변수 선택(Feature Selection)

- 다중공선성(Multicollinearity)

  독립변수들 간에 강한 상관관계가 나타나는 문제

  예측 값의 분산이 커짐(회귀 모형의 적합성이 떨어짐)

  결정 계수의 값이 과대하게 나타날 수 있음

  설명력은 좋으나 예측력이 떨어질 수 있음

  다중공선성 문제를 해소하기 위해 변수 선택법 적용 필요

변수 선택(Feature Selection)

- 전진 선택법(Forward Selection)

  설명 변수가 하나도 없는 모델(Null Model)에서부터 시작

  가장 유의미한 변수를 하나씩 추가해 나가는 방법 (F-통계량 사용)

  한번 선택된 변수는 제거되지 않음

- 후진 소거법(Backward Elimination)

  모든 변수를 사용하여 구축한 모델(Full Model)에서 유의미하지 않은 변수를 하나씩 제거해 나가는 방법

  한번 제거된 변수는 다시 선택될 가능성이 없음

변수 선택(Feature Selection)

- 단계적 선택법(Stepwise Selection)

  설명 변수가 하나도 없는 모델에서부터 시작하여 전진선택법과 후진소거법을 번갈아 가며 수행

  전진 선택법 및 후진 소거법에 상대적으로 시간이 많이 소요되지만
  보다 우수한 예측 성능을 나타내는 변수 집합을 찾아낼 가능성이 높음

  한번 선택되거나 제거된 변수라도 다시 선택/제거될 가능성이 있음

  변수의 수는 초기에는 일반적으로 증가하나 중반 이후에는 증가와 감소를 반복

변수 선택(Feature Selection)

- 변수 선택법 적용 – SelectKBest

그룹 내 분산이 작고 그룹 간 분산이 클 경우 값이 커지는 F-value를 이용하여 변수를 선택

각 변수마다 F값을 구해 F값이 큰 변수를 기준으로 변수를 선택하는 방법

```
from sklearn.feature_selection import SelectKBest, f_classif

selector = SelectKBest(score_func=f_classif, k=10)
```

```
selector.fit(boston.data,boston.target)
```

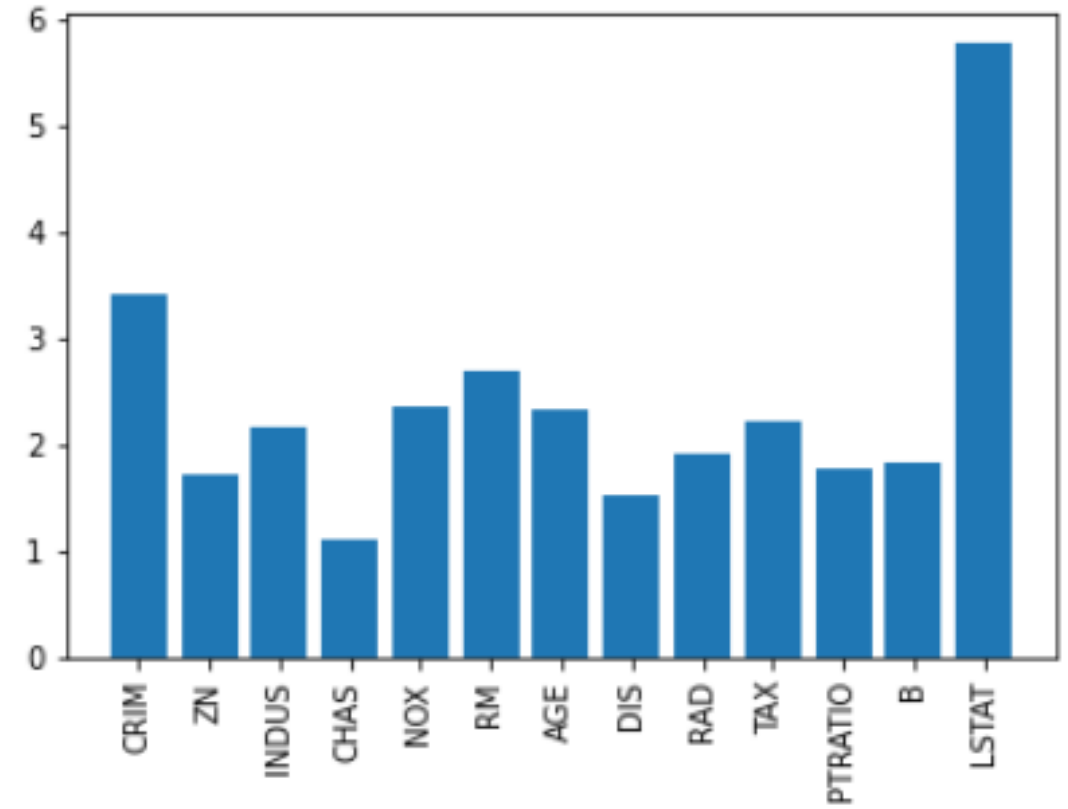
```
SelectKBest()
```

```
selector.scores_
```

```
array([3.41392257, 1.69825318, 2.15122   , 1.1023291 , 2.33899533,
       2.6759978 , 2.32954454, 1.50668687, 1.91464853, 2.21005614,
       1.77625065, 1.81833191, 5.75215088])
```

변수 선택(Feature Selection)

- 변수 선택법 적용 – SelectKBest



	column	score
12	LSTAT	5.752151
0	CRIM	3.413923
5	RM	2.675998
4	NOX	2.338995
6	AGE	2.329545
9	TAX	2.210056
2	INDUS	2.151220
8	RAD	1.914649
11	B	1.818332
10	PTRATIO	1.776251
1	ZN	1.698253
7	DIS	1.506687
3	CHAS	1.102329

변수 선택(Feature Selection)

- 변수 선택법 적용 – RFE(Recursive Feature Elimination)

Backward 방식중 하나

모든 변수를 우선 다 포함시킨 후 반복해서 학습을 진행하면서 중요도가 낮은 변수를 하나씩 제거

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
```

```
from sklearn.feature_selection import RFE
```

```
rfe = RFE(model)
rfe = rfe.fit(boston.data, boston.target)
```

```
pd.DataFrame({'feature' : boston.feature_names,
              '선택여부' : rfe.support_,
              'ranking' : rfe.ranking_}).sort_values('ranking')
```

변수 선택(Feature Selection)

- 변수 선택법 적용 – RFE(Recursive Feature Elimination)

	feature	선택여부	ranking
3	CHAS	True	1
4	NOX	True	1
5	RM	True	1
7	DIS	True	1
10	PTRATIO	True	1
12	LSTAT	True	1
8	RAD	False	2
0	CRIM	False	3
2	INDUS	False	4
1	ZN	False	5
9	TAX	False	6
11	B	False	7
6	AGE	False	8

08-3. feature_selection.ipynb 실습