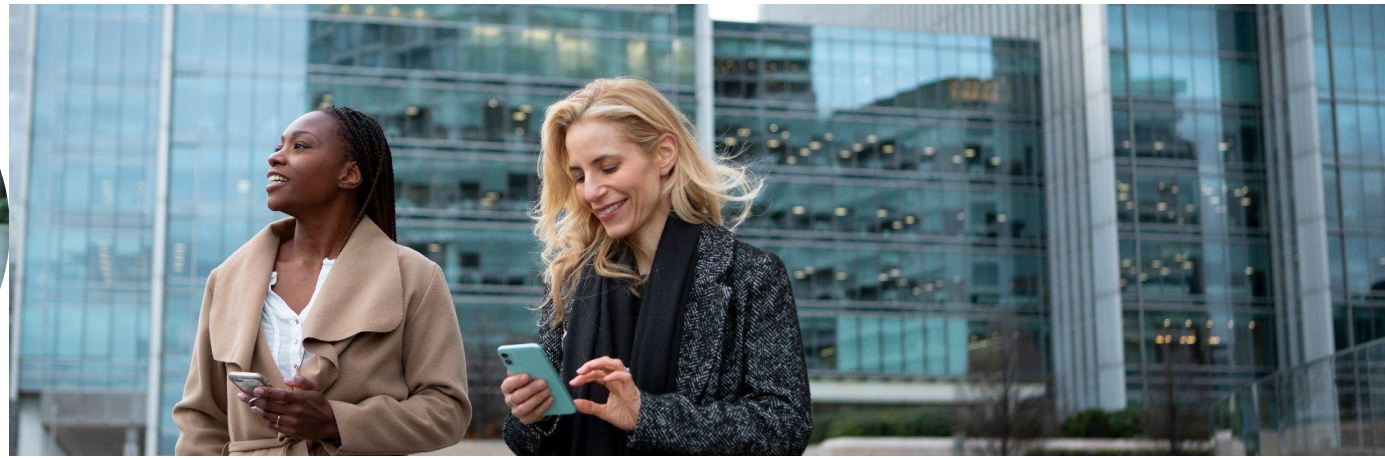
Backward Elimination

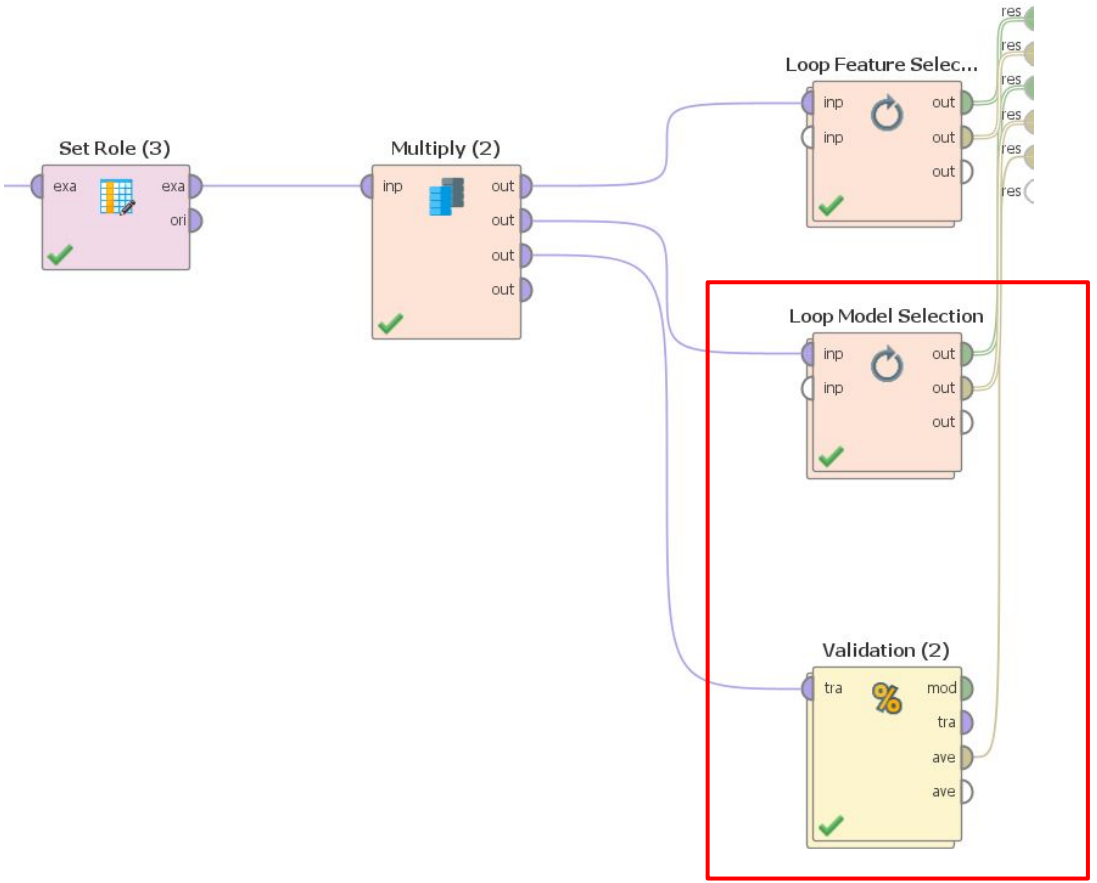


accuracy: 18.59% +/- 1.74% (micro average: 18.59%)

	true NO	true YES	class precision
pred. NO	116	11	91.34%
pred. YES	1622	257	13.68%
class recall	6.67%	95.90%	

05 Model




Set Role (3)
exa
exa
ori
Multiply (2)
inp
out
out
out
out
Loop Feature Selec...
inp
inp
out
out
out
res
res
res
res
res
res
Loop Model Selection
inp
inp
out
out
out
Validation (2)
tra
mod
tra
ave
ave

Imbalance Data

Label variable

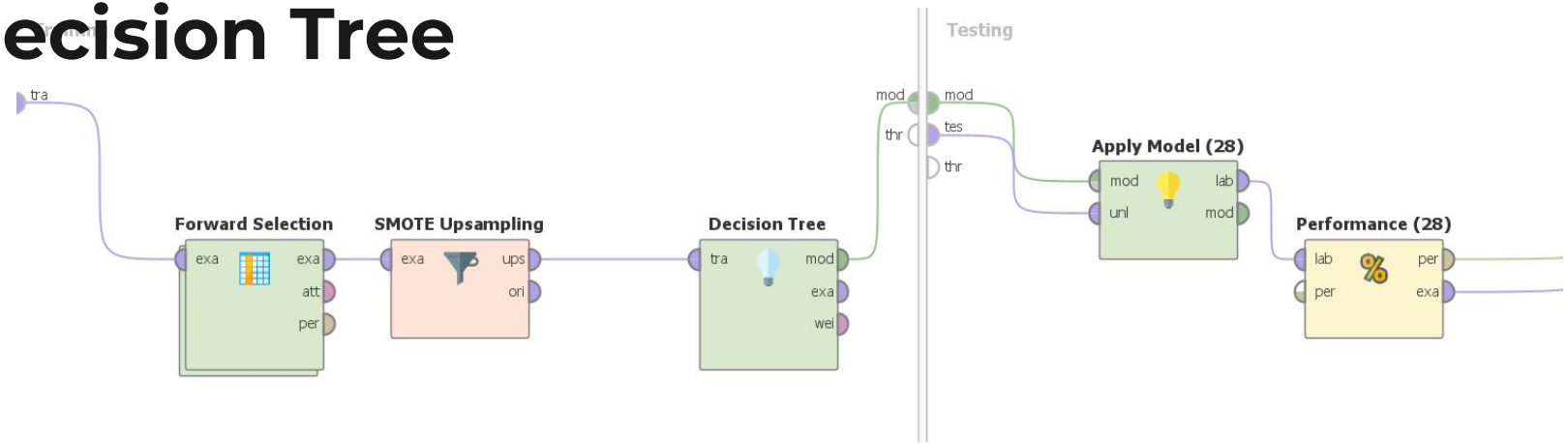
		Least	Most	Values
Nominal	0	YES (268)	NO (1738)	NO (1738), YES (268)

Baseline : Random Guessing Baseline
= p^2 + (1-p)^2

All = 1738 + 268 = 2,006
No = 1738/2006 = 0.8664
Yes = 268/2006 = 0.1336

Baseline = (0.8664)^2 + (0.1336)^2 = 0.8035

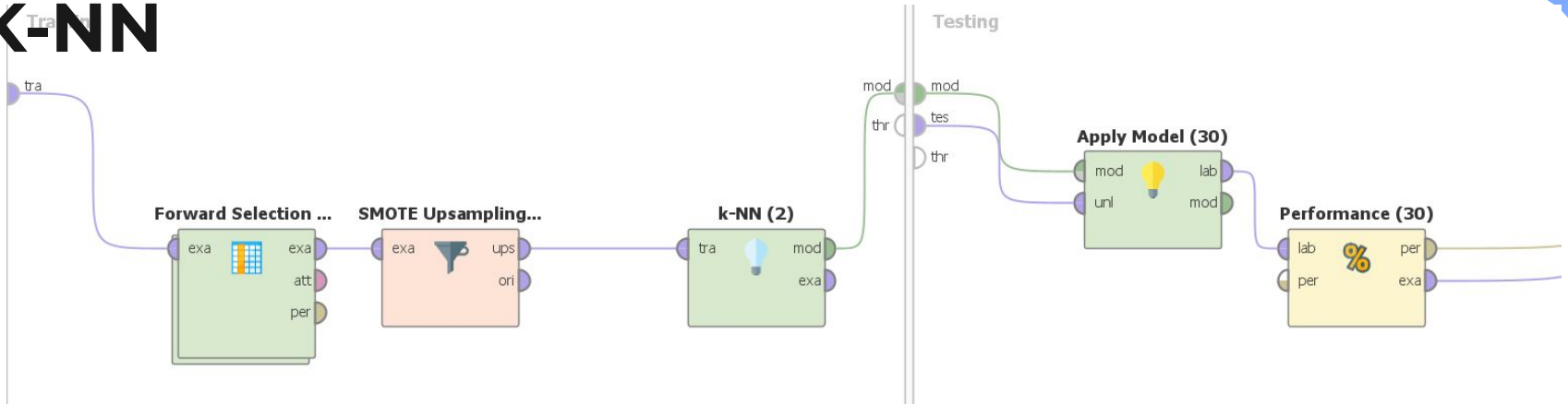
Decision Tree



accuracy: 84.05% +/- 1.61% (micro average: 84.05%)

	true NO	true YES	class precision
pred. NO	1675	257	86.70%
pred. YES	63	11	14.86%
class recall	96.38%	4.10%	

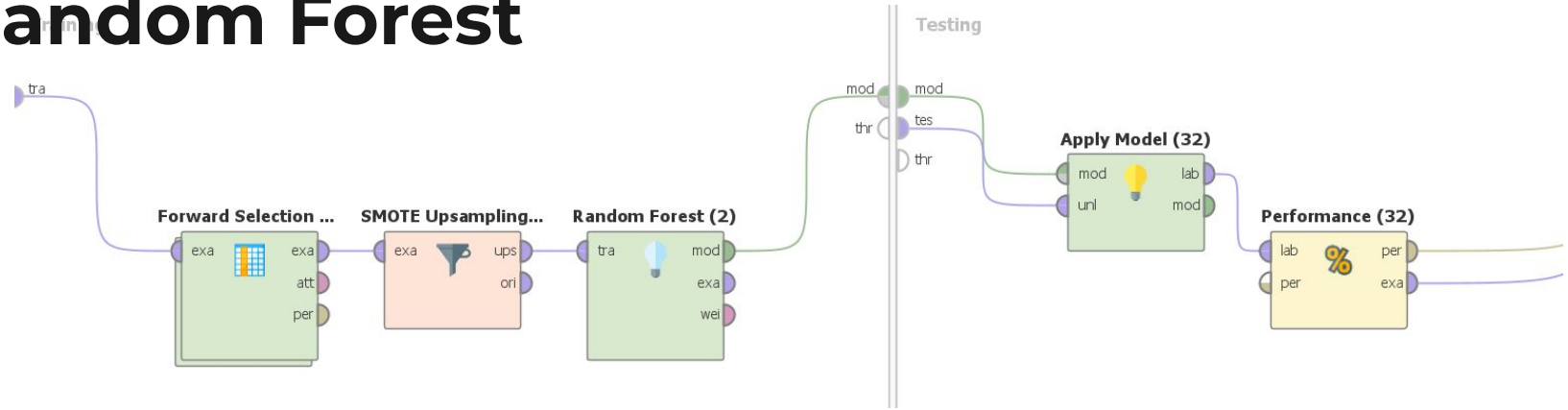
K-NN



accuracy: 86.19% +/- 0.57% (micro average: 86.19%)

	true NO	true YES	class precision
pred. NO	1728	267	86.62%
pred. YES	10	1	9.09%
class recall	99.42%	0.37%	

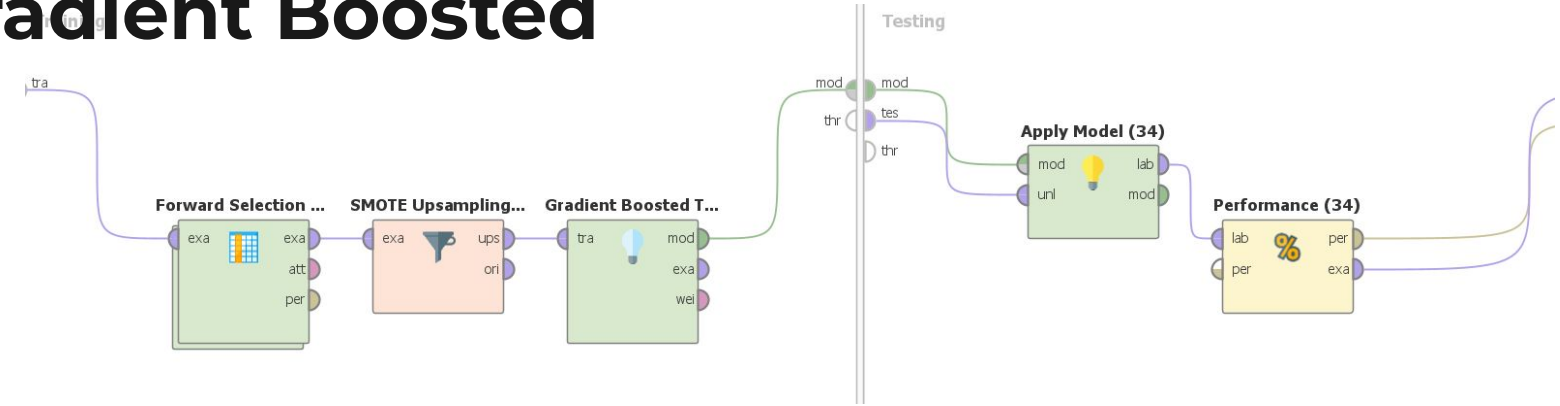
Random Forest



accuracy: 83.70% +/- 1.34% (micro average: 83.70%)

	true NO	true YES	class precision
pred. NO	1663	252	86.84%
pred. YES	75	16	17.58%
class recall	95.68%	5.97%	

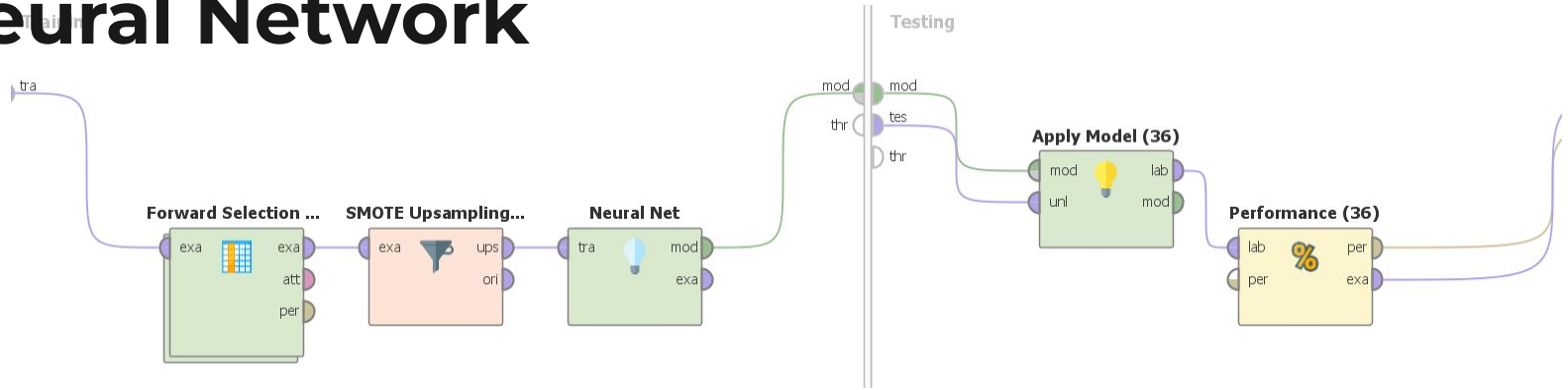
Gradient Boosted



accuracy: 79.86% +/- 3.01% (micro average: 79.86%)

	true NO	true YES	class precision
pred. NO	1581	247	86.49%
pred. YES	157	21	11.80%
class recall	90.97%	7.84%	

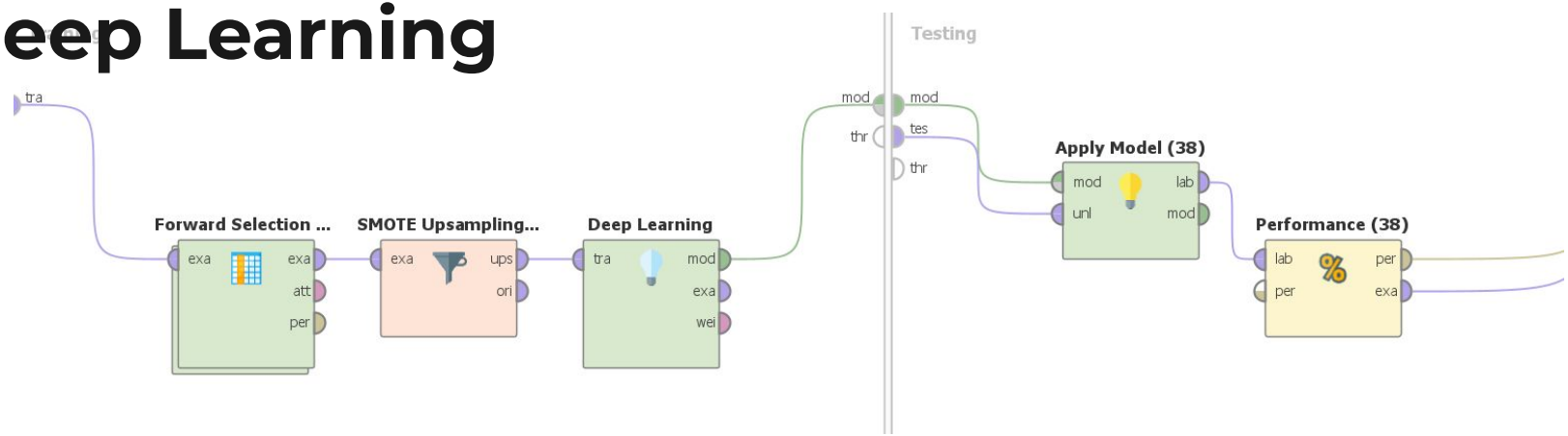
Neural Network



accuracy: 55.55% +/- 36.11% (micro average: 55.53%)

	true NO	true YES	class precision
pred. NO	998	152	86.78%
pred. YES	740	116	13.55%
class recall	57.42%	43.28%	

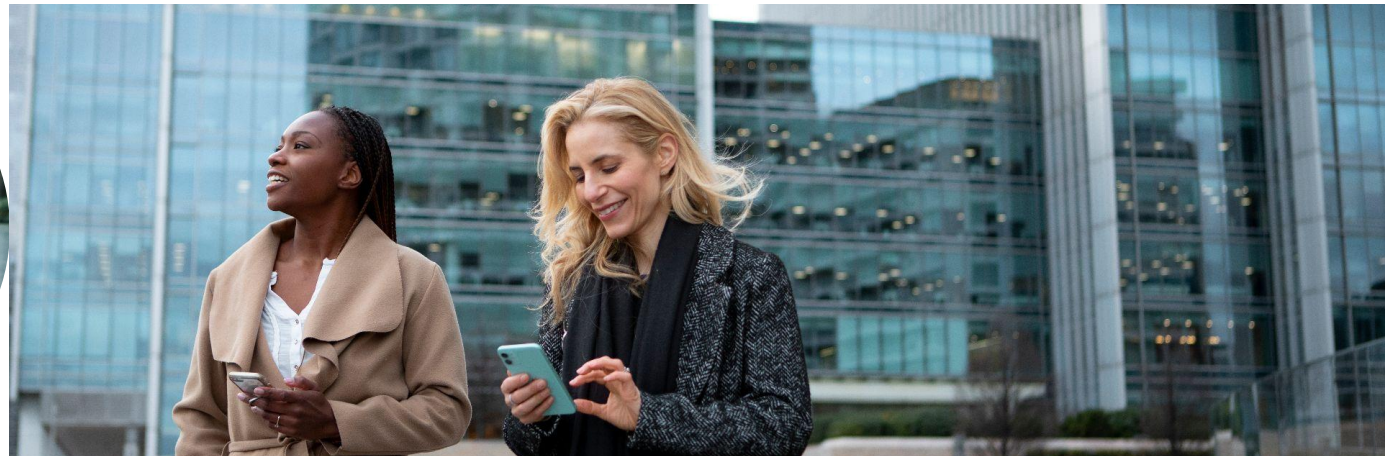
Deep Learning



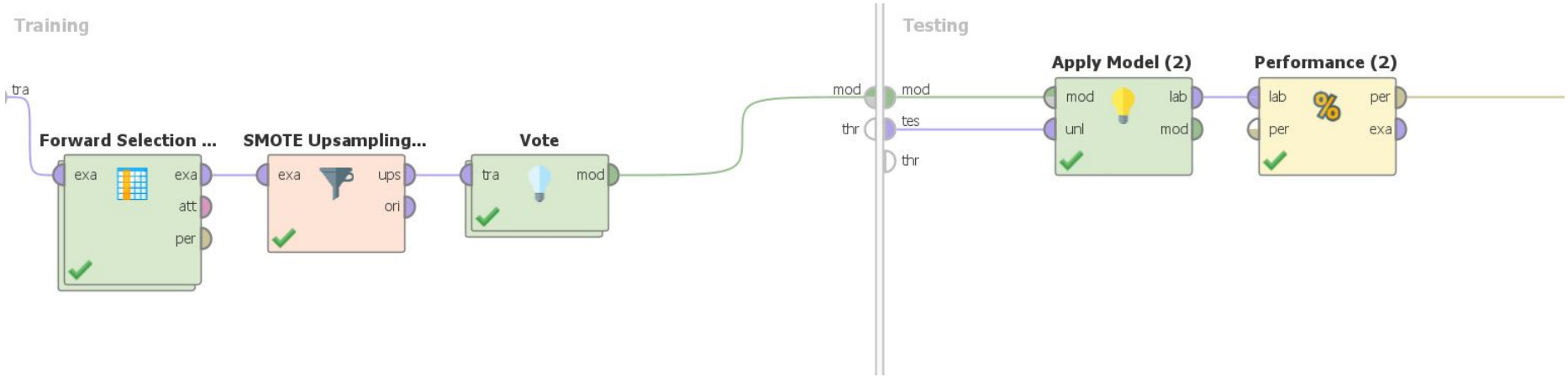
accuracy: 80.71% +/- 3.57% (micro average: 80.71%)

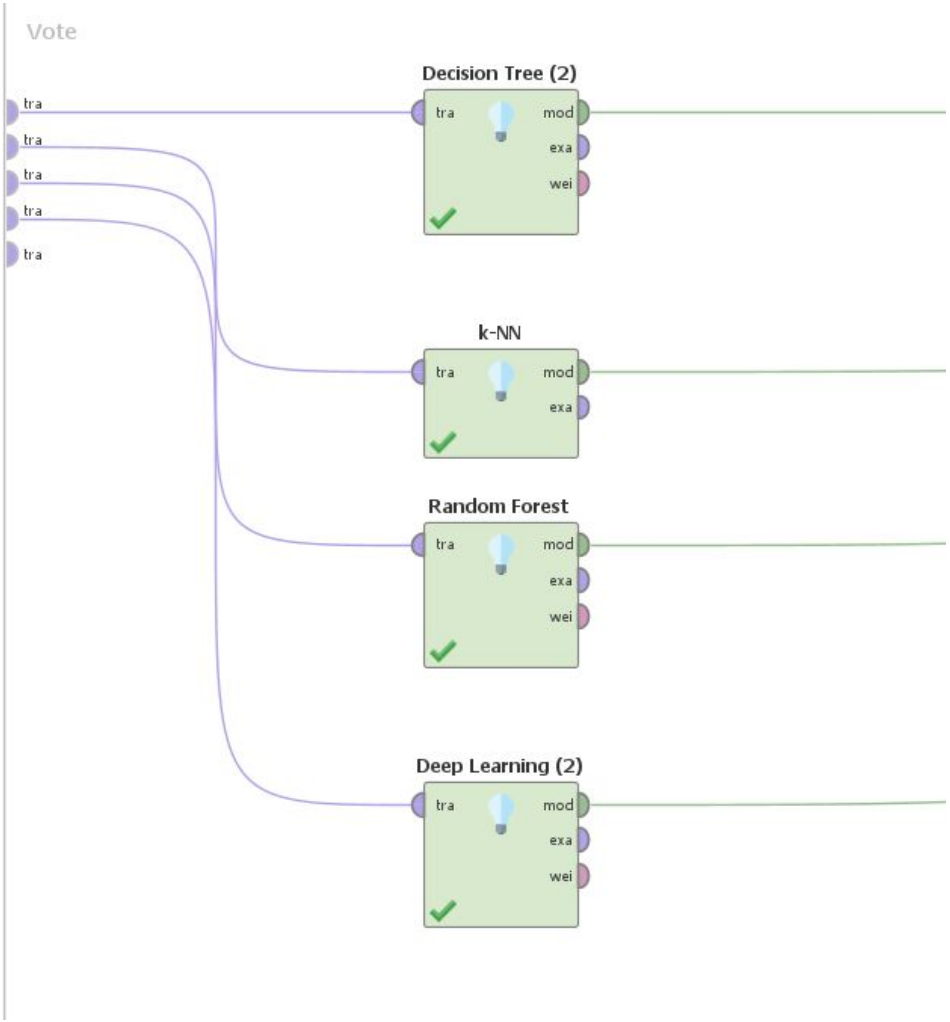
	true NO	true YES	class precision
pred. NO	1600	249	86.53%
pred. YES	138	19	12.10%
class recall	92.06%	7.09%	

06 Voting



Voting (Ensemble)




Vote
tra
tra
tra
tra
tra
Decision Tree (2)
tra
mod
exa
wei
k-NN
tra
mod
exa
Random Forest
tra
mod
exa
wei
Deep Learning (2)
tra
mod
exa
wei

Model Result

accuracy: 82.53%

	true NO	true YES	class precision
pred. NO	490	74	86.88%
pred. YES	31	6	16.22%
class recall	94.05%	7.50%	