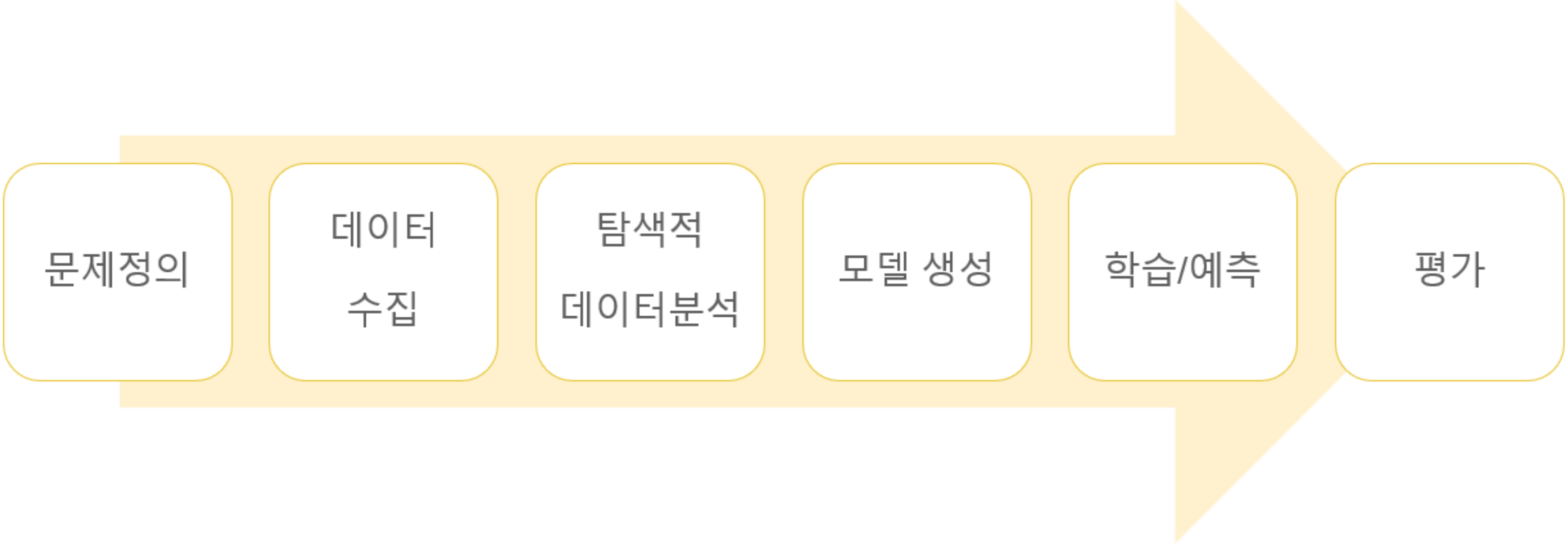
지도학습: 분류 (실습)

홍익 대학교
Hyun-Sun Ryu

분류 프로세스



문제정의

데이터 수집

탐색적 데이터분석

모델 생성

학습/예측

평가

분류 프로세스

문제정의 데이터수집 탐색적 데이터분석 모델생성 학습/예측 평가

- Titanic 생존자 예측 모델 만들기

< titanic 탑승자 데이터를 바탕으로 생존자와 사망자 분류하기 >

분류 프로세스

문제정의	**데이터수집**	탐색적 데이터분석	모델생성	학습/예측	평가

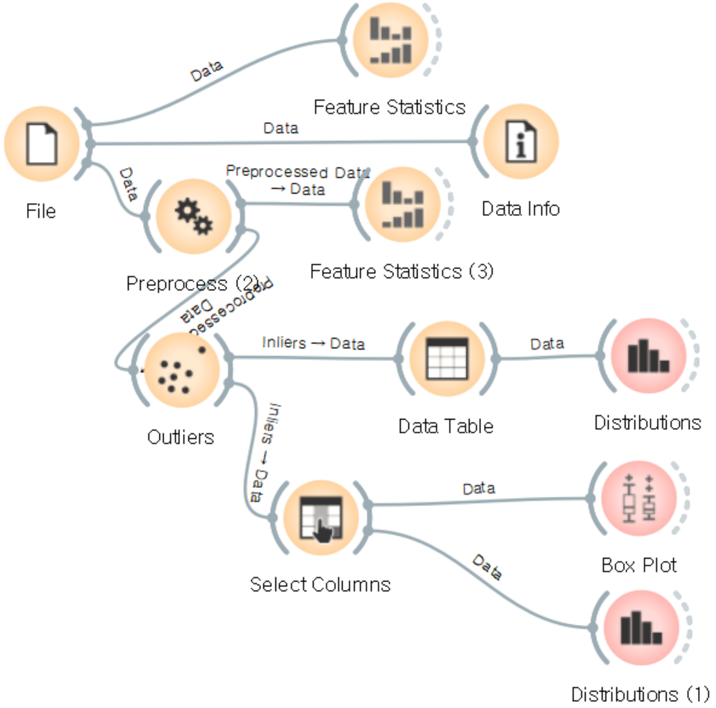
- 캐글 누리집 – compete항목에서 다운로드

https://www.kaggle.com/c/titanic/data

- 모델 훈련을 위하여 train 데이터가 필요함

분류 프로세스

문제정의 | 데이터수집 | **탐색적데이터분석** | 모델생성 | 학습/예측 | 평가



분류 프로세스

문제정의

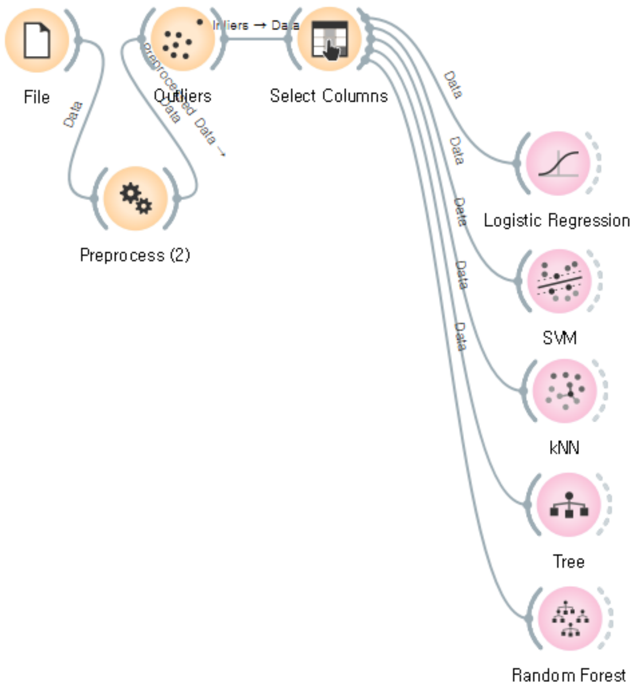
데이터수집

탐색적데이터분석

모델생성

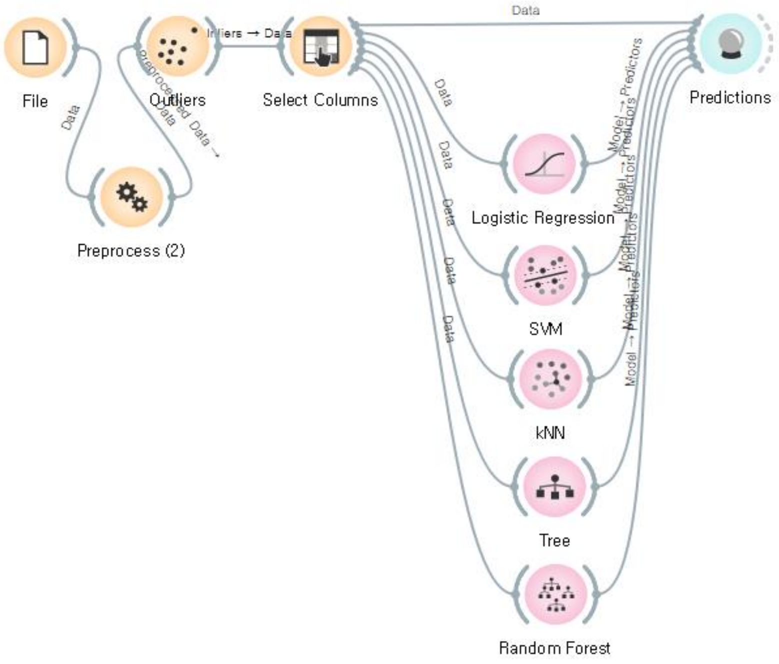
학습/예측

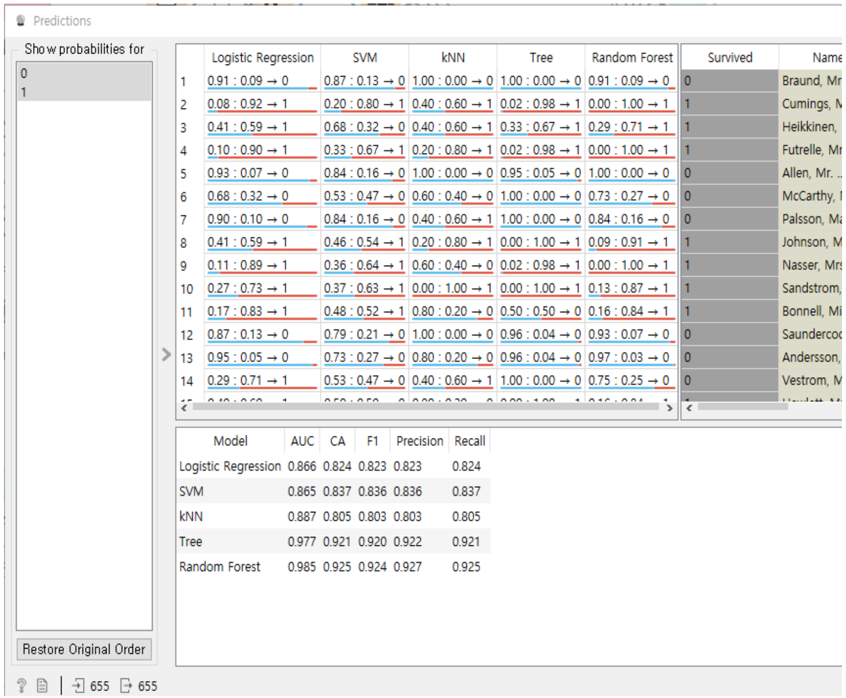
평가



분류 프로세스

문제정의 | 데이터수집 | 탐색적데이터분석 | 모델생성 | **학습/예측** | 평가



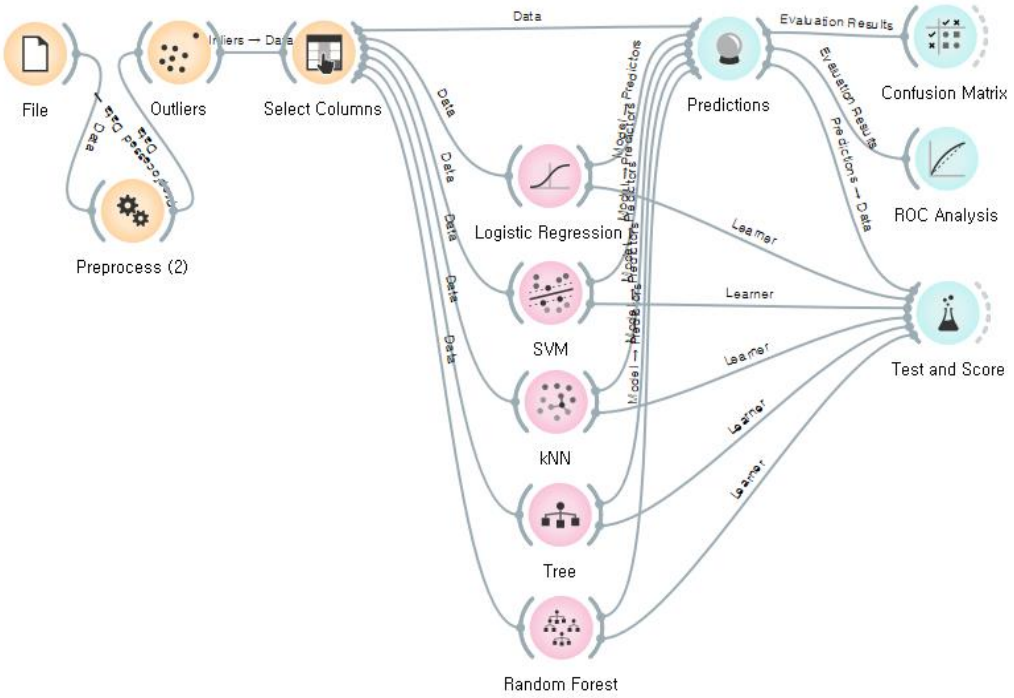


	Logistic Regression	SVM	kNN	Tree	Random Forest	Survived	Name
1	0.91 : 0.09 → 0	0.87 : 0.13 → 0	1.00 : 0.00 → 0	1.00 : 0.00 → 0	0.91 : 0.09 → 0	0	Braund, Mr
2	0.08 : 0.92 → 1	0.20 : 0.80 → 1	0.40 : 0.60 → 1	0.02 : 0.98 → 1	0.00 : 1.00 → 1	1	Cumings, M
3	0.41 : 0.59 → 1	0.68 : 0.32 → 0	0.40 : 0.60 → 1	0.33 : 0.67 → 1	0.29 : 0.71 → 1	1	Heikkinen,
4	0.10 : 0.90 → 1	0.33 : 0.67 → 1	0.20 : 0.80 → 1	0.02 : 0.98 → 1	0.00 : 1.00 → 1	1	Futrelle, Mr
5	0.93 : 0.07 → 0	0.84 : 0.16 → 0	1.00 : 0.00 → 0	0.95 : 0.05 → 0	1.00 : 0.00 → 0	0	Allen, Mr. ...
6	0.68 : 0.32 → 0	0.53 : 0.47 → 0	0.60 : 0.40 → 0	1.00 : 0.00 → 0	0.73 : 0.27 → 0	0	McCarthy,
7	0.90 : 0.10 → 0	0.84 : 0.16 → 0	0.40 : 0.60 → 1	1.00 : 0.00 → 0	0.84 : 0.16 → 0	0	Palsson, Ma
8	0.41 : 0.59 → 1	0.46 : 0.54 → 1	0.20 : 0.80 → 1	0.00 : 1.00 → 1	0.09 : 0.91 → 1	1	Johnson, M
9	0.11 : 0.89 → 1	0.36 : 0.64 → 1	0.60 : 0.40 → 0	0.02 : 0.98 → 1	0.00 : 1.00 → 1	1	Nasser, Mrs
10	0.27 : 0.73 → 1	0.37 : 0.63 → 1	0.00 : 1.00 → 1	0.00 : 1.00 → 1	0.13 : 0.87 → 1	1	Sandstrom,
11	0.17 : 0.83 → 1	0.48 : 0.52 → 1	0.80 : 0.20 → 0	0.50 : 0.50 → 0	0.16 : 0.84 → 1	1	Bonnell, Mi
12	0.87 : 0.13 → 0	0.79 : 0.21 → 0	1.00 : 0.00 → 0	0.96 : 0.04 → 0	0.93 : 0.07 → 0	0	Saundercoc
13	0.95 : 0.05 → 0	0.73 : 0.27 → 0	0.80 : 0.20 → 0	0.96 : 0.04 → 0	0.97 : 0.03 → 0	0	Andersson,
14	0.29 : 0.71 → 1	0.53 : 0.47 → 0	0.40 : 0.60 → 1	1.00 : 0.00 → 0	0.75 : 0.25 → 0	0	Vestrom, M

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.866	0.824	0.823	0.823	0.824
SVM	0.865	0.837	0.836	0.836	0.837
kNN	0.887	0.805	0.803	0.803	0.805
Tree	0.977	0.921	0.920	0.922	0.921
Random Forest	0.985	0.925	0.924	0.927	0.925

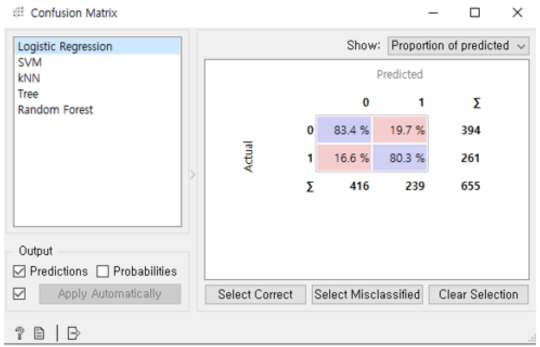
분류 프로세스

문제정의 | 데이터수집 | 탐색적데이터분석 | 모델생성 | 학습/예측 | **평가**



분류 프로세스

문제정의 | 데이터수집 | 탐색적데이터분석 | 모델생성 | 학습/예측 | **평가**



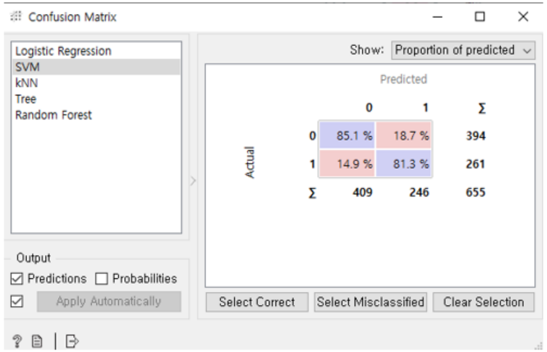
Confusion Matrix

Logistic Regression
SVM
kNN
Tree
Random Forest

Show: Proportion of predicted

		Predicted		
		0	1	Σ
Actual	0	83.4 %	19.7 %	394
	1	16.6 %	80.3 %	261
	Σ	416	239	655

Output
☑ Predictions □ Probabilities
☑ Apply Automatically



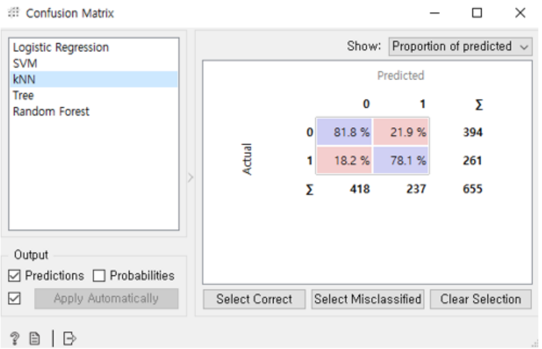
Confusion Matrix

Logistic Regression
SVM
kNN
Tree
Random Forest

Show: Proportion of predicted

		Predicted		
		0	1	Σ
Actual	0	85.1 %	18.7 %	394
	1	14.9 %	81.3 %	261
	Σ	409	246	655

Output
☑ Predictions □ Probabilities
☑ Apply Automatically



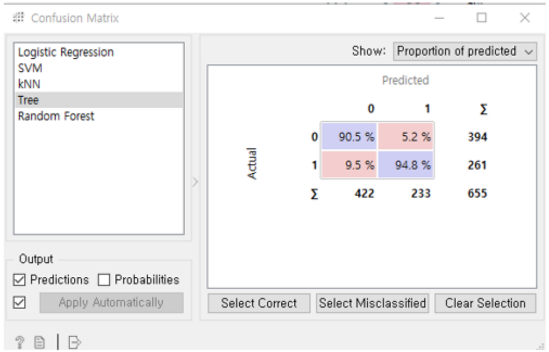
Confusion Matrix

Logistic Regression
SVM
kNN
Tree
Random Forest

Show: Proportion of predicted

		Predicted		
		0	1	Σ
Actual	0	81.8 %	21.9 %	394
	1	18.2 %	78.1 %	261
	Σ	418	237	655

Output
☑ Predictions □ Probabilities
☑ Apply Automatically



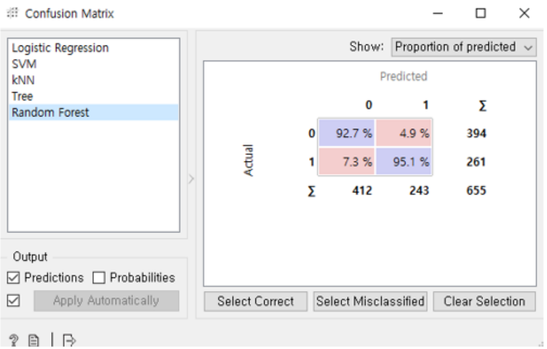
Confusion Matrix

Logistic Regression
SVM
kNN
Tree
Random Forest

Show: Proportion of predicted

		Predicted		
		0	1	Σ
Actual	0	90.5 %	5.2 %	394
	1	9.5 %	94.8 %	261
	Σ	422	233	655

Output
☑ Predictions □ Probabilities
☑ Apply Automatically



Confusion Matrix

Logistic Regression
SVM
kNN
Tree
Random Forest

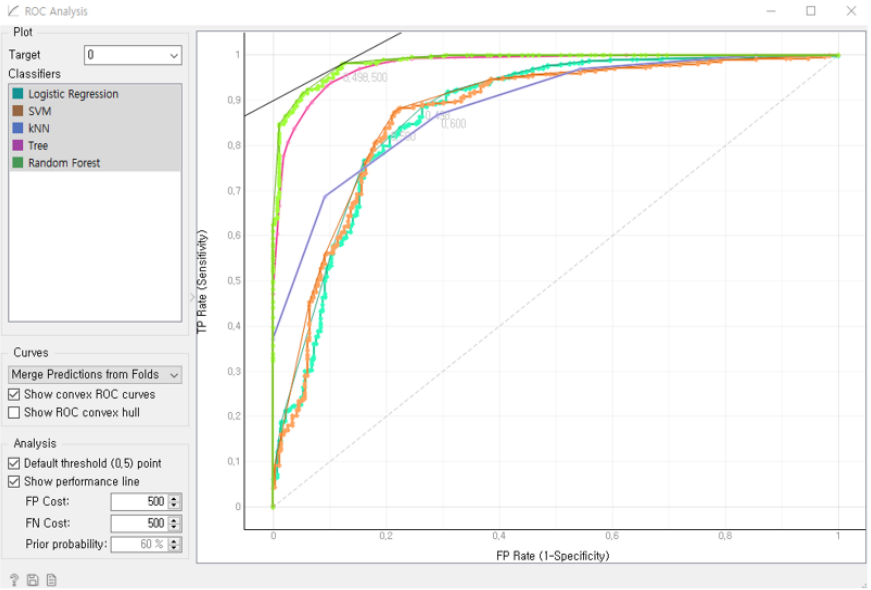
Show: Proportion of predicted

		Predicted		
		0	1	Σ
Actual	0	92.7 %	4.9 %	394
	1	7.3 %	95.1 %	261
	Σ	412	243	655

Output
☑ Predictions □ Probabilities
☑ Apply Automatically

분류 프로세스

문제정의 | 데이터수집 | 탐색적데이터분석 | 모델생성 | 학습/예측 | **평가**



Plot

Target: 0

Classifiers
- 🟦 Logistic Regression
- 🟫 SVM
- 🟦 kNN
- 🟪 Tree
- 🟩 Random Forest

Curves

Merge Predictions from Folds

☑ Show convex ROC curves

☐ Show ROC convex hull

Analysis

☑ Default threshold (0,5) point

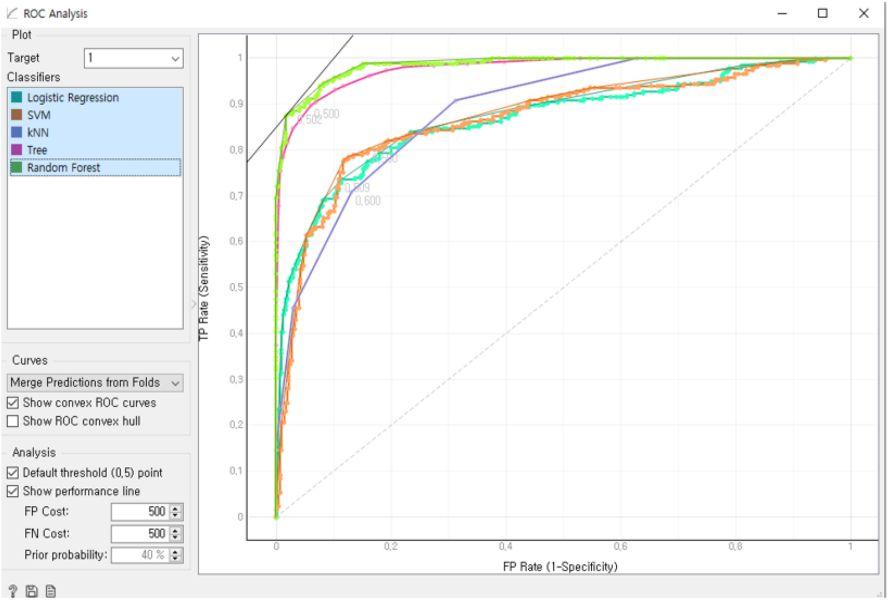
☑ Show performance line

FP Cost: 500

FN Cost: 500

Prior probability: 60 %

사망 ROC



Plot

Target: 1

Classifiers
- 🟦 Logistic Regression
- 🟫 SVM
- 🟦 kNN
- 🟪 Tree
- 🟩 Random Forest

Curves

Merge Predictions from Folds

☑ Show convex ROC curves

☐ Show ROC convex hull

Analysis

☑ Default threshold (0,5) point

☑ Show performance line

FP Cost: 500

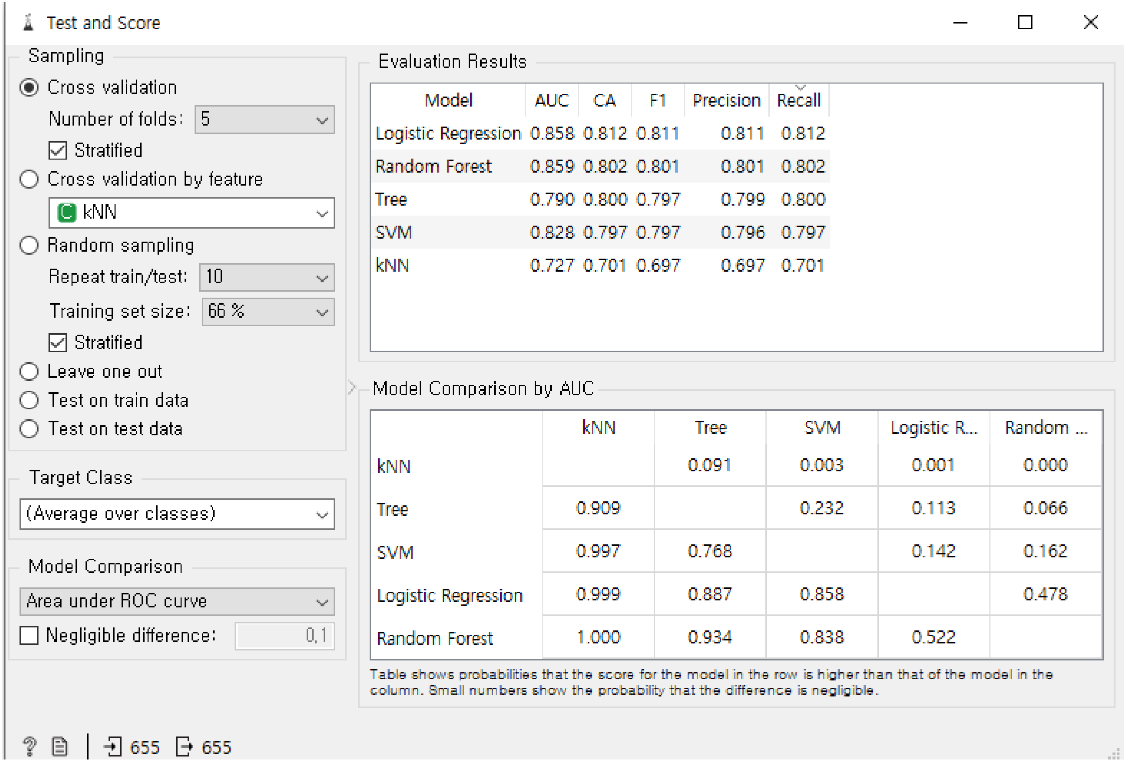
FN Cost: 500

Prior probability: 40 %

생존 ROC

분류 프로세스

문제정의 | 데이터수집 | 탐색적데이터분석 | 모델생성 | 학습/예측 | **평가**



Test and Score

Sampling

● Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

kNN

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Target Class

(Average over classes)

Model Comparison

Area under ROC curve

□ Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.858	0.812	0.811	0.811	0.812
Random Forest	0.859	0.802	0.801	0.801	0.802
Tree	0.790	0.800	0.797	0.799	0.800
SVM	0.828	0.797	0.797	0.796	0.797
kNN	0.727	0.701	0.697	0.697	0.701

Model Comparison by AUC

	kNN	Tree	SVM	Logistic R...	Random ...
kNN		0.091	0.003	0.001	0.000
Tree	0.909		0.232	0.113	0.066
SVM	0.997	0.768		0.142	0.162
Logistic Regression	0.999	0.887	0.858		0.478
Random Forest	1.000	0.934	0.838	0.522	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

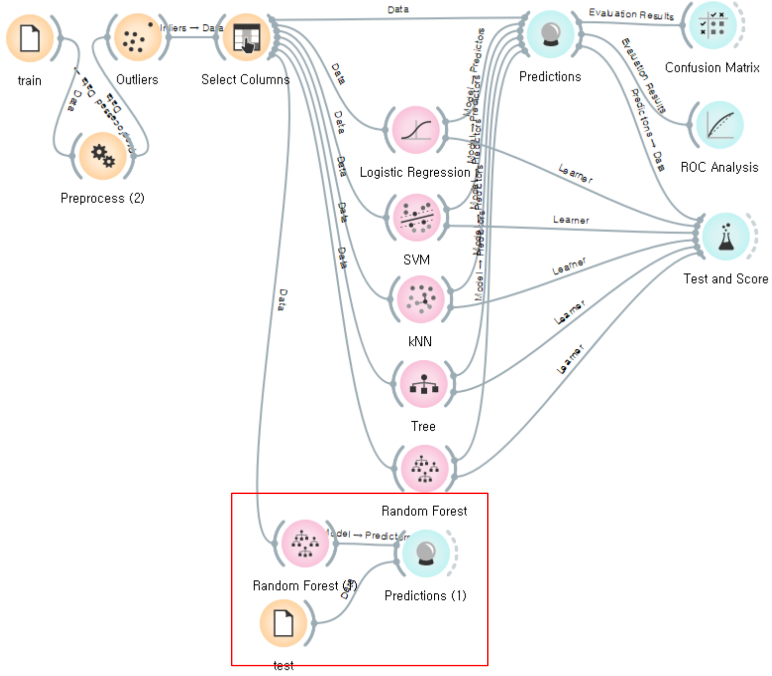
분류 프로세스

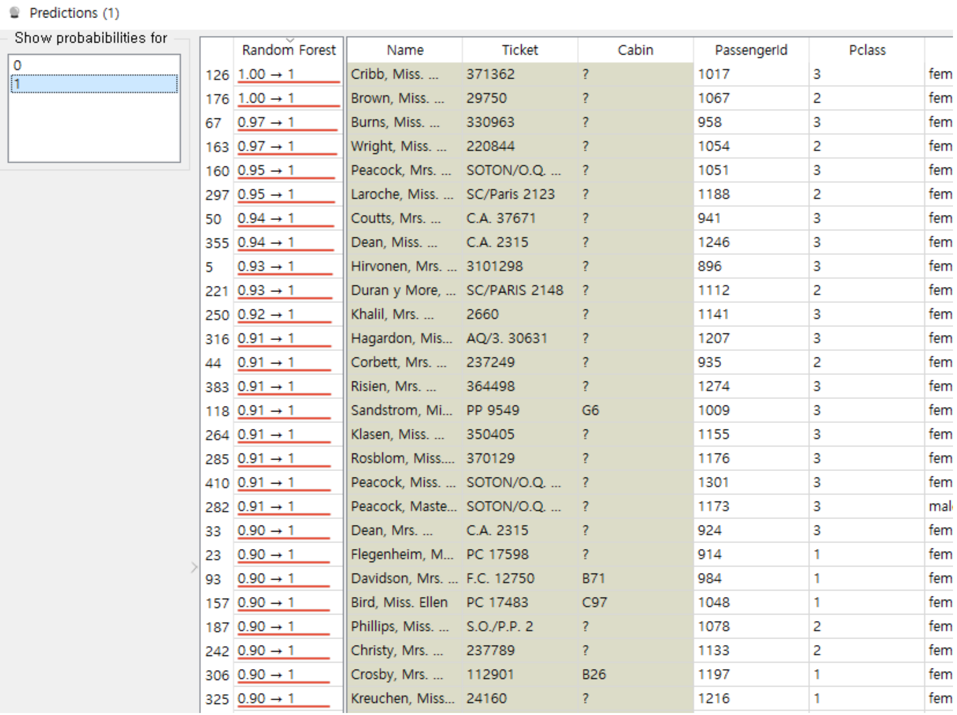
문제정의 | 데이터수집 | 탐색적데이터분석 | 모델생성 | 학습/예측 | **평가**

Test and Score, Confusion matrix, ROC analysis 위젯으로 분류 모델을 평가한 결과 Random forest의 분류 결과가 가장 신뢰도가 높다고 판단

분류 프로세스

문제정의 | 데이터수집 | 탐색적데이터분석 | 모델생성 | **학습/예측** | 평가





	Random Forest	Name	Ticket	Cabin	PassengerId	Pclass	
126	1.00 → 1	Cribb, Miss. ...	371362	?	1017	3	fem
176	1.00 → 1	Brown, Miss. ...	29750	?	1067	2	fem
67	0.97 → 1	Burns, Miss. ...	330963	?	958	3	fem
163	0.97 → 1	Wright, Miss. ...	220844	?	1054	2	fem
160	0.95 → 1	Peacock, Mrs. ...	SOTON/O.Q. ...	?	1051	3	fem
297	0.95 → 1	Laroche, Miss. ...	SC/Paris 2123	?	1188	2	fem
50	0.94 → 1	Coutts, Mrs. ...	C.A. 37671	?	941	3	fem
355	0.94 → 1	Dean, Miss. ...	C.A. 2315	?	1246	3	fem
5	0.93 → 1	Hirvonen, Mrs. ...	3101298	?	896	3	fem
221	0.93 → 1	Duran y More, ...	SC/PARIS 2148	?	1112	2	fem
250	0.92 → 1	Khalil, Mrs. ...	2660	?	1141	3	fem
316	0.91 → 1	Hagardon, Mis...	AQ/3. 30631	?	1207	3	fem
44	0.91 → 1	Corbett, Mrs. ...	237249	?	935	2	fem
383	0.91 → 1	Risien, Mrs. ...	364498	?	1274	3	fem
118	0.91 → 1	Sandstrom, Mi...	PP 9549	G6	1009	3	fem
264	0.91 → 1	Klasen, Miss. ...	350405	?	1155	3	fem
285	0.91 → 1	Rosblom, Miss....	370129	?	1176	3	fem
410	0.91 → 1	Peacock, Miss. ...	SOTON/O.Q. ...	?	1301	3	fem
282	0.91 → 1	Peacock, Maste...	SOTON/O.Q. ...	?	1173	3	mal
33	0.90 → 1	Dean, Mrs. ...	C.A. 2315	?	924	3	fem
23	0.90 → 1	Flegenheim, M...	PC 17598	?	914	1	fem
93	0.90 → 1	Davidson, Mrs. ...	F.C. 12750	B71	984	1	fem
157	0.90 → 1	Bird, Miss. Ellen	PC 17483	C97	1048	1	fem
187	0.90 → 1	Phillips, Miss. ...	S.O./P.P. 2	?	1078	2	fem
242	0.90 → 1	Christy, Mrs. ...	237789	?	1133	2	fem
306	0.90 → 1	Crosby, Mrs. ...	112901	B26	1197	1	fem
325	0.90 → 1	Kreuchen, Miss...	24160	?	1216	1	fem

질문 있나요?

hsryu13@hongik.ac.kr