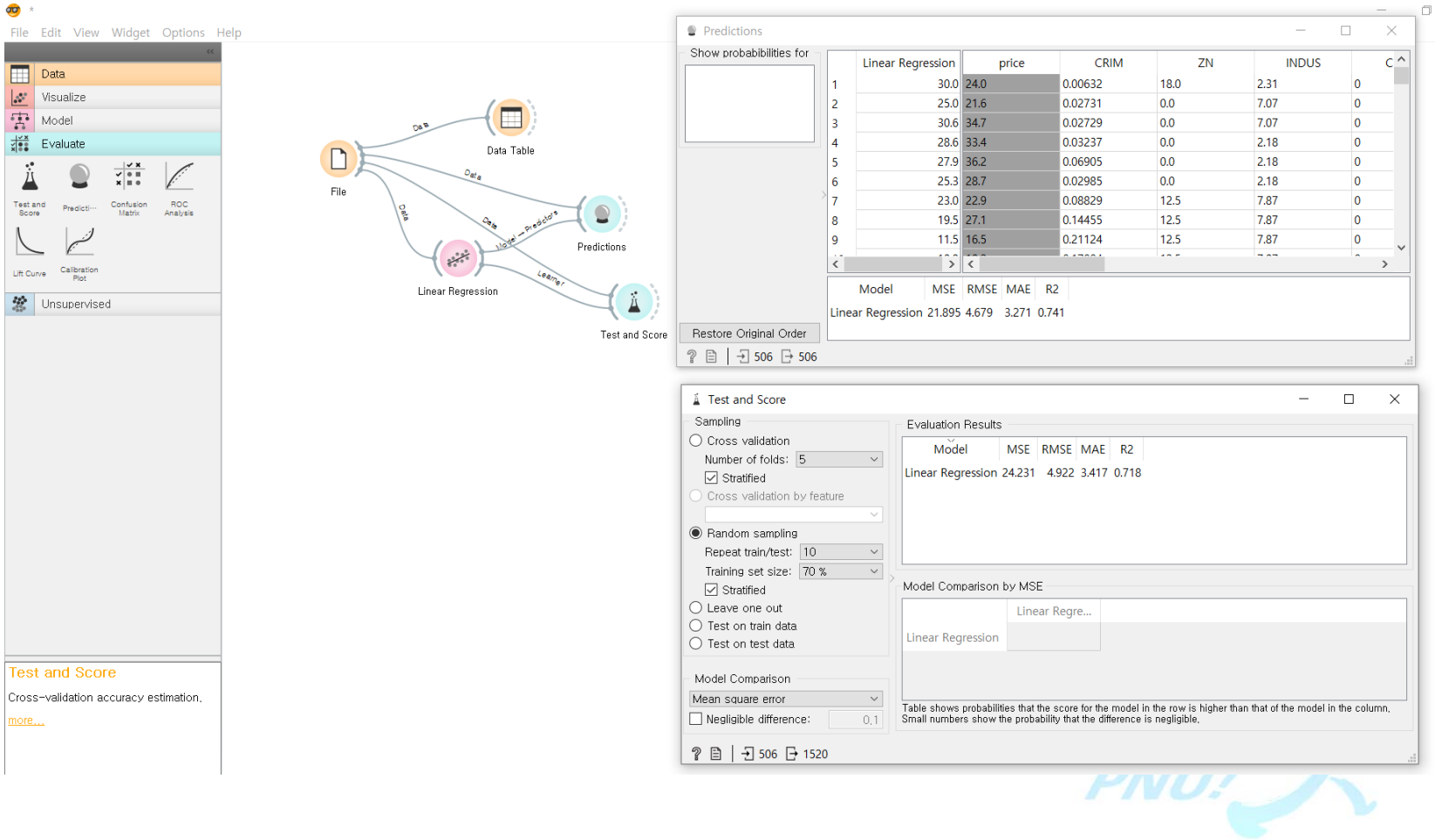
지도학습-회귀모델

Boston housing price

- **평균 주택가격 예측**
  - 13개의 특성(feature)

  - CRIM : 근방 범죄율
  - ZN : 주택지 비율
  - INDUS : 상업적 비즈니스에 활용되지 않는 농지 면적
  - CHAS : 경계선에 강에 있는지 여부
  - NOX : 산화 질소 농도
  - RM : 자택당 평균 방 갯수
  - AGE : 1940 년 이전에 건설된 비율
  - DIS : 5 개의 보스턴 고용 센터와의 거리에 다른 가중치 부여
  - RAD : radial 고속도로와의 접근성 지수
  - TAX : 10000달러당 재산세
  - PTRATIO : 지역별 학생-교사 비율
  - B : 지역의 흑인 지수 (1000(B - 0.63)^2), B는 흑인의 비율.
  - LSTAT : 빈곤층의 비율
  - price : 1978년 보스턴 주택 가격, 506개 주택 가격 중앙값(단위 천달러)

오렌지3를 활용한 분석


File Edit View Widget Options Help
Data
Visualize
Model
Evaluate
Unsupervised
Test and Score
Cross-validation accuracy estimation.
more...
File
Data Table
Predictions
Linear Regression
Test and Score
Predictions
Show probabilities for
Restore Original Order
Linear Regression price CRIM ZN INDUS
1 30.0 24.0 0.00632 18.0 2.31 0
2 25.0 21.6 0.02731 0.0 7.07 0
3 30.6 34.7 0.02729 0.0 7.07 0
4 28.6 33.4 0.03237 0.0 2.18 0
5 27.9 36.2 0.06905 0.0 2.18 0
6 25.3 28.7 0.02985 0.0 2.18 0
7 23.0 22.9 0.08829 12.5 7.87 0
8 19.5 27.1 0.14455 12.5 7.87 0
9 11.5 16.5 0.21124 12.5 7.87 0
Model MSE RMSE MAE R2
Linear Regression 21.895 4.679 3.271 0.741
506 506
Test and Score
Sampling
Cross validation
Number of folds: 5
Stratified
Cross validation by feature
Random sampling
Repeat train/test: 10
Training set size: 70 %
Stratified
Leave one out
Test on train data
Test on test data
Model Comparison
Mean square error
Negligible difference: 0.1
Evaluation Results
Model MSE RMSE MAE R2
Linear Regression 24.231 4.922 3.417 0.718
Model Comparison by MSE
Linear Regre...
Linear Regression
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.
506 1520

회귀 모델 평가지표

- **종류**
  - **MAE (Mean Absolue Error)**
    - 실제 값과 예측 값의 차이를 절댓값으로 변환해 평균한 것
  - **MSE (Mean Squared Error)**
    - 실제 값과 예측 값의 차이를 제곱해 평균한 것
  - **RMSE (Root Mean Squared Error)**
    - MSE 값은 오류의 제곱을 구하므로 실제 오류 평균보다 더 커지는 특성이 있어 MSE에 루트를 씌운 값
- 값이 작을수록 예측값과 실제값의 차이가 없다는 뜻

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

파이썬을 활용한 분석

- 판다스(pandas)
  - 데이터 분석을 위한 파이썬 라이브러리
- 사이킷런(scikit-learn)
  - 머신러닝을 위한 파이썬 라이브러기

파이썬을 활용한 분석

- 파이썬에서 머신러닝 분석을 할 때 유용하게 사용할 수 있는 라이브러리

사이킷런 회귀모델

- 데이터 수집
  - 라이브러리 추가
    - from sklearn.datasets import load_boston
  - 학습모델에 사용가능한 데이터프레임만들기
    - df = pd.DataFrame(data=boston.data, columns=boston.feature_names)
    - df['price'] = boston.target

사이킷런 회귀모델

- **학습데이터 분리**
  - **from sklearn.model_selection import train_test_split**
    - Train, Test 데이터 분리하기
    - x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=10)

사이킷런 회귀모델

- **선형회귀모델 학습**
  - 라이브러리 추가
    - from sklearn.linear_model import LinearRegression
  - 모델 만들기
    - model = LinearRegression()
  - 학습하기
    - model.fit(x_train, y_train)

사이킷런 회귀모델

- 평가하기
  - 라이브러리 추가
    - from sklearn.metrics import mean_squared_error
    - import numpy as np
  - 예측하기
    - y_train_predict = model.predict(x_train)
  - 평가하기
    - rmse = (np.sqrt(mean_squared_error(y_train, y_train_predict)))