

데이터 과학 기초: 과제 #2.

Kaggle Data Analysis with Orange

컴퓨터학부 우성현

1. 자전거 대여 수요 예측

(<https://www.kaggle.com/c/bike-sharing-demand>)

1.1. 개요

▶ Bike sharing systems

: A means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

1.2. 데이터셋

Source				
File: bike_train.csv				
Info				
10886 instance(s)				
12 feature(s) (no missing values)				
Data has no target variable.				
0 meta attribute(s)				
Columns (Double click to edit)				
	Name	Type	Role	Values
1	datetime	datetime	feature	
2	season	categorical	feature	
3	holiday	categorical	feature	0, 1
4	workingday	categorical	feature	0, 1
5	weather	categorical	feature	
6	temp	numeric	feature	
7	atemp	numeric	feature	
8	humidity	numeric	feature	
9	windspeed	numeric	feature	
10	casual	numeric	feature	
11	registered	numeric	feature	
12	count	numeric	target	

Source				
File: bike_test.csv				
Info				
6493 instance(s)				
9 feature(s) (no missing values)				
Data has no target variable.				
0 meta attribute(s)				
Columns (Double click to edit)				
	Name	Type	Role	Values
1	datetime	datetime	feature	
2	season	categorical	feature	
3	holiday	categorical	feature	0, 1
4	workingday	categorical	feature	0, 1
5	weather	categorical	feature	
6	temp	numeric	feature	
7	atemp	numeric	feature	
8	humidity	numeric	feature	
9	windspeed	numeric	feature	

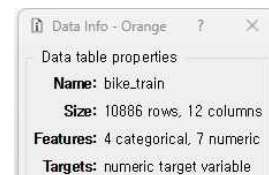
데이터셋의 각 변수별 Type과 Role, 변수에 대한 설명은 다음 표를 참조하도록 한다.

No.	Name	Type	Role	Description
1	datetime	datetime	feature	hourly date + timestamp
2	season	categorical	feature	1=봄 2=여름 3=가을 4=겨울
3	holiday	categorical	feature	0=국경일이 아닌 날, 1=국경일(토요일, 일요일은 제외)
4	workingday	categorical	feature	0=주말 또는 휴일, 1=일하는 날
5	weather	categorical	feature	1=맑음 2=안개 3=가벼운 비 또는 눈 4=심한 비 또는 눈
6	temp	numeric	feature	온도(단위: 섭씨)
7	atemp	numeric	feature	체감온도(단위: 섭씨)
8	humidity	numeric	feature	상대습도
9	windspeed	numeric	feature	풍속
10	casual	numeric	feature	사전에 등록하지 않은 사용자가 대여한 횟수
11	registered	numeric	feature	사전에 등록한 사용자가 대여한 횟수
12	count	numeric	target	대여 횟수

1.3. 탐색적 데이터 분석

1) File에서 bike_train.csv 파일 로드

- Type과 Role을 표를 참조하여 변경 후 [Apply] 클릭
- Data Info 연결: Data Set Size의 Row, Column 갯수는 각각 얼마인가?
▶ 10886 rows, 12 columns
- Features에서 Categorical, Numeric은 각각 얼마인가?
▶ 4 categorical, 7 numeric



Data table properties	
Name:	bike_train
Size:	10886 rows, 12 columns
Features:	4 categorical, 7 numeric
Targets:	numeric target variable

2) File에서 bike_test.csv 파일 로드

- Type과 Role을 표를 참조하여 변경 후 [Apply] 클릭
- Data Info 연결: Data Set Size의 Row 갯수는 얼마인가?
▶ 6493 rows
- Features에서 Categorical, Numeric은 각각 얼마인가?
▶ 4 categorical, 5 numeric



Data table properties	
Name:	bike_test
Size:	6493 rows, 9 columns
Features:	4 categorical, 5 numeric

3) File:bike_train에 Feature Statistics 연결

- temp의 평균값과 중앙값은 각각 얼마인가? Mean : 20.2309 / Median : 20.50
- count의 평균값과 중앙값은 각각 얼마인가? Mean : 191.57 / Median : 145
- 정규분포(normal dist.)를 따르는 것처럼 보이는 feature는 무엇인가? temp, atemp, humidity
- 균일분포(uniform dist.)를 따르는 것처럼 보이는 feature는 무엇인가? season
- 멱함수 분포(power-law dist.)를 따르는 것처럼 보이는 feature는 무엇인가? casual, registered, count



4) File:bike_test에 Feature Statistics 연결

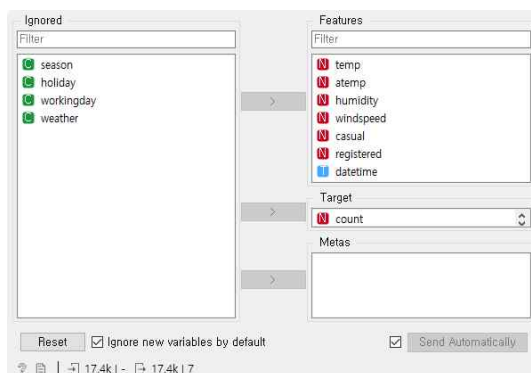
- temp의 평균값과 중앙값은 각각 얼마인가? Mean : 20.6206 / Median : 21.32
- windspeed의 평균값과 중앙값은 각각 얼마인가? Mean : 12.631157 / Median : 11.0014

5) Concatenate로 File:bike_train과 File:bike_test를 병합한 후

- Data Info 연결: Data Set Size의 Row 갯수는 얼마인가? 17379 rows
- Feature Statistics 연결: count의 평균값과 중앙값은 각각 얼마인가? Mean:191.57 / Median : 145



5) Concatenate한 데이터에 Select Columns로 numeric, datetime feature만 선택



- Correlations 연결: Pearson 상관관계수의 절대값이 가장 높은 두 변수는? registered, casual

1	+0.967	count	registered
2	+0.780	casual	count
3	+0.315	count	temp
4	+0.314	atemp	count
5	-0.253	count	humidity
6	+0.237	count	datetime
7	+0.093	count	windspeed

- 위에서 두 변수의 상관관계수가 가장 높은 이유는? target인 count(대여 횟수)는 독립변수 registered(사전에 등록한 사용자가 대여한 횟수)와 casual(사전에 등록하지 않은 사용자가 대여한 횟수)의 합이기에 당연히 두 독립변수와는 양의 상관관계를 가지며 높은 Pearson 상관관계수를 가진다.

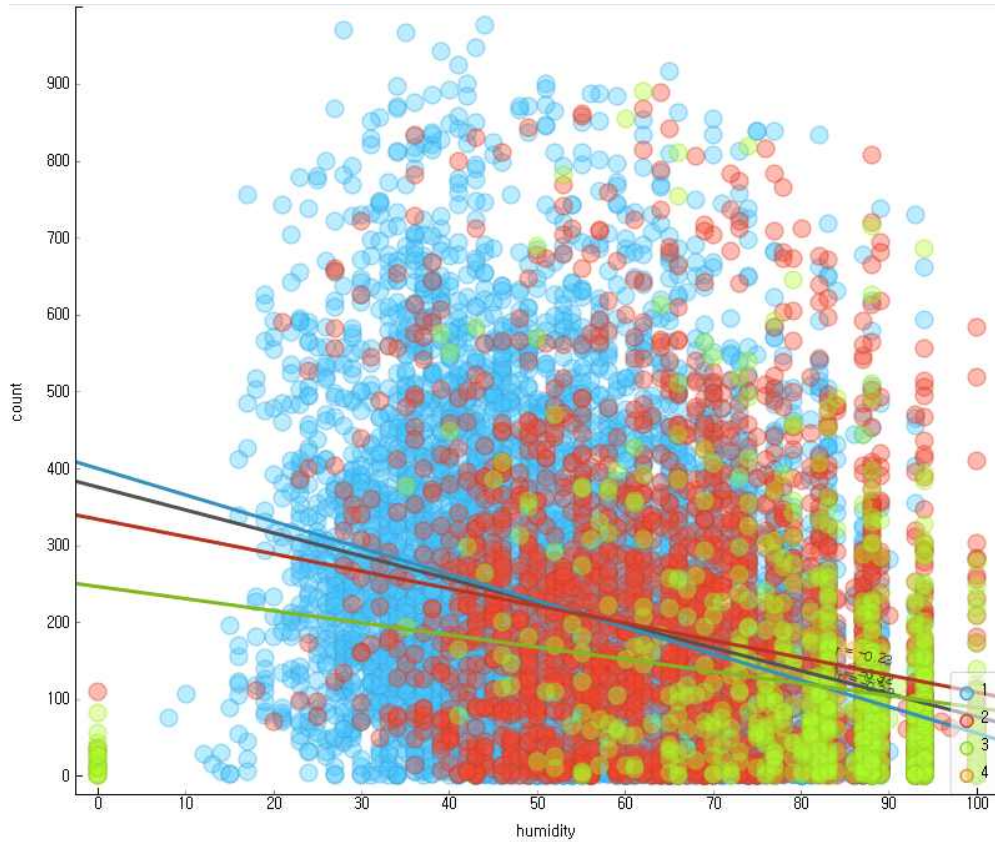
- temp와 count의 상관관계수는 얼마인가? +0.315

- temp와 count의 상관관계수를 보고 알 수 있는 것은?

강하지는 않지만 양의 상관관계를 보인다. Scatter Plot을 보면 자전거 타기 좋은 온도인 12~30도 구간에 온도가 높아질수록, 많은 대여 횟수를 보인다.

- count와 humidity의 상관관계를 분석하고, 그 의미를 해석하라.

count와 humidity의 Pearson 상관계수는 -0.253 으로 약한 음의 상관관계를 가지고 있다. humidity가 높아질수록 대여 횟수는 줄어드는 경향이 있다. Scatter Plot(train data set)을 통해 데이터를 자세히 보면, 습도가 낮은 구간에는 맑은 날씨(1)가 많이 분포하지만, 습도가 높아질수록 좋지 않은 날씨가 분포되어있음을 확인할 수 있다. 이는 자전거 타기에 좋지 않은 환경이 습도가 높은 경우가 많기 때문이라고 생각할 수 있지만 상관관계가 곧 인과관계를 말하는 것은 아니기에 정확한 인과관계는 독립변수를 통제된 상황에서 분석해봐야 알 수 있다.



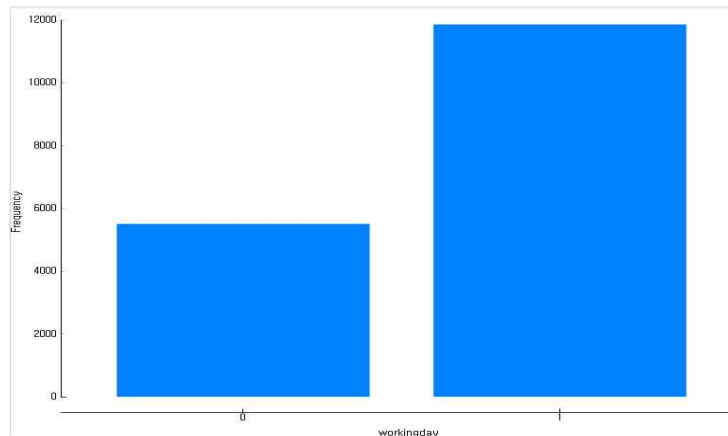
- count와 windspeed의 상관관계를 분석하고, 그 의미를 해석하라.

count와 windspeed의 상관계수는 $+0.093$ 으로 아주 약한 양의 상관관계를 가지고 있고, 7개의 독립변수의 상관계수의 절댓값이 가장 낮다. 이를 보건대 자전거를 대여하는 것에 있어 풍속은 그리 큰 상관관계가 없다는 것을 알 수 있다.

이를 토대로 자전거를 타는 것에 있어서 일반인은 풍속을 크게 고려하지 않고, 전문가는 일반적으로 풍속을 고려할 것이지만 개인 자전거를 이용해 data set에 포함되지 않았을까? 라는 추측해 볼 수 있지만, 정확한 분석을 위해서는 대여 자전거를 이용하는 이용자의 정보와 용도, 전문가들이 풍속에 신경을 쓰는지 등에 대한 추가적인 정보가 필요하다.

6) Concatenate한 데이터에 Distributions 연결:

- workingday에 대한 분포를 확인하라:



- 위 분포로부터 해석할 수 있는 것은?

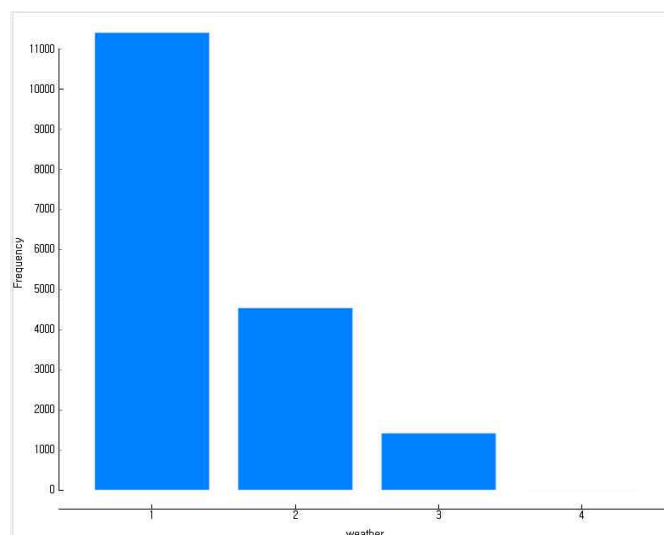
일하는 날과 그렇지 않은 날의 대여 횟수의 비율은 대략 2:1로 일하는 날에 더 많은 대여 횟수를 보인다. 그렇지만 보통의 경우, 일하는 날은 5일이고, 주말은 2일이기에 각 빈도수를 일수로 나누게 되면, 일일 대여 횟수는 주말이 더 높다는 것을 알 수 있다.

- 위 해석으로부터 통찰할 수 있는 것은?

대여 자전거를 이용해 평일에 하지 못하는 활동을 주말에 한다는 것을 알 수 있다. 평일에는 학교나 직장 등 자전거를 탈 수 없는 고정적인 시간이 있기에 물리적으로 주말에 더 많은 이용량을 보인다고 생각해볼 수 있지만, 정확한 것은 시간대별 이용량과 용도 등의 자료를 추가적인 확인을 할 필요가 있다.

서비스를 운영하는 회사는 평일에 대여 자전거 이용의 제한 사항과 해결 방안을, 주말의 대여량을 늘리기 위해 주로 대여하는 장소에 더 많은 자전거를 배치하거나 이벤트를 하는 등 서비스 수익 확대를 위해 여러 가지를 고려해볼 수 있다.

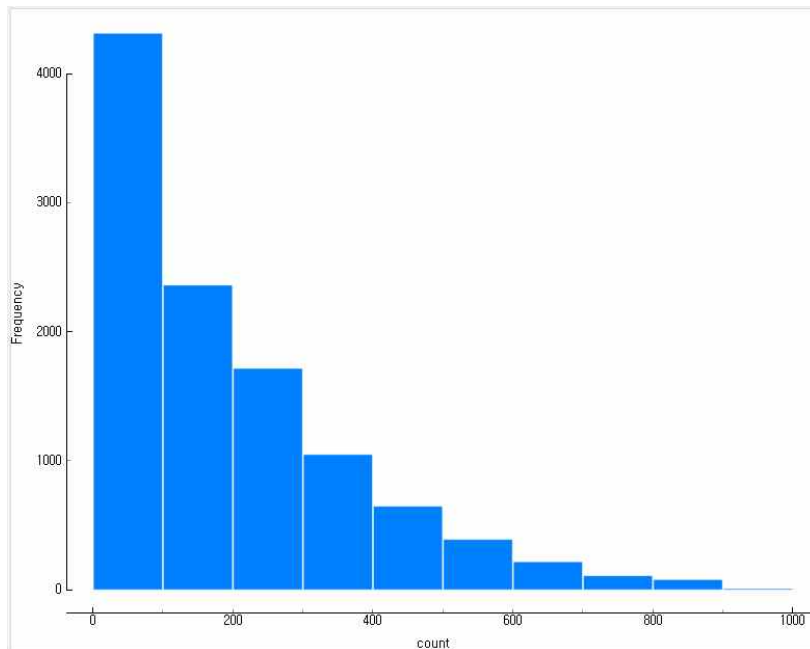
- weather에 대한 분포를 확인하라:



- 위 분포로부터 해석할 수 있는 것과 통찰할 수 있는 것은?

날씨가 맑을수록 더 많은 대여가 발생했다. weather 4에 해당하는 심한 비 또는 눈이 온 경우에는 전체 기간 중 단 3건의 대여만 존재한다. 자전거는 날씨에 많은 영향을 받는다. 좋은 날씨와 자전거 대여 횟수 간 많은 상관관계가 있기에 날씨 예보에 따른 자전거 배치를 고려한다면, 자전거 유지 보수 비용과 수익 증대를 기대해볼 수 있다.

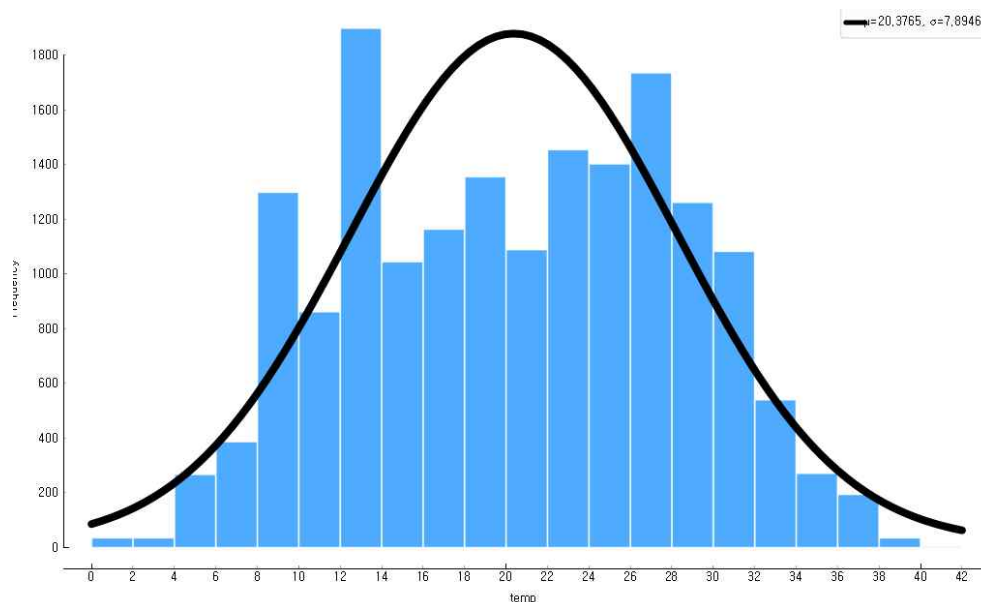
- count에 대한 분포를 확인하라: (Binwidth=100)



- 위 분포로부터 해석할 수 있는 것과 통찰할 수 있는 것은?

count의 분포를 보면 멱함수 분포를 확인할 수 있다. count는 시간당 대여 횟수에 대한 변수로 특정한 시간대에는 많은 횟수를 보이지만, 대부분의 시간대에는 0~100의 총대여 횟수를 보인다. data table을 함께 보면 자전거 타는 행위가 시간대에 많은 영향을 받으며 위의 분포는 이러한 특성이 반영되어 있음을 알 수 있다.

- temp에 대한 분포를 확인하라: (Binwidth=2)



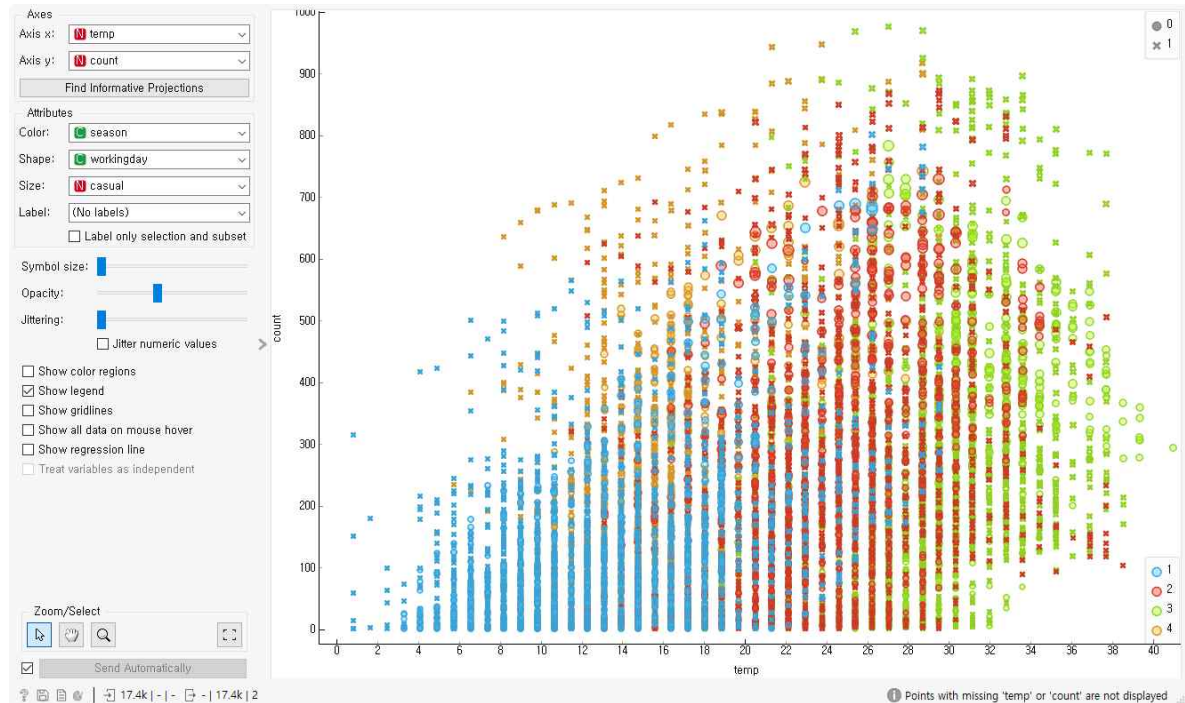
- 위 분포로부터 해석할 수 있는 것과 통찰할 수 있는 것은?

temp의 분포는 대략적으로 정규분포의 형태를 보인다. 완전한 정규분포가 아닌 것은 자전거를 대여하는 빈도가 온도에 완전한 상관관계가 있지 않기 때문이다. Pearson 상관계수 역시 +0.315로 약한 양의 상관관계가 있다. 이를 통해 자전거를 이용함에 있어 좋은 온도는 있지만, 이 좋은 온도의 범위는 상당히 넓고 이 범위에만 들어오면 자전거를 이용하는데 큰 제한을 받지 않는다고

볼 수 있다. 또한 시간대와 함께 생각한다면 자전거 대여 횟수가 많은 시간대는 주로 위의 범위에 속하기에 이러한 분포가 나온다고 추측해볼 수 있다.

7) Concatenate한 데이터에 Scatter Plot 연결:

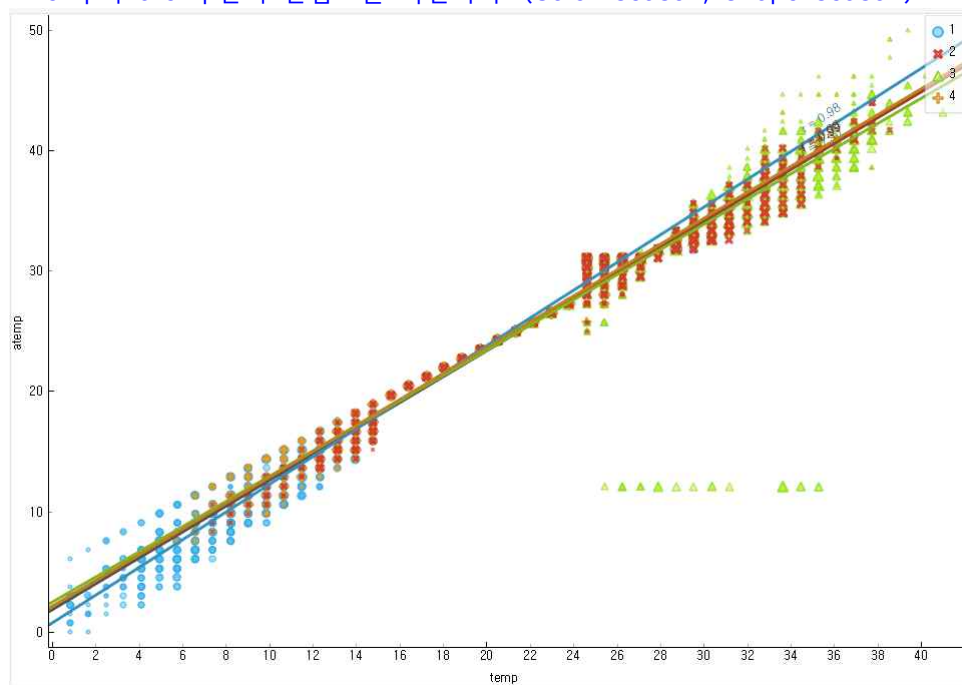
- temp와 count간의 산점도를 확인하라: (Color=season, Shape=workingset)



- 위 산점도로부터 파악할 수 있는 것은?

전체적으로 온도와 대여 횟수는 양의 상관관계를 알 수 있다. 그리고 season 별로 대략적인 온도 분포와 대여 횟수를 알 수 있다. 또한 가장 많은 대여 횟수를 기록한 지점들을 보면 workingday가 1(일하는 날)임을 파악할 수 있다.

- temp와 atemp간의 산점도를 확인하라: (Color=season, Shape=season)

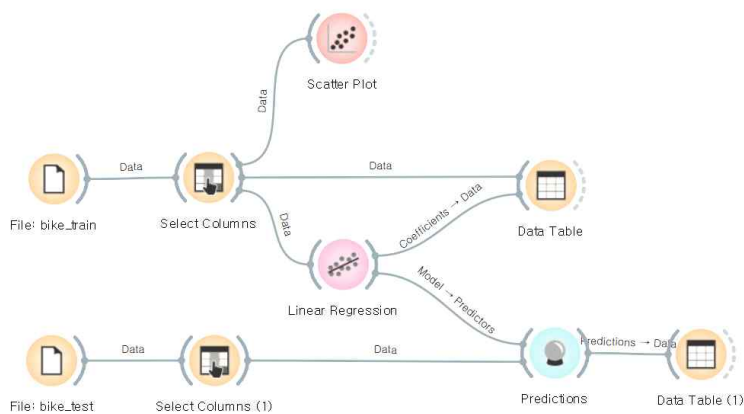


- 위 산점도로부터 파악할 수 있는 것은?

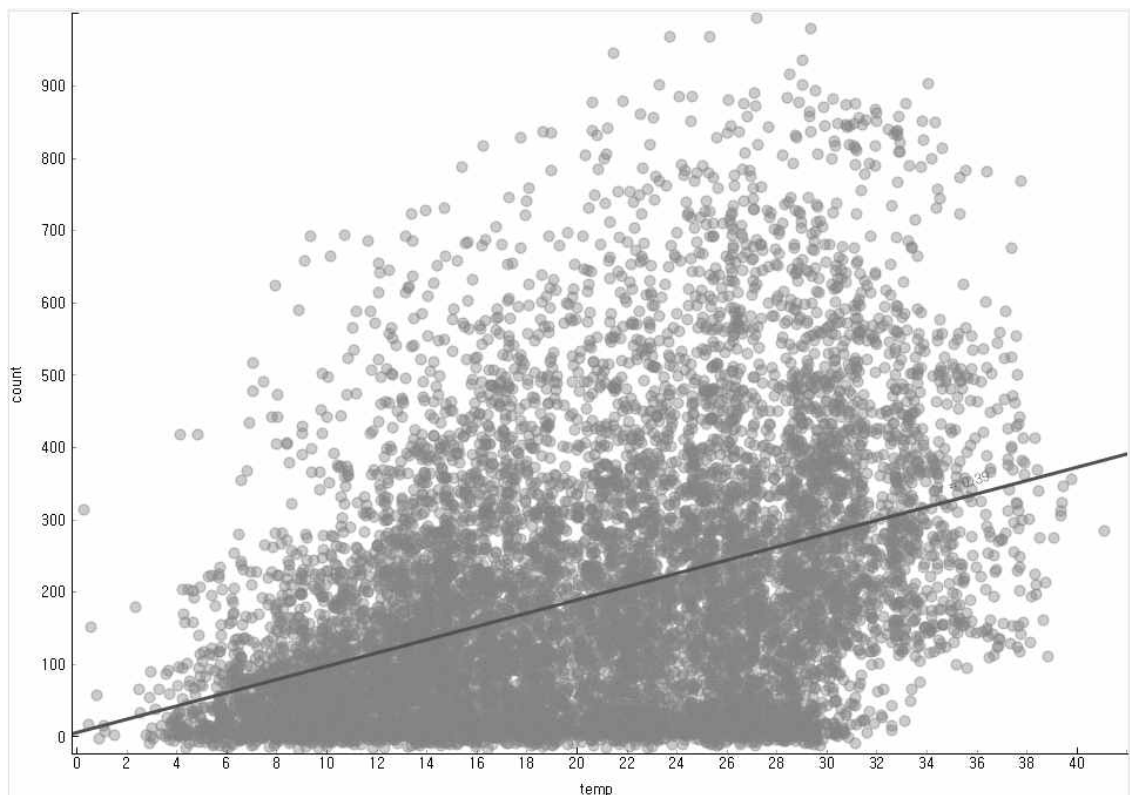
온도와 체감온도의 Pearson 상관계수는 +0.99로 아주 강한 양의 상관관계를 가지고 있다. 주로 봄에는 실제 온도보다 더 낮은 체감온도를 보이고, 가을에는 실제 온도보다 더 높은 체감온도를 느낀다.

1.4. 선형 회귀 분석과 예측

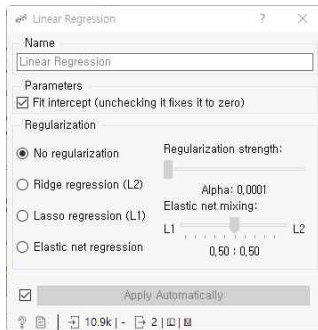
bike_train.csv 파일로부터 (temp, count) 컬럼만 선택하여 선형 회귀로 학습을 한 후에, bike_test.csv 파일로부터 (temp) 컬럼만 선택하여 예측을 해본다. 실습을 위한 워크플로우는 다음과 같다.



1) 위와 같은 워크플로우에서 Scatter Plot이 어떻게 나타나는가?



2) Linear Regression을 아래와 같이 설정했을 때,



- 선형 회귀식의 절편(intercept)과 temp 변수의 기울기(slope)는 각각 얼마인가?

▶ intercept: 6.04621 / temp: 9.17054

coefficients		bike_train
	name	coef
1	intercept	6.04621
2	temp	9.17054

- 이때, temp의 값이 10.66이라면 Prediction이 예측하는 값은 얼마인가? 103.804

- 이때, temp의 값이 9.84라면 Prediction이 예측하는 값은 얼마인가? 96.2843

	Linear Regressor	temp
1	103.804	10.66
2	96.2843	9.84

3) Select Columns에서 (temp, atemp, humidity, windspeed)를 선택하고, 다중변수 선형회귀를 위와 동일한 조건에서 적용했다. (train에서는 target을 count로 뒀음)

- 선형 회귀식의 절편(intercept)과 temp 변수의 기울기(slope)는 각각 얼마인가? 164.37/2.41

coefficients		bike_train
	name	coef
1	intercept	164.373
2	temp	2.41228
3	atemp	5.90999
4	humidity	-2.7305
5	windspeed	0.592099

- atemp, humidity, windspeed 변수의 기울기는 각각 얼마인가?

▶ atemp: 5.91 /humidity: -2.73 /windspeed: 0.59

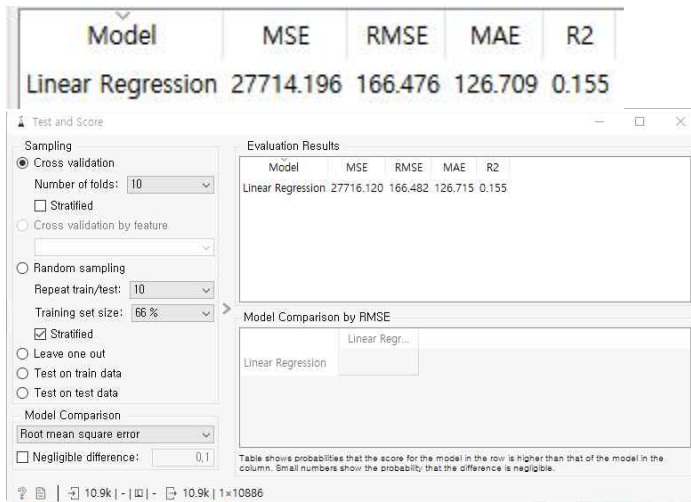
4) (temp, atemp, humidity, windpeed)의 값이 다음과 같을 때, 다중변수 선형회귀의 예측값은 각각 얼마인가?

- (10.66, 11.365, 56, 26.0027): 119.743

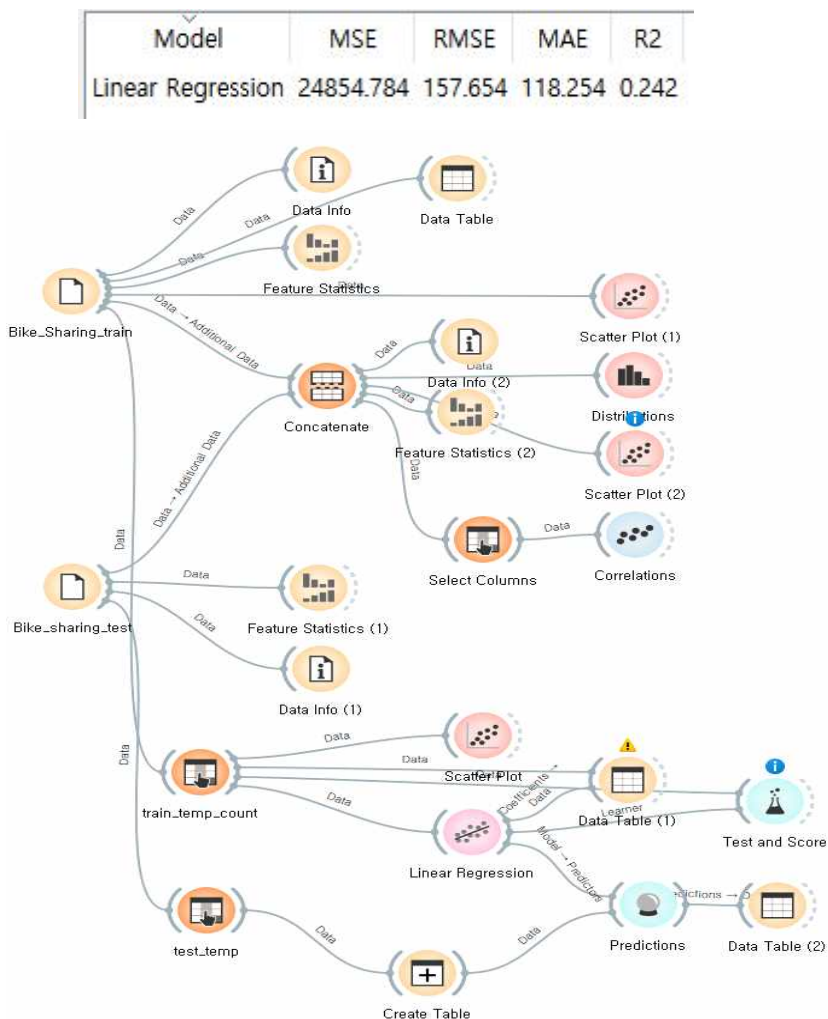
- (9.84, 11.365, 60, 15.0013): 100.329

	Linear Regressor	temp	atemp	humidity	windspeed
1	119.743	10.66	11.365	56	26.0027
2	100.329	9.84	11.365	60	15.0013

- 5) bike_train에서 Select Columns를 (temp, count)만 했을때,
 - Linear Regression의 MSE, RMSE, MAE, R2의 값은 얼마인가?
 ▶ MSE :27714.196, RMSE :166.476, MAE:126.709, R2:0.1555



- 6) Select Columns에서 (temp, atemp, humidity, windspeed)를 선택했을 때,
 - Linear Regression의 MSE, RMSE, MAE, R2의 값은 얼마인가?
 ▶ MSE:24854.784, RMSE:157.654, MAE:118.254, R2:0.242



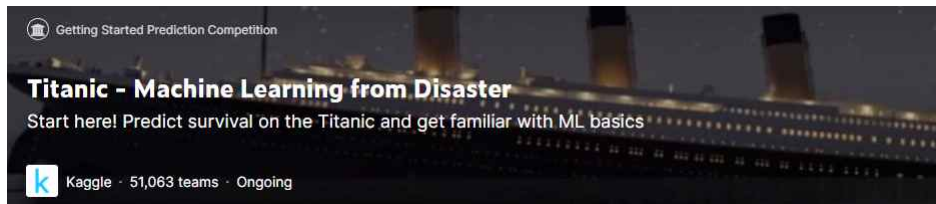
2. 타이타닉 호의 생존자 예측

2.1. 개요

본 과제에서는 캐글의 타이타닉 데이터셋으로 분류 알고리즘을 비교해 보고, MNIST 필기체 인식 머신러닝 알고리즘을 실습한다.

2.2. 타이타닉호의 생존자 예측

<https://www.kaggle.com/c/titanic>



과제에 사용할 데이터셋은 캐글에서 직접 다운로드한다.

Titanic Dataset:

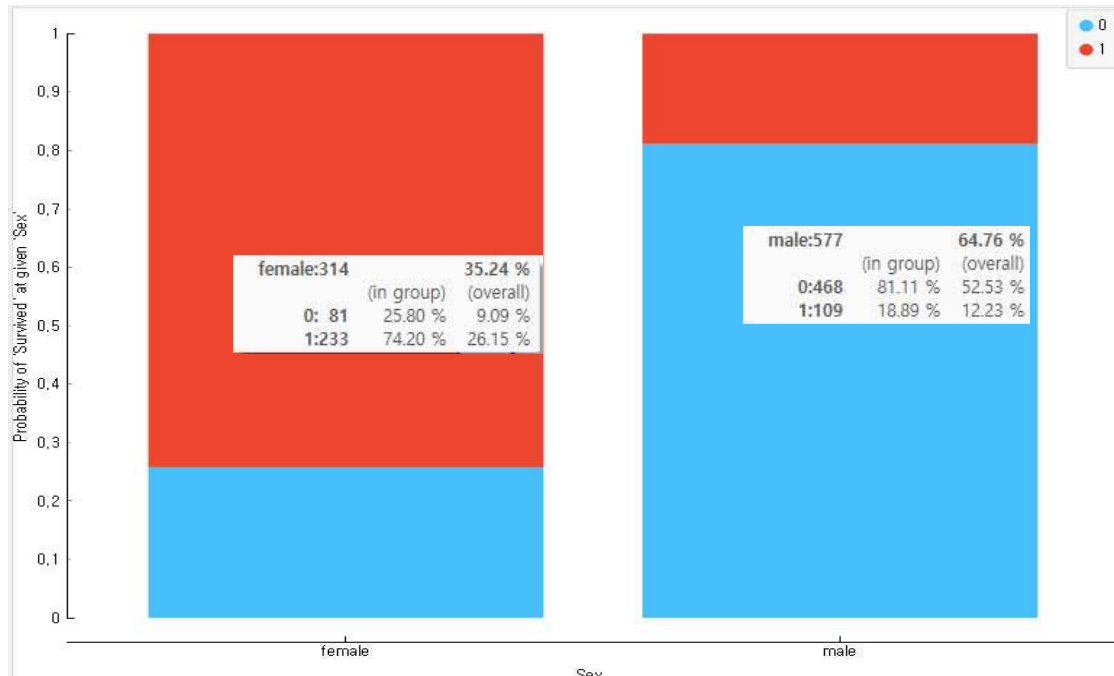
titanic_train.csv: 훈련용 데이터

titanic_test.csv: 시험용 데이터

데이터셋의 각 변수별 Type과 Role, 변수에 대한 설명은 다음 표를 참조하도록 한다.

No.	Name	Type	Role	Description
1	PassengerId	numeric	feature	승객의 아이디
2	Survived	categorical	target	생존유무: 0=사망, 1=생존
3	Pclass	numeric	feature	객실 등급: 1=1등급, 2=2등급, 3=3등급
4	Sex	categorical	feature	성별: female=여성, male=남성
5	Age	numeric	feature	나이
6	SibSp	numeric	feature	함께 탑승한 형제자매, 배우자 수의 합
7	ParCh	numeric	feature	함께 탑승한 부모, 자녀 수의 합
8	Fare	numeric	feature	운임
9	Embarked	categorical	feature	탑승 항구: C=Cherbourg, Q=Queenstown, S=Southampton
10	Name	text	meta	승객 이름
11	Ticket	text	meta	티켓 번호
12	Cabin	text	meta	객실 번호

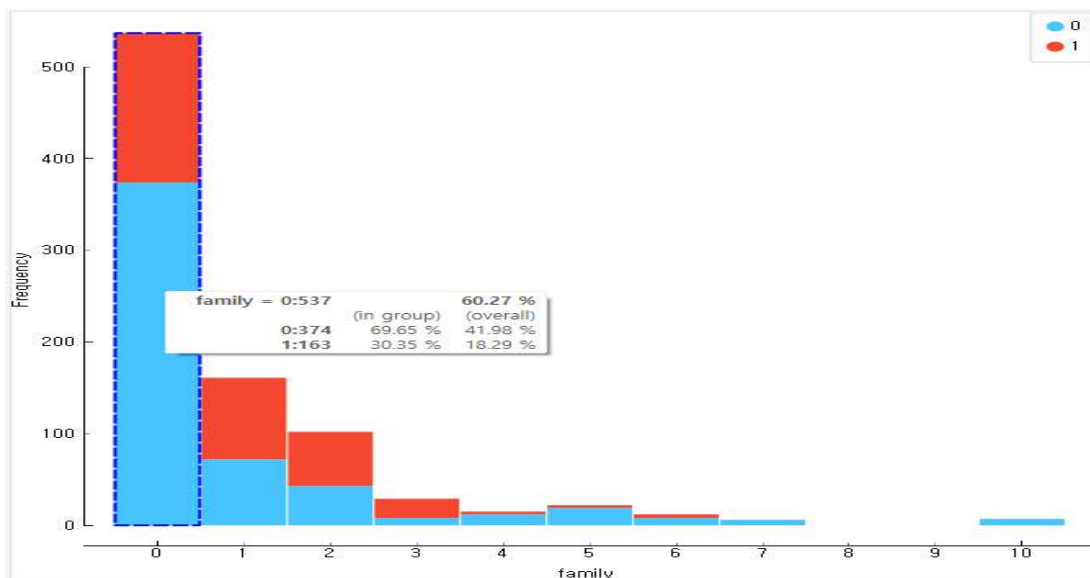
- 1) 타이타닉호에서 남성과 여성 중에 어느 성별이 더 많이 생존했는가? 그 비율은 어떤가?
 - Visualize / Distributions로 성별 생존률을 보이고, 남성의 생존률, 여성의 생존률을 비교하라.
 (생존률을 보일 때, Stack Columns와 Show probabilities를 선택하면 훨씬 보기 좋다.)



▶ 여성 생존률은 74.20%, 남성은 생존률 18.89%로 여성이 더 높은 생존률을 보인다.

- 2) 함께 탑승한 부모, 자녀, 형제자매, 배우자의 수가 많을수록 생존에 도움이 되었는가? 그 이유는?

가족 수	0	1	2	3	4	5	6	7	8	9	10
생존률(%)	30.3	55.2	57.8	72.4	20.0	13.6	33.3	0.0	-	-	0.0



▶ 가족수별 생존률을 보면, 가족이 없는 경우에는 30.3%의 생존률을 보인다. 하지만 가족이 1~3명까지는 55.2~72.4%까지 생존률이 높아졌다가 4명 이상의 경우 낮은 생존률을 보인다. 이로 보건대, 가족이 없는 것보다 3명 정도 있을 때는 생존률이 높아진다고 볼 수 있다. 하지만 더 많아지는 경우에는 생존률이 급격히 떨어져, 단순히 가족 수가 많을수록 생존률이 높아진다고 보기는 어렵다.

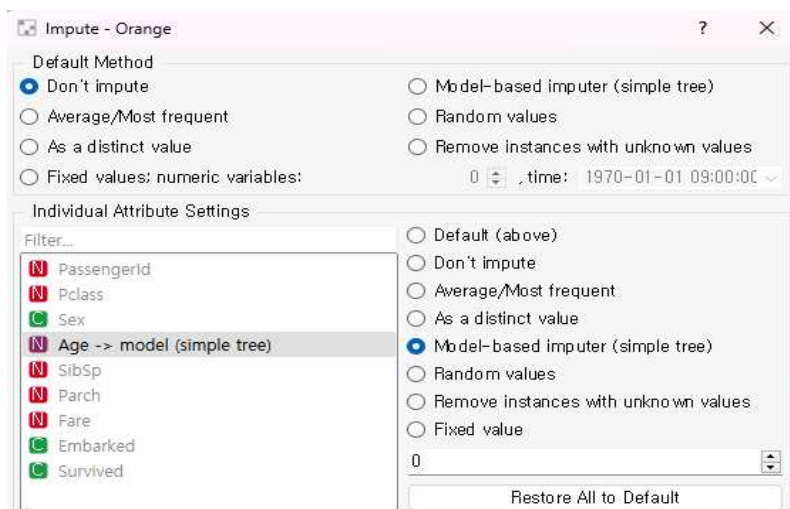
3) 예측 결과의 품질을 높이기 위해서 어떤 전처리가 필요한가? 어떻게 할 수 있는가?

- Impute: Age 변수의 Missing Value를 기존 값들도 채워보라. (Impute 설정화면 제시)
- 이외에 최소한 하나 이상의 전처리를 하고, 그 결과를 설명하라.

① don't impute

Model	AUC	CA	F1	Precision	Recall
kNN	0.615	0.650	0.601	0.639	0.650
Tree	0.826	0.795	0.791	0.793	0.795
Neural Network	0.867	0.816	0.812	0.816	0.816
Logistic Regression	0.850	0.796	0.794	0.794	0.796

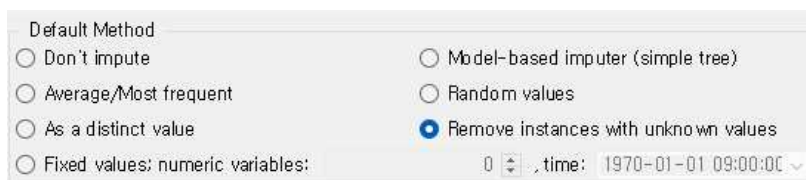
② Age - Model-based imputer



Model	AUC	CA	F1	Precision	Recall
kNN	0.626	0.653	0.606	0.644	0.653
Tree	0.823	0.790	0.788	0.788	0.790
Neural Network	0.869	0.822	0.818	0.822	0.822
Logistic Regression	0.855	0.796	0.794	0.794	0.796

▶ model-based imputer는 확률 분포상에서 sampling 하는 방법으로, 시행 후 Test and Score의 표에서 확인 할 수 있듯, Tree model을 제외하고는 전반적으로 조금씩 높아진 성능을 확인할 수 있다.

③ Default Method - Remove instances with unknown values



Model	AUC	CA	F1	Precision	Recall
kNN	0.622	0.654	0.610	0.642	0.654
Tree	0.844	0.801	0.798	0.799	0.801
Neural Network	0.863	0.823	0.819	0.824	0.823
Logistic Regression	0.854	0.796	0.795	0.795	0.796

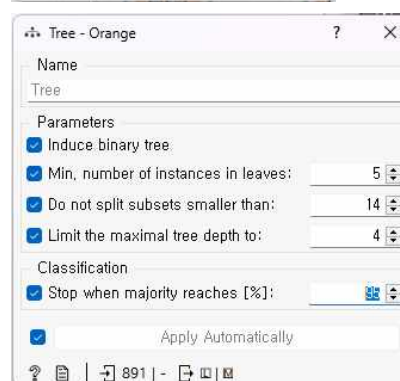
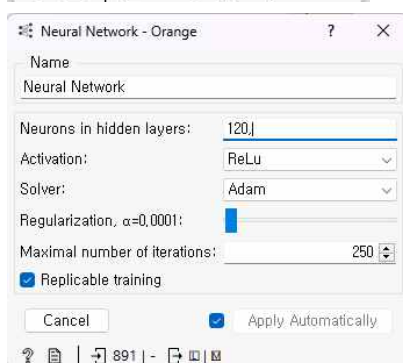
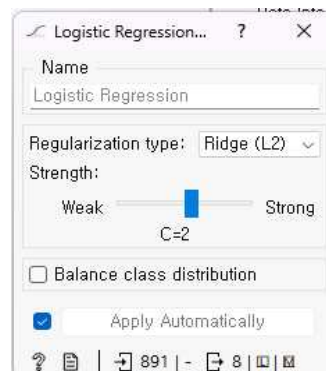
▶ 모든 결측치에 대해 제거해주는 옵션을 선택한 경우, Tree model의 예측 성능을 높여줄 수 있다.

- 4) 위에서 전처리한 데이터셋에 다음 학습 모델들을 함께 적용하고, 성능을 비교 분석하라.
- kNN, Logistic Regression, (Decision) Tree, Neural Network
 - 전처리를 통해 얻은 최선의 결과를 얻을 수 있는 적절한 feature들을 선택하라.
 - 각 모델별로 최선의 성능을 얻는 파라미터를 찾고, 설정 화면을 캡처하여 제시하라.
 - Test and Score와 Confusion Matrix의 결과 화면을 캡처하여 제시하라.

① 적절한 feature 선택 -> PassengerId, Embarked 제거

		#	Gain ratio	ReliefF
1	Sex	2	0.231	0.036
2	Pclass		0.057	0.066
3	Fare		0.033	0.006
4	SibSp		0.024	0.009
5	Embarked	3	0.019	0.018
6	Parch		0.018	0.010
7	PassengerId		0.002	0.036
8	Age		0.001	0.022

② 모델별 파라미터 셋팅



③ Test and Score

** 전처리 전

Model	AUC	CA	F1	Precision	Recall
kNN	0.622	0.654	0.610	0.642	0.654
Tree	0.844	0.801	0.798	0.799	0.801
Neural Network	0.863	0.823	0.819	0.824	0.823
Logistic Regression	0.854	0.796	0.795	0.795	0.796

** 파라미터 셋팅 후

Model	AUC	CA	F1	Precision	Recall
kNN	0.816	0.764	0.759	0.762	0.764
Tree	0.854	0.825	0.823	0.824	0.825
Neural Network	0.863	0.827	0.824	0.827	0.827
Logistic Regression	0.851	0.798	0.796	0.796	0.798

⑤ Confusion Matrix

** KNN

		Predicted		
		0	1	Σ
Actual	0	476	73	549
	1	137	205	342
Σ		613	278	891

** Logistic Regression

		Predicted		
		0	1	Σ
Actual	0	472	77	549
	1	103	239	342
Σ		575	316	891

** Tree

		Predicted		
		0	1	Σ
Actual	0	487	62	549
	1	94	248	342
Σ		581	310	891

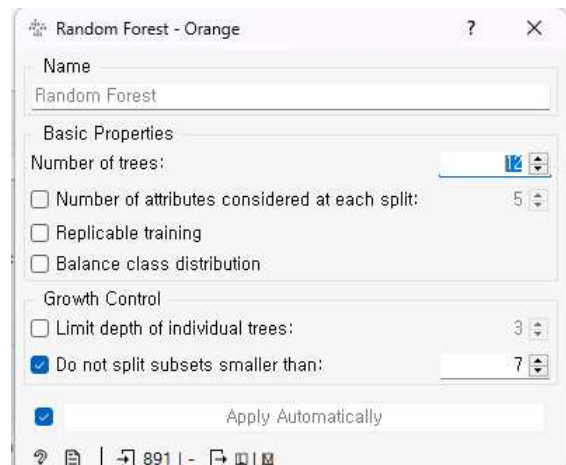
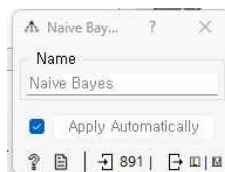
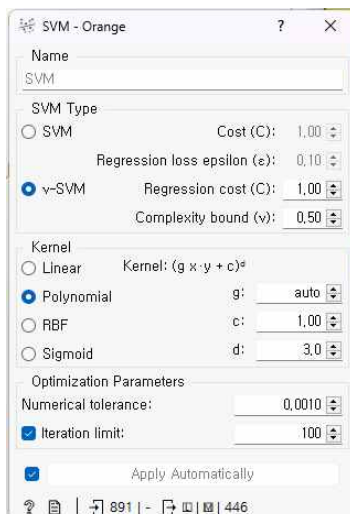
** Neural Network

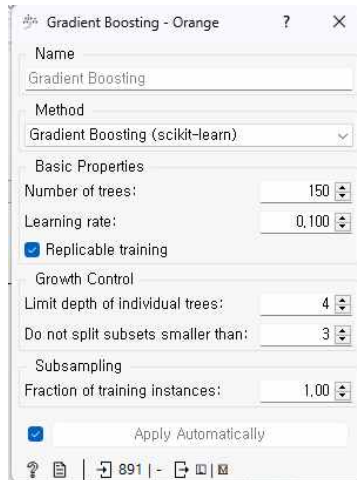
		Predicted		
		0	1	Σ
Actual	0	499	50	549
	1	105	237	342
Σ		604	287	891

▶ feature를 선택할 때, PassengerId와 Embarked 변수를 제거했더니, KNN model의 성능이 급격하게 높아졌으나, 나머지 model에서는 조금씩 떨어졌다. 분류 모델의 성능 평가 지표들의 전반적인 점수는 전처리 전, 후 모두 Neural Network model이 가장 높은 성능을 보여주고 있다. Parameter setting에서 Hidden layer의 수를 120으로 설정해주니 최적의 성능을 보여주었다. 하지만 모델의 특성상 많은 시간이 필요했다.

5) 추가로, 다음 학습 모델들을 함께 적용하고, 성능을 비교 분석하라.

- SVM(Support Vector Machine), Naive Bayes, Random Forest, GradientBoosting
- 각 모델별로 설정 화면을 캡처하여 제시하라. (파라미터는 디폴트로 두어도 된다.)





- Test and Score와 Confusion Matrix의 결과 화면을 캡처하여 제시하라.

Evaluation results for target (None, show average over classes)						
Model	AUC	CA	F1	Precision	Recall	
Gradient Boosting	0.872	0.835	0.833	0.834	0.835	
Random Forest	0.872	0.836	0.834	0.835	0.836	
Neural Network	0.862	0.815	0.812	0.814	0.815	
Logistic Regression	0.851	0.799	0.797	0.797	0.799	
Tree	0.849	0.814	0.812	0.812	0.814	
SVM	0.828	0.806	0.803	0.804	0.806	
Naive Bayes	0.821	0.751	0.749	0.749	0.751	
kNN	0.814	0.758	0.752	0.754	0.758	

**SVM(Support Vector Machine)

		Predicted		Σ
		0	1	
Actual	0	482	67	549
	1	106	236	342
Σ		588	303	891

**Naive Bayes

		Predicted		Σ
		0	1	
Actual	0	447	102	549
	1	120	222	342
	Σ	567	324	891

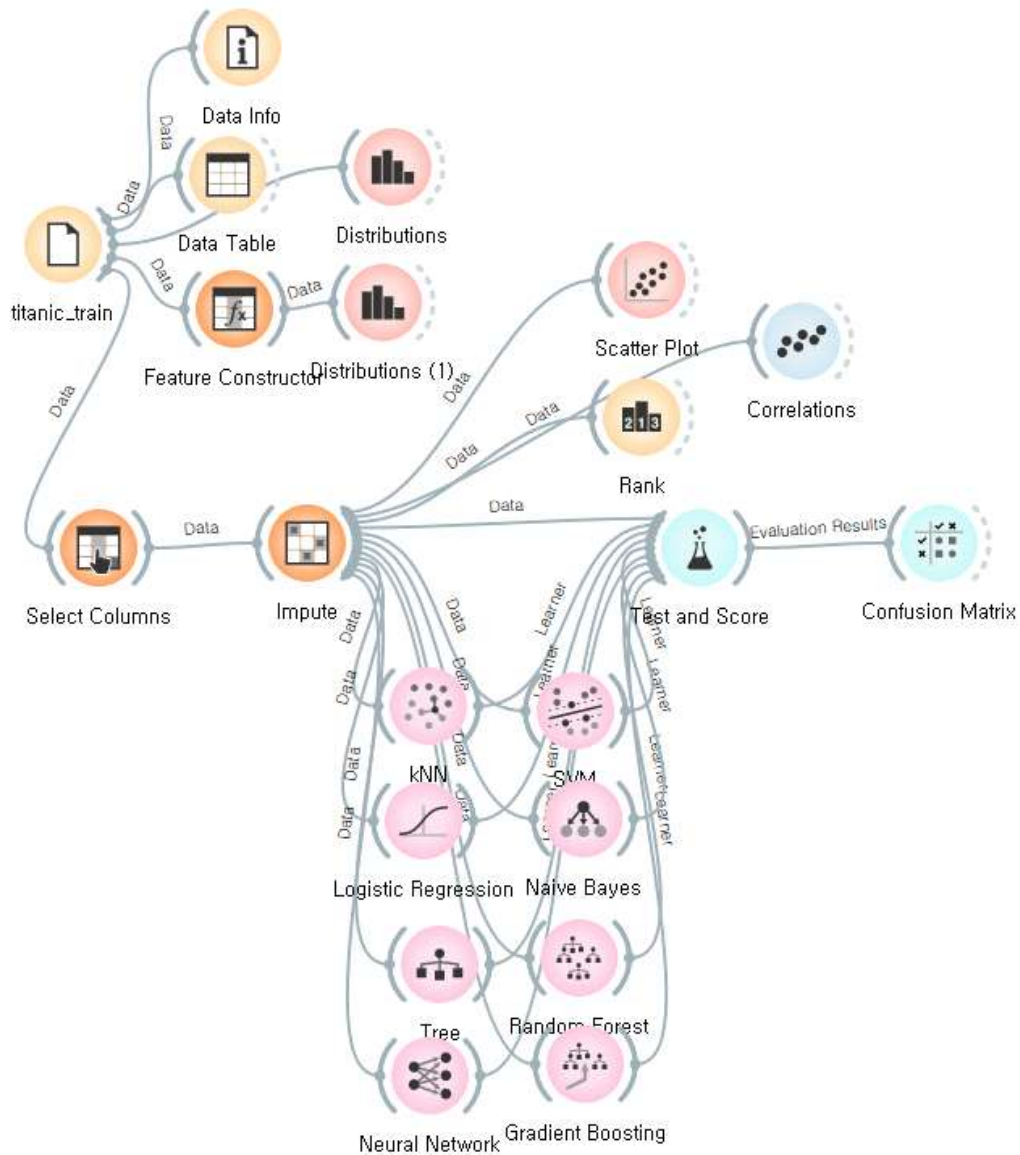
**Random Forest

		Predicted		Σ
		0	1	
Actual	0	493	56	549
	1	90	252	342
	Σ	583	308	891

**GradientBoosting

		Predicted		Σ
		0	1	
Actual	0	490	59	549
	1	88	254	342
Σ		578	313	891

▶ Random Forest와 GradientBoosting이 가장 높은 Test Score를 기록했다. 타이타닉에서 생존을 예측하는 것은 완전한 운의 영역이라 생각했었는데, 머신러닝을 통해 타이타닉 생존자 예측을 이렇게 높은 확률로 예측이 가능하다는 것이 대단하게 느껴진다.



3. 기타

- 1) 위 실습 과제를 참고하여, “캐글 자전거 수요 및 타이타닉 분석 보고서”를 작성하여 제출할 것.
- 2) 보고서 내용에 위 실습 과제의 **파란색 질문에 대한 답은 반드시 포함**되어야 함.
- 3) 보고서의 분량과 형식은 자유이나, 반드시 **LMS에 PDF 파일로 제출**할 것. (마감기한 준수!)
- 4) Orange를 사용할 것을 권장하지만, **R이나 Python을 사용해도 무방**함. (가산점은 없음)
- 5) 본인이 스스로 분석해서 작성하였음을 알 수 있도록 **화면 캡처, 소스 코드 등을 반드시 포함**할 것.
(질문의 답에만 간단히 대답하는 성의가 없는 보고서는 감점할 수도 있음)
(그렇다고 화려하게 많은 분량을 작성한다고 해서 가산점도 없음)