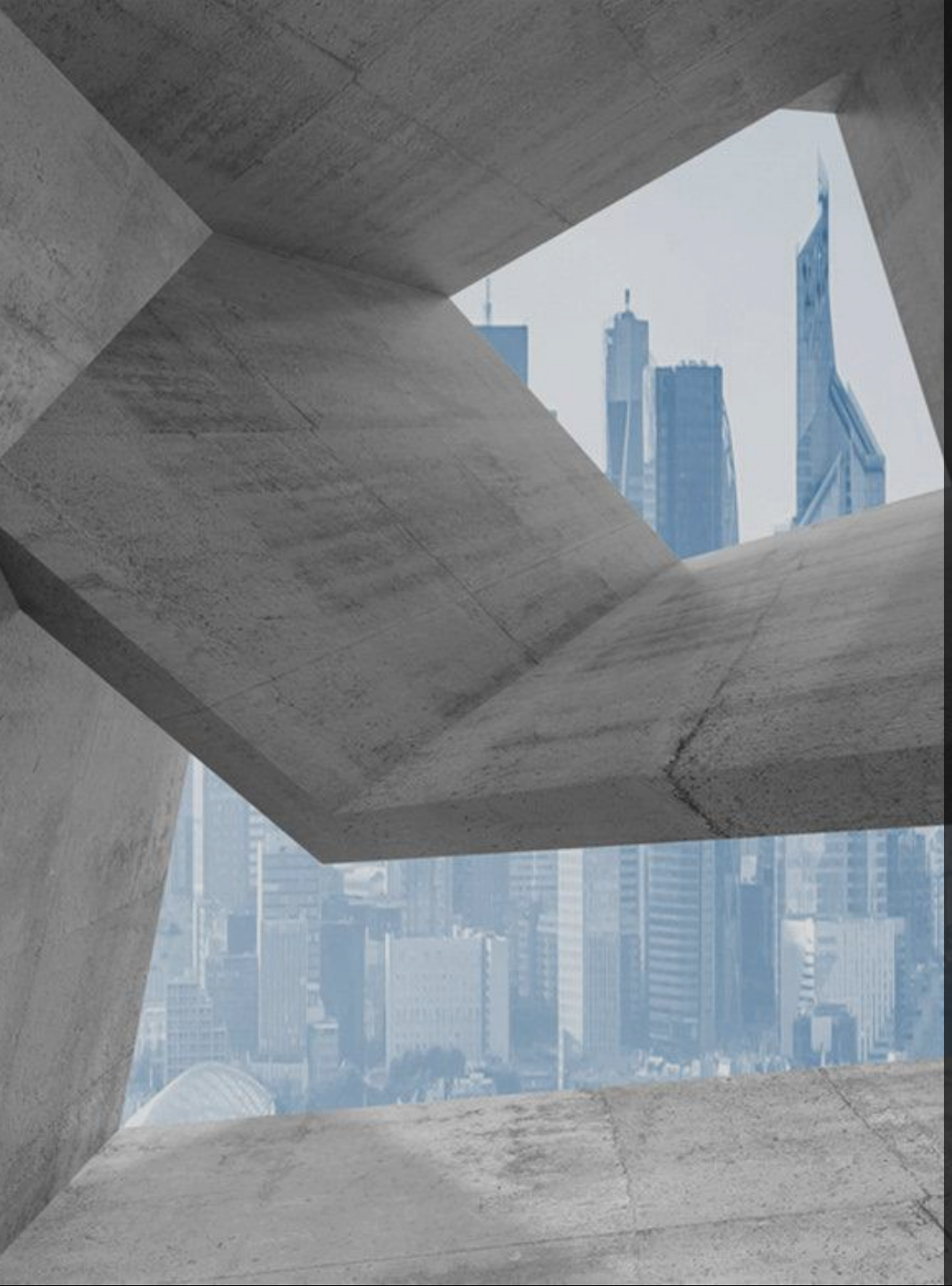


2019.01.25 빅데이터 핀테크 과정 2기

콘크리트의 압축강도에 미치는 영향인자에 대한 고찰

5조

노진현 문명진 박지연 이한얼

KNOW YOUR CONCRETE

8월 14일, 이탈리아 제노아
다리 붕괴 참사, 35명 사망

“강화 콘크리트, 일반 교량과 다른 배합 비율을 가져 계산 착오..

바닷바람과 산업단지의 오염물질이 구조를 더 약하게 해”

– 구조공학 전문가 브렌치치 교수

https://www.youtube.com/watch?v=b7WeoT46Ve4

Our Goal

콘크리트의 압축강도는 재령(age)와 최소 7가지 재료(ingredient)의 비선형함수로 표현된다.
설명변수는 8가지 (Discrete 변수 1개와 Float 변수 7개)인데 데이터셋은 단 1,030개뿐이다.

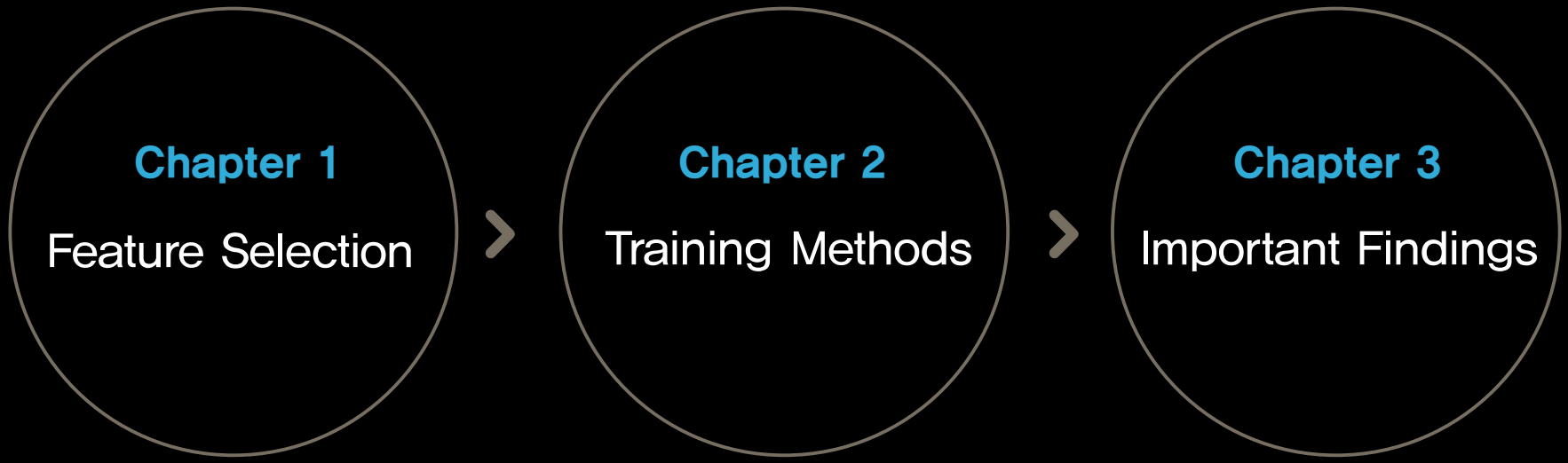
```
Concrete Compressive Strength

---------------------------------

Data Type: multivariate

Abstract: Concrete is the most important material
in civil engineering. The
concrete compressive strength is a highly
nonlinear function of age and
ingredients. These ingredients include cement,
blast furnace slag, fly ash,
water, superplasticizer, coarse aggregate, and
fine aggregate.
```

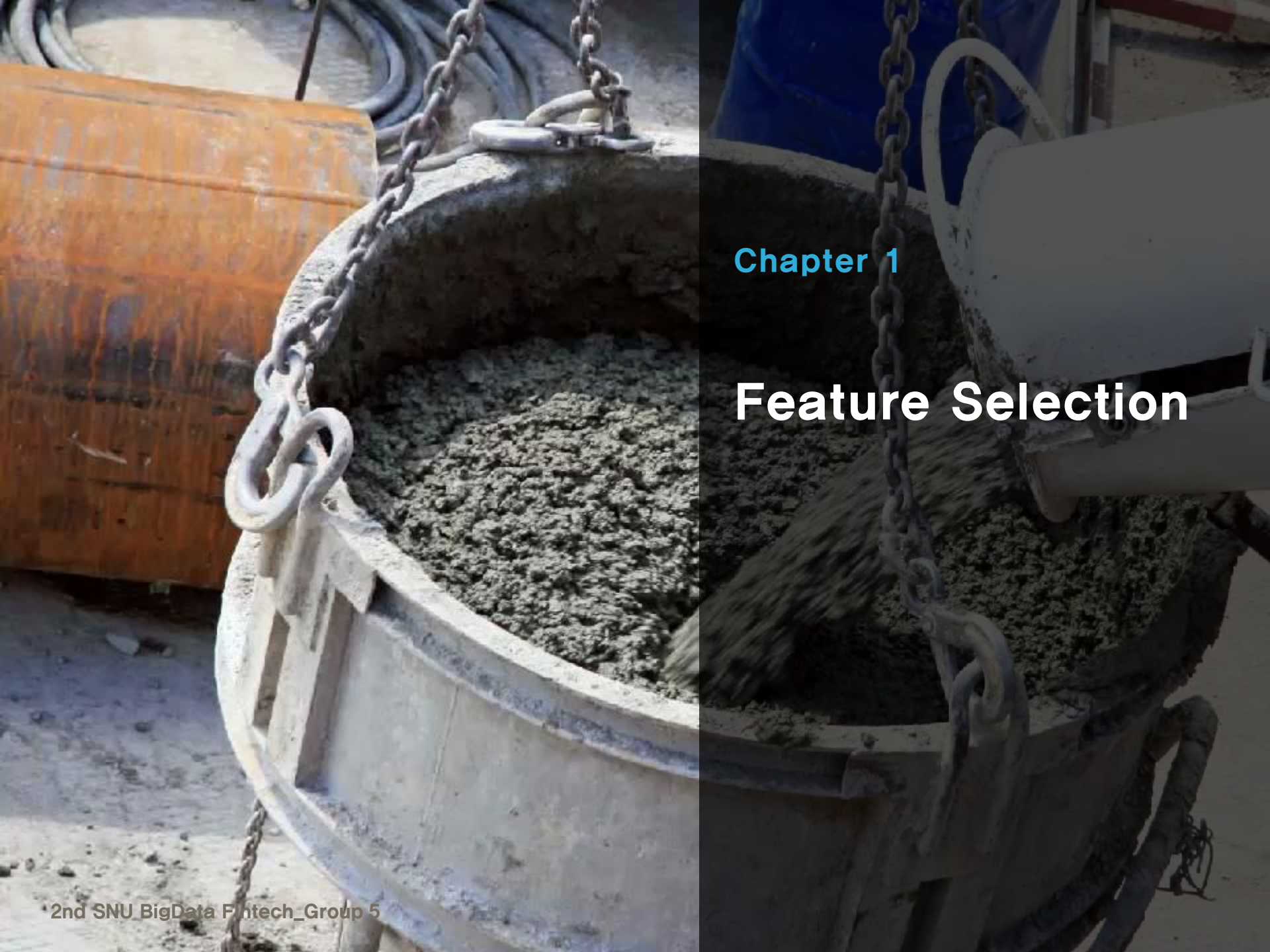
```
Data Characteristics:

The actual concrete compressive strength (MPa) for
a given mixture under a
specific age (days) was determined from
laboratory. Data is in raw form (not scaled).

Summary Statistics:

Number of instances (observations): 1030
Number of Attributes: 9
Attribute breakdown: 8 quantitative input
variables, and 1 quantitative output variable
Missing Attribute Values: None

---------------------------------
```

어떻게 하면 콘크리트 압축강도를 정확하게 추정해서 안전을 보장할까?

•출처 https://archive.ics.uci.edu/ml/machine-learning-databases/concrete/compressive/Concrete_Readme.txt

Journey


Chapter 1
Feature Selection
Chapter 2
Training Methods
Chapter 3
Important Findings



Chapter 1

Feature Selection

Feature Transformation – I

물/시멘트비 증감은 콘크리트 **압축강도 저하요인 11가지* 중 1번째**로 큰 요인이다.
콘크리트 설계 과정의 모수로서 온습도 등 외부 요인이나 계량오차에도 **덜 민감하다.**

두 개 독립 변수 –〉 W/C 변수로 변환 `data0['w/c'] = data['water']/data['cement']`

	cement	slag	flyash	water	superplasticizer	coarseaggregate	fineaggregate	age	csMPa
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.986111
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.887366
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.269535
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.052780
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.296075

표 6. 콘크리트 압축강도 저하

저하요인	
1.물/시멘트비 증가	▸양질의 원재료 사용 ▸골재표면수 안정화
2.공기량 증가	▸모래입도, 점토량의 안정화

	cement	slag	flyash	water	superplasticizer	coarseaggregate	fineaggregate	age	w/c	age_c	age_f	age_class_1	age_class_2	age_class_3	csMPa
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	0.300000	class_2	2	0	1	0	79.99
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	0.300000	class_2	2	0	1	0	61.89
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	0.685714	class_3	3	0	0	1	40.27
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	0.685714	class_3	3	0	0	1	41.05
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	0.966767	class_3	3	0	0	1	44.30

•출처 [감수율에 따른 압축강도와 물–시멘트비 관계에 관한 연구], 한국콘크리트학회 : 콘크리트학회 논문집 26권 5호, 1–2

Feature Transformation – II

28일 재령강도는 표준적인 양생환경에서 콘크리트의 품질을 나타내는 공식 지표이다.
한 달 전까지는 압축강도가 나날이 큰 폭으로 증가하다 28일 기점으로 안정화되기 때문이다.

단일 이산형 변수 –〉 3개 범주형 변수로 변환

```
for i in data['age']:  if i 〈 28 :
                       data1['age'][cnt] = 'class_1'; (...)
```

	cement	slag	flyash	water	superplasticizer	coarseaggregate	fineaggregate	age	csMPa
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.986111
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.887366
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.269535
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.052780
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.296075

배합설계

- 필요한 모든 재료들의 성질을 잘 파악하고 있어야 하다
  하여 가장 효율적인 배합을 해야 함

설계강도(f_{ck}) 및 배합강도(f_{cr})의 결정

- 배합강도는 현장 콘크리트가 설계기준 강도 이하가 될
  하로 되는 확률은 0.13% 이하가 되게 충분히 크게 할
- 배합강도(f_{cr})는 보통 28일 재령 압축강도(f_{28})로 정함

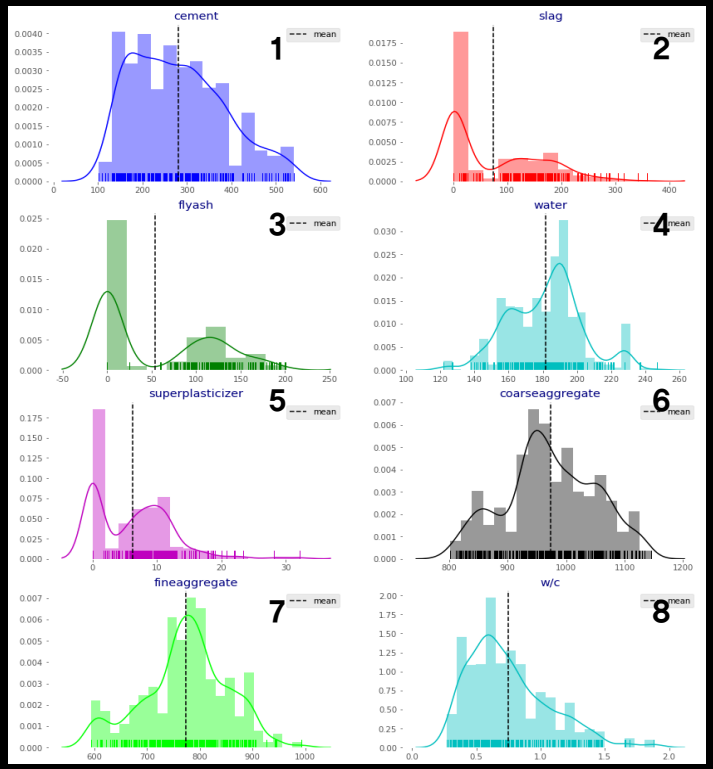
	cement	slag	flyash	water	superplasticizer	coarseaggregate	fineaggregate	age	w/c	age_c	age_f	age_class_1	age_class_2	age_class_3	csMPa
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	0.300000	class_2	2	0	1	0	79.99
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	0.300000	class_2	2	0	1	0	61.89
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	0.685714	class_3	3	0	0	1	40.27
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	0.685714	class_3	3	0	0	1	41.05
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	0.966767	class_3	3	0	0	1	44.30

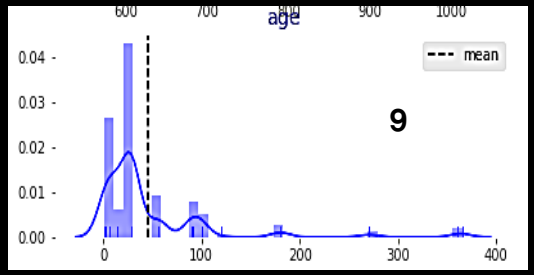
•출처 [감수율에 따른 압축강도와 물–시멘트비 관계에 관한 연구], 한국콘크리트학회 : 콘크리트학회 논문집 26권 5호, 1–2

Exploratory Data Analysis – I

차원 축소에 앞서 새로 변환한 변수를 포함, 10개 변수의 분포의 특성을 시각화해보았다.
Null값이 없었고, age 데이터가 실제로 3개 구간에 균일하게 분포해, 더미변수화에 유리했다.

New Variable
1 cement
2 slag
3 flyash
4 water
5 super plasticizer
6 coarse aggregate
7 fine aggregate
8 w/c
9 age
y = csMPa





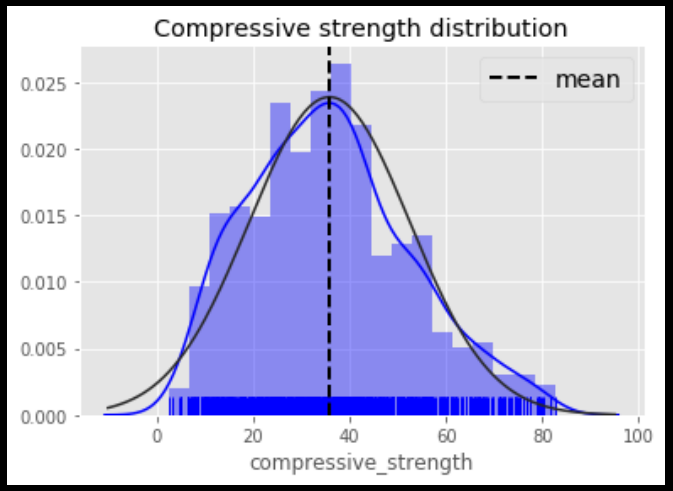
x 〈 28 class_1 (324개)
x = 28 class_2 (425개)
x 〉 28 class_3 (281개)

```
cnt = 0;
for i in data1['age']:
    if i < 28 :
            data1['age'][cnt] = 'class_1';
    elif i == 28 :
            data1['age'][cnt] = 'class_2';
    else :
            data1['age'][cnt] = 'class_3';
    cnt+=1;
```

Exploratory Data Analysis – II

csMPa, 즉 타겟변수인 콘크리트 압축강도의 경우 그 분포가 정규분포곡선에 근접하고 있다.
설명변수를 통해 회귀추정하기에 이미 적합하므로, y는 로그/큐브 등 추가변환하지 않았다.

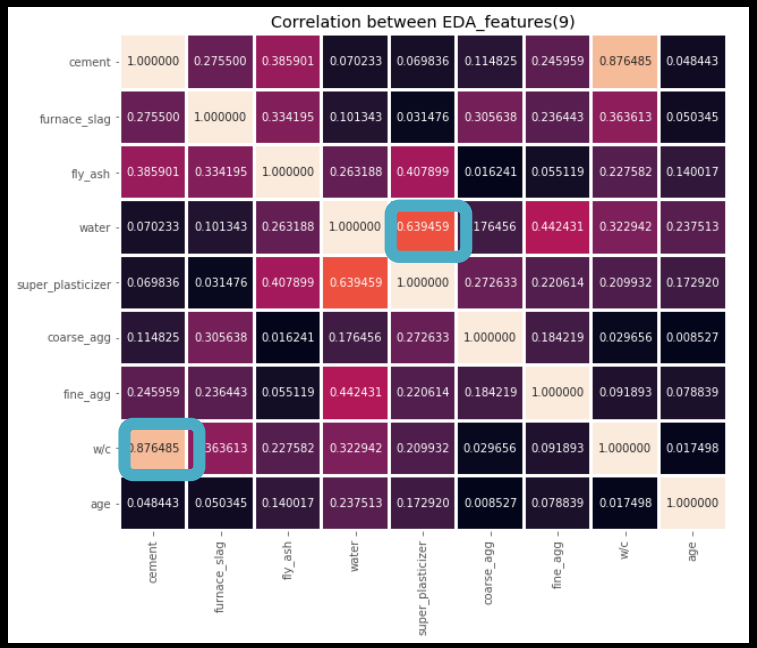
	count	mean	std	min	25%	50%	75%	max
cement	1030.0	281.167864	104.506364	102.000000	192.375000	272.900000	350.000000	540.000000
slag	1030.0	73.895825	86.279342	0.000000	0.000000	22.000000	142.950000	359.400000
flyash	1030.0	54.188350	63.997004	0.000000	0.000000	0.000000	118.300000	200.100000
water	1030.0	181.567282	21.354219	121.800000	164.900000	185.000000	192.000000	247.000000
superplasticizer	1030.0	6.204660	5.973841	0.000000	0.000000	6.400000	10.200000	32.200000
coarseaggregate	1030.0	972.918932	77.753954	801.000000	932.000000	968.000000	1029.400000	1145.000000
fineaggregate	1030.0	773.580485	80.175980	594.000000	730.950000	779.500000	824.000000	992.600000
age	1030.0	45.662136	63.169912	1.000000	7.000000	28.000000	56.000000	365.000000
w/c	1030.0	0.748266	0.314005	0.266893	0.533333	0.675349	0.935165	1.882353
age_f	1030.0	1.958252	0.765640	1.000000	1.000000	2.000000	3.000000	3.000000
age_class_1	1030.0	0.314563	0.464567	0.000000	0.000000	0.000000	1.000000	1.000000
age_class_2	1030.0	0.412621	0.492545	0.000000	0.000000	0.000000	1.000000	1.000000
age_class_3	1030.0	0.272816	0.445623	0.000000	0.000000	0.000000	1.000000	1.000000
csMPa	1030.0	35.817961	16.705742	2.330000	23.710000	34.445000	46.135000	82.600000

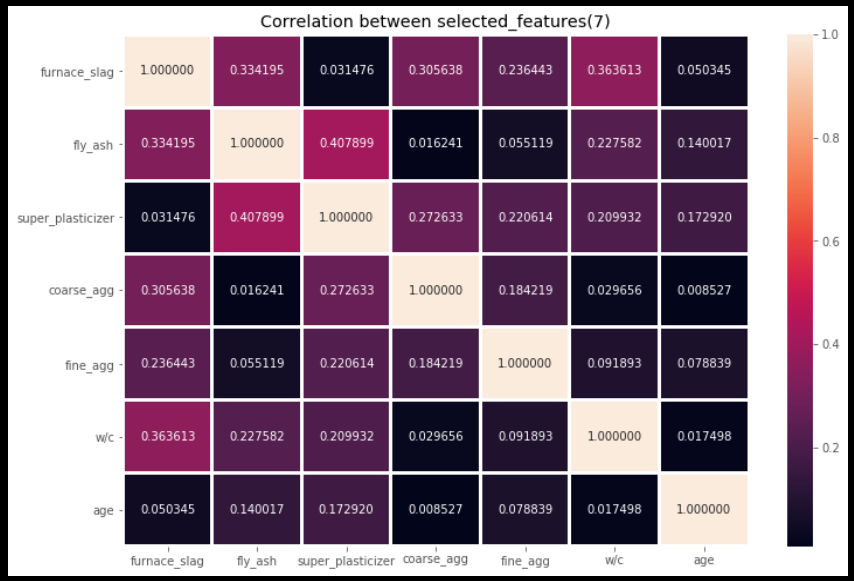


이상 9가지 변수 중에 어떤 변수를 선택하면 간단명료하게 회귀추정할 수 있을까?

Feature Selection – I

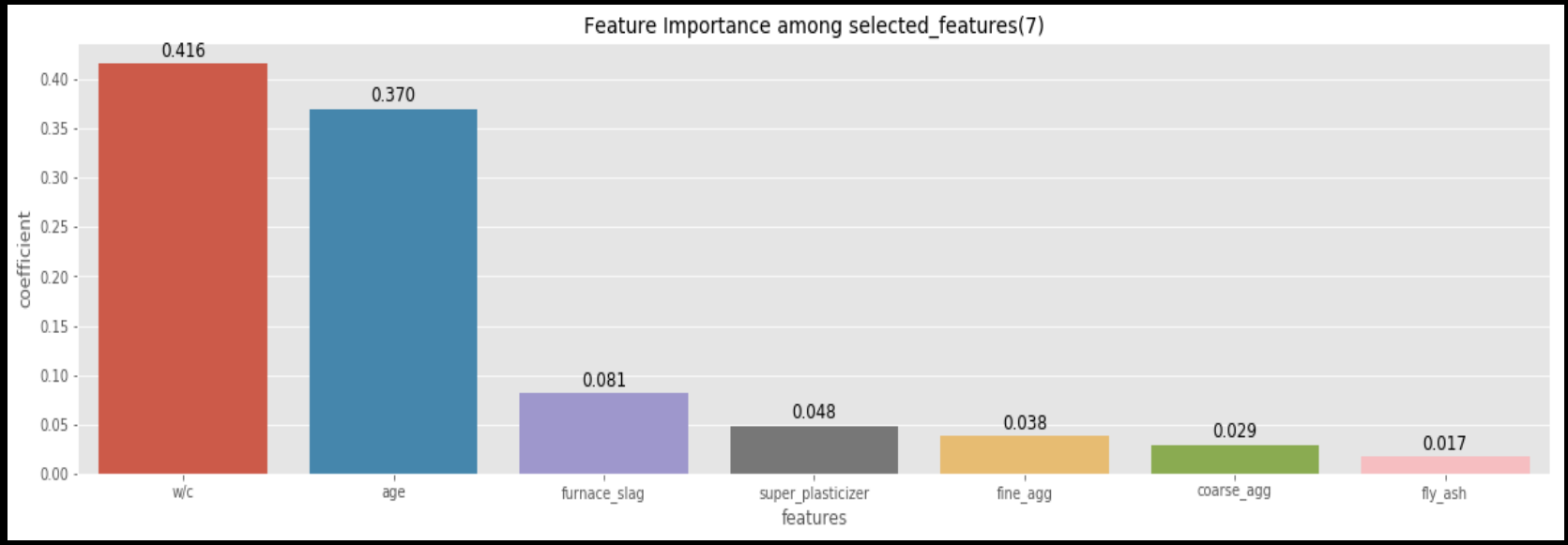
9개 변수 중 W/C 변수의 경우 Cement 변수와 상관계수가 매우 높다(0.876). W/C 변수를 넣을 경우, Water 변수, Cement 변수를 제외하였다. (마침 Water 변수도 비슷한 결합작용을 하는 Superplasticizer 변수 등과 상관관계가 높았고) 이로써 전반적인 상관관계가 낮아졌다.





Feature Selection – II

또다른 차원 축소의 가능성을 확인하기 위해 변수별 중요도를 RandomForestClassifer모델 통해 분석한 결과, W/C 변수와 Age 변수만으로 이미 과반수가 넘는 설명력을 갖고 있다.



features['coefficient'] =
RandomForestRegressor.feature_importances_

Preprocessing Summary

→ 상대 비율 Water/Cement 변수 변환 　 - - -▸ 범주형으로 Age 더미변수화 　  Feature Importance 분석

Variable	Description	Type
cement	**Gradients of concrete**	**Float**
slag	Gradients of concrete	Float
flyash	Gradients of concrete	Float
water	**Gradients of concrete**	**Float**
super plasticizer	Gradients of concrete	Float
coarse aggregate	Gradients of concrete	Float
fine aggregate	Gradients of concrete	Float
age	**Age**	**Integer**
compressive _strength	**Gradients of concrete**	**Float**

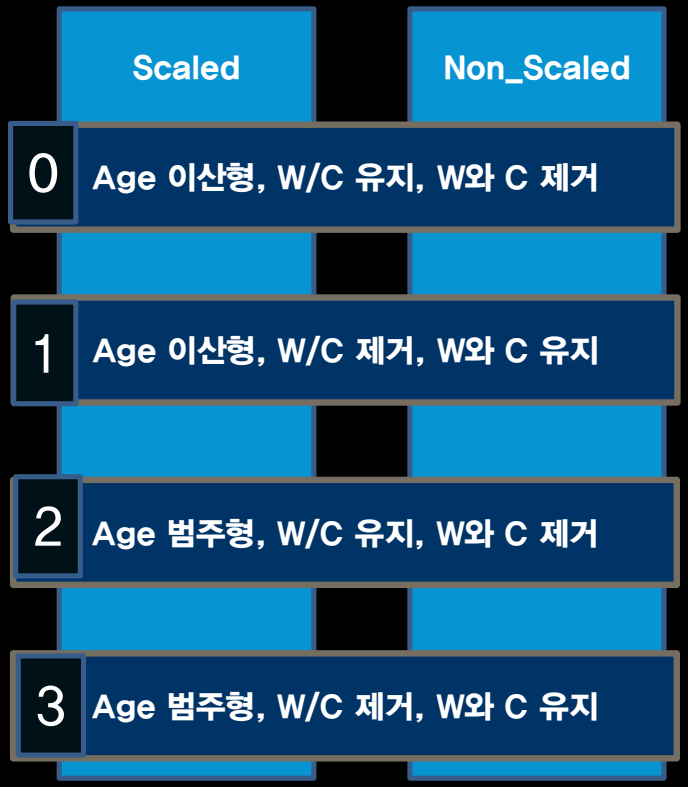
Variable	Description	Type
cement ✕	**Gradients of concrete**	**Float**
slag	Gradients of concrete	Float
flyash	Gradients of concrete	Float
water ✕	**Gradients of concrete**	**Float**
super plasticizer	Gradients of concrete	Float
coarse aggregate	Gradients of concrete	Float
fine aggregate	Gradients of concrete	Float
age ②	**Age**	**Categorical**
compressive _strength	**Gradients of concrete**	**Float**
w/c ①	**Feature Transformation (= water/cement Ratio)**	**Float**

Chapter 2

Training Methods

Dataset Selection

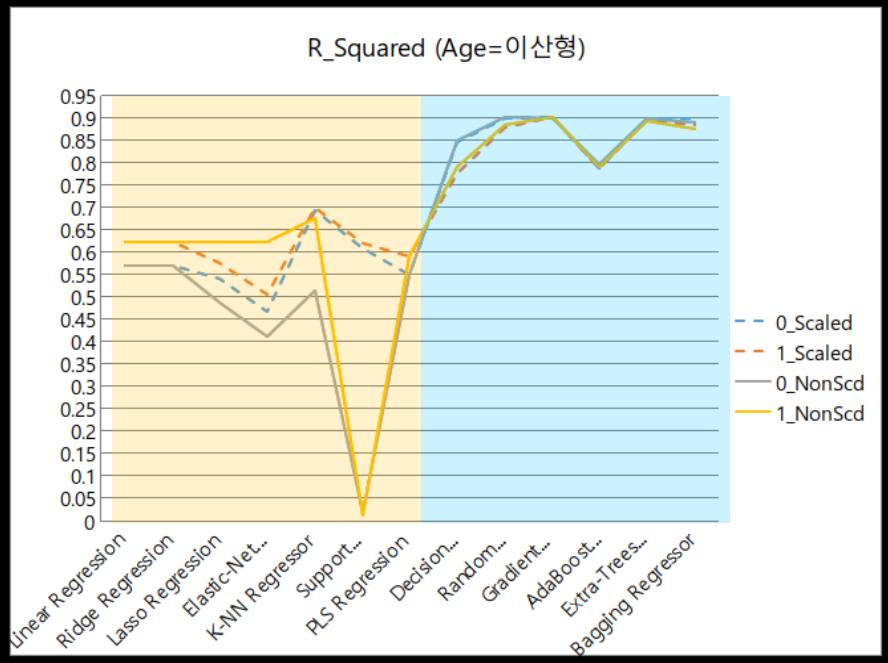
8개의 데이터셋 (Feature Transformation-1,2, StandardScaler 교차적용)에 대해
13개의 회귀모델(Default Parameter)에 최적인 데이터셋 1개씩을 선택한다. (R^2 기준)

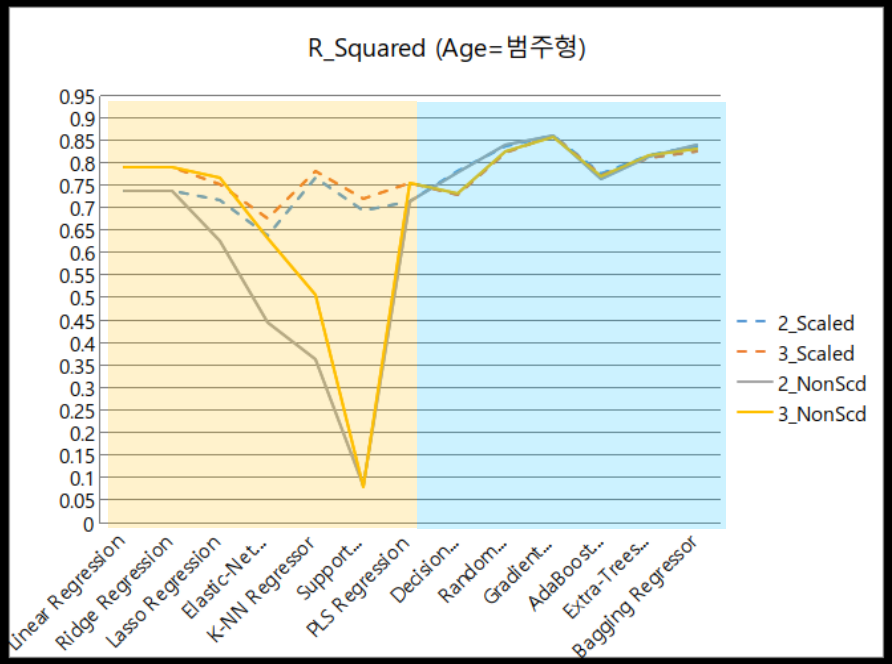



```
feature_name_all = ['cement', 'water', 'furnace_slag', 'fly_ash', 'super_plasticizer',
       'coarse_agg', 'fine_agg', 'age', 'w/c', 'compressive_strength']

#0번 조합 (Age 이산형, w/c변수 유지, water변수 제외, cement변수 제외)
feature_name_0 = ['furnace_slag', 'fly_ash', 'super_plasticizer',
                  'coarse_agg', 'fine_agg', 'w/c',
                  'age', 'compressive_strength']

#1번 조합 (Age 이산형, w/c변수 제거, water변수 유지, cement변수 유지)
feature_name_1 = ['cement', 'water', 'furnace_slag', 'fly_ash', 'super_plasticizer',
                  'coarse_agg', 'fine_agg',
                  'age', 'compressive_strength']

#2번 조합 (Age 범주형, w/c변수 유지, water변수 제외, cement변수 제외)
feature_name_2 = ['furnace_slag', 'fly_ash', 'super_plasticizer',
                  'coarse_agg', 'fine_agg', 'w/c',
                  'age_class_1','age_class_2', 'age_class_3', 'compressive_strength']

#3번 조합 (Age 범주형, w/c변수 제거, water변수 유지, cement변수 유지)
feature_name_3 = ['cement', 'water', 'furnace_slag', 'fly_ash', 'super_plasticizer',
                  'coarse_agg', 'fine_agg',
                  'age_class_1', 'age_class_2', 'age_class_3', 'compressive_strength']
```

Dataset Selection

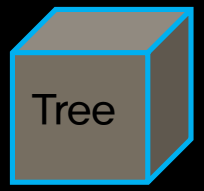
Non-Tree 계열은 3번(Scaled)--, Tree 계열은 0번(Non-Scd)--이 최적 데이터셋이다.





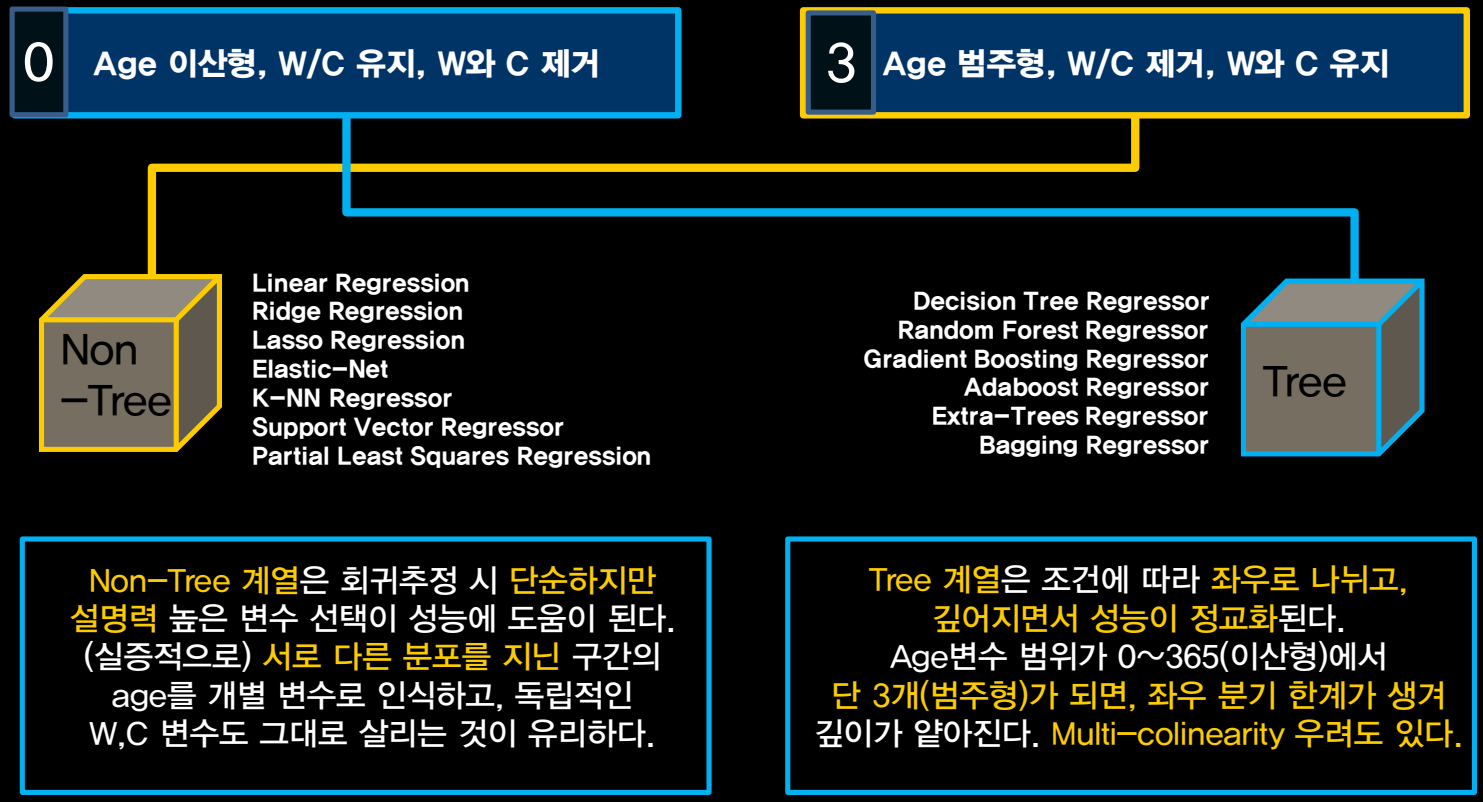
Linear Regression
Ridge Regression
Lasso Regression
Elastic-Net
K-NN Regressor
Support Vector Regressor
Partial Least Squares Regression

Decision Tree Regressor
Random Forest Regressor
Gradient Boosting Regressor
Adaboost Regressor
Extra-Trees Regressor
Bagging Regressor



2 Types of Regression

Non-Tree 계열은 Coeff_(가중치)로, Tree-계열은 Leaves와 Depths(분기)를 통해 회귀 추정하여 변수의 변환(ex. Age 더미변수화)이나 선택(ex. W/C변수)에 다르게 반응한다.



Train, Validate and Test

1. Spllit Dataset (Train: 80%, Test: 20%)

2. 5-Fold (cross validation = 5)

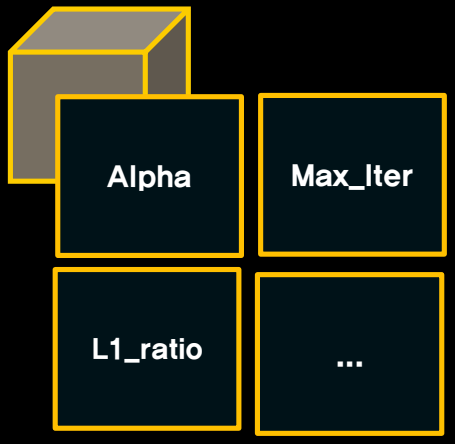
3. Test Train Data on Default Model

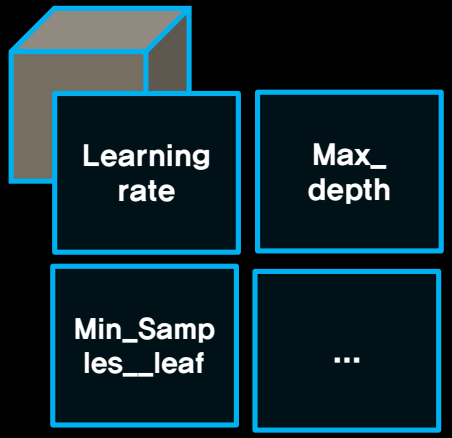
4. Parameter Tuning on GridSearchCV

5. Test Train Data on Best Model
(with Best Parameters from 4)

6. Test Test Data on Best Model

Parameter Tuning






```
names = ['Linear Regression', 'Ridge Regression', 'Lasso Regression',
         'ElasticNet', 'K Neighbors Regressor', 'svr']
models = [LinearRegression(), Ridge(alpha=1.0,max_iter=1000),
          Lasso(alpha=0.001,max_iter=1000),
          ElasticNet(alpha=0.001,max_iter=1000,l1_ratio = 0.5),
          KNeighborsRegressor(n_neighbors=3, p=2),
          SVR(C=100.0, epsilon=1.0, kernel='rbf')]

for name, model in zip(names, models):
    input_scores_bestparam(name, model, X_train_scaled, y_train_raveled)
```

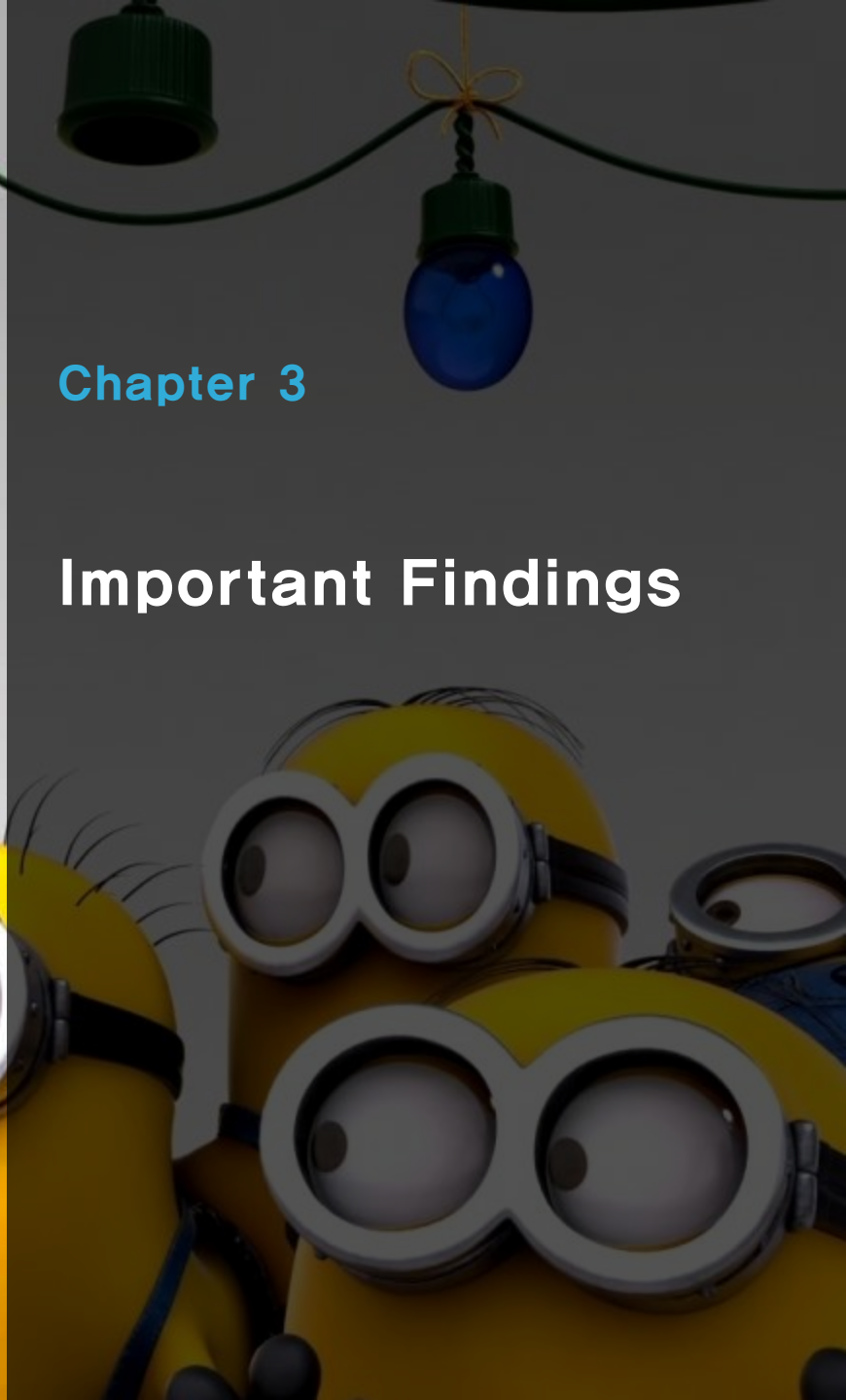
```
param_grid = {'n_estimators':range(20,300,10),
              #'max_depth':[10], #range(5,16,2),
              #'min_samples_split':[100], #range(200,1001,200),
              'learning_rate':[0.01, 0.1, 0.2, 0.3,0.4,0.6,0.8]}
```

```
param_grid = {
              'max_depth':range(5,25,2),
              'min_samples_split':range(2,50,1), #range(200,1001,200),
              'min_samples_leaf':[1,2,5,10]}
```

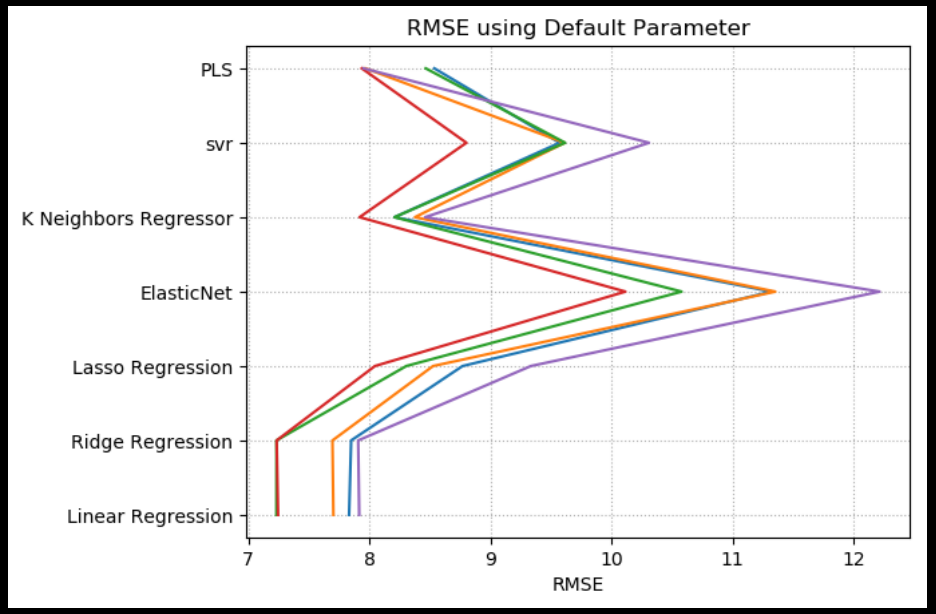
```
param_grid = {'n_estimators':range(20,1001,10),
              'max_depth':range(2,16,2)}
              #'min_samples_split':[100], #range(200,1001,200),
              #'learning_rate':[0.2]}
```

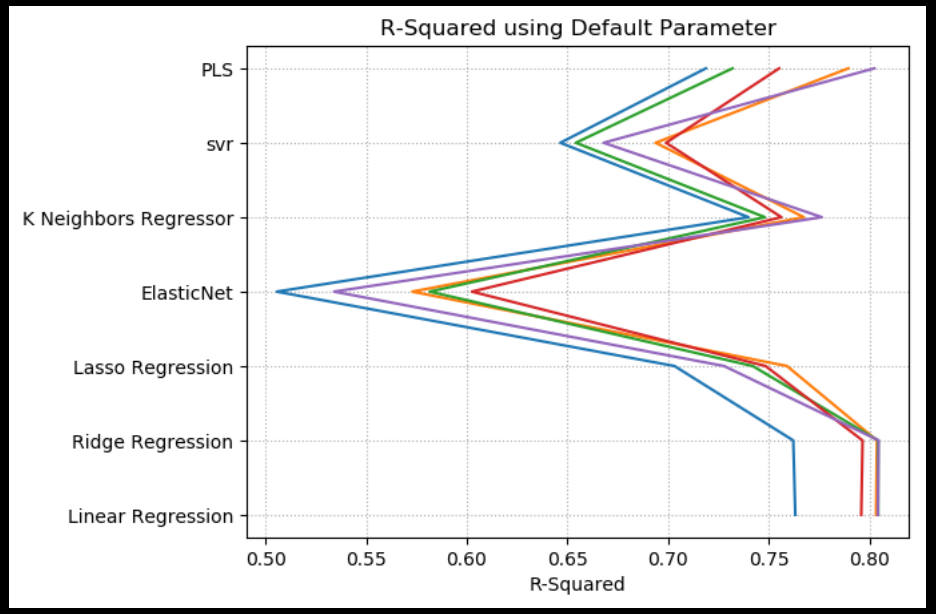
Chapter 3

Important Findings

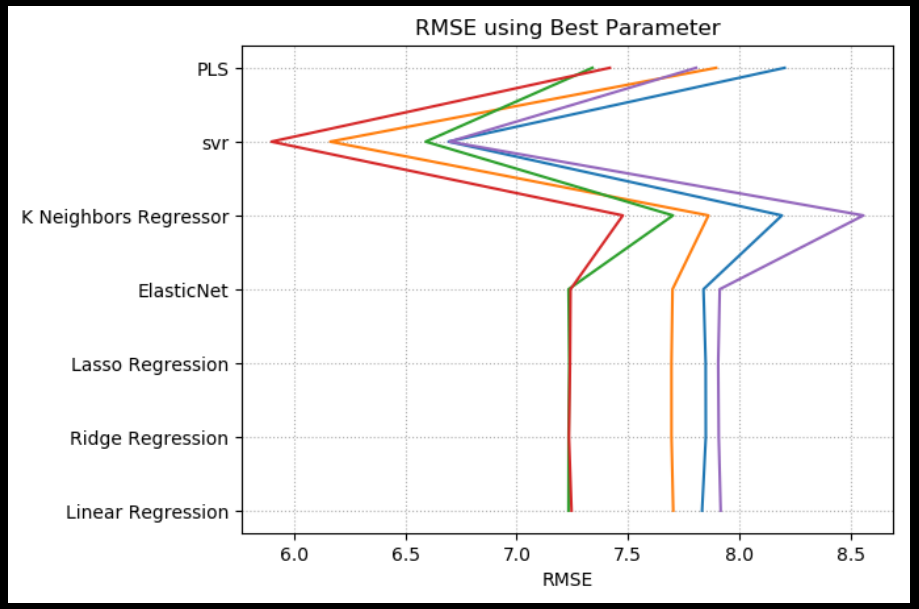


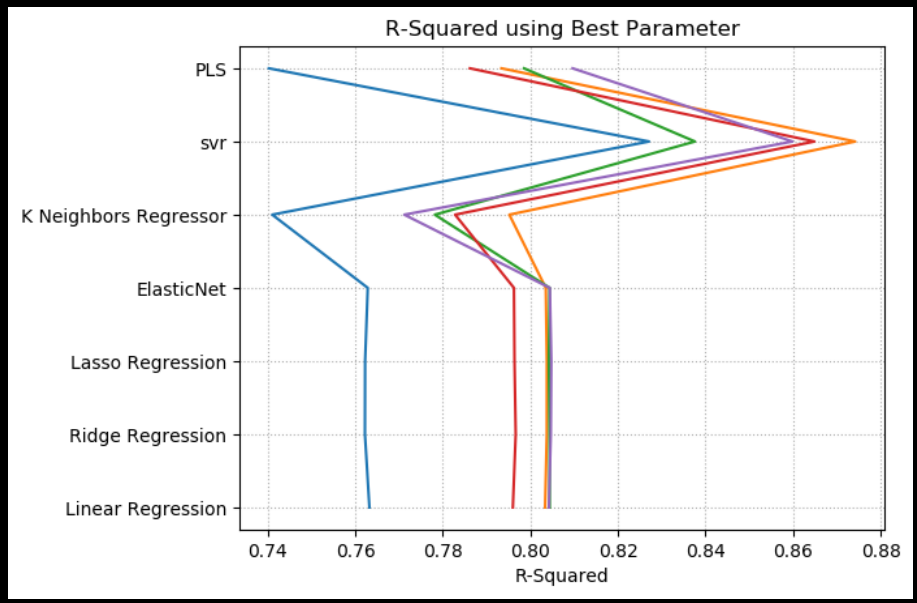
Non_Tree_Default





Non_Tree_Best





SVR = Support Vector Regressor

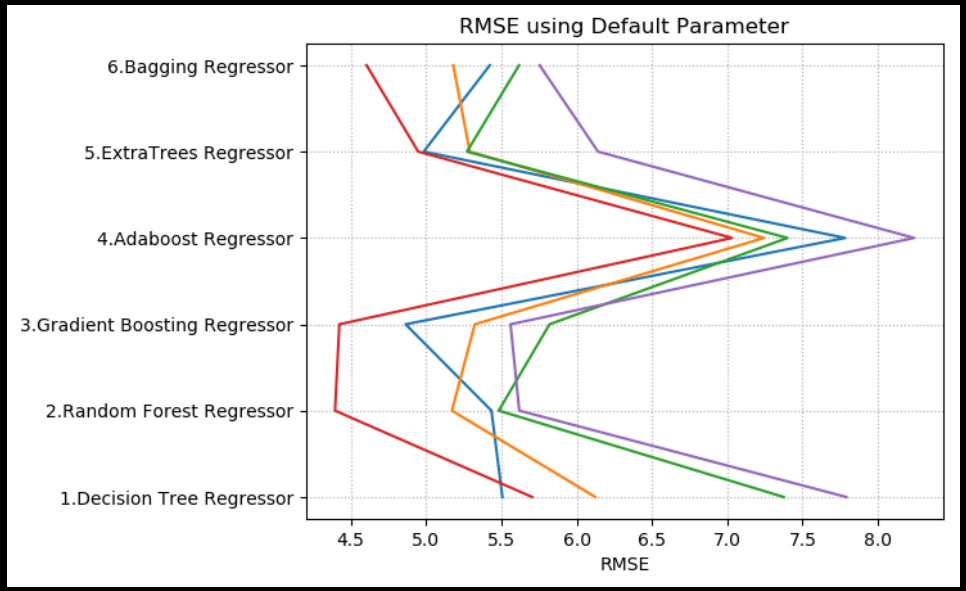
Support Vector Regressor은 특성 스케일링에 민감하다. 이는 고차원의 Kernel 함수 (epsilon loss function)를 통해 Loss Function을 쓰기 때문이어서 약간의 파라미터 조정으로 Generalization Bound가 유의미하게 축소된다.

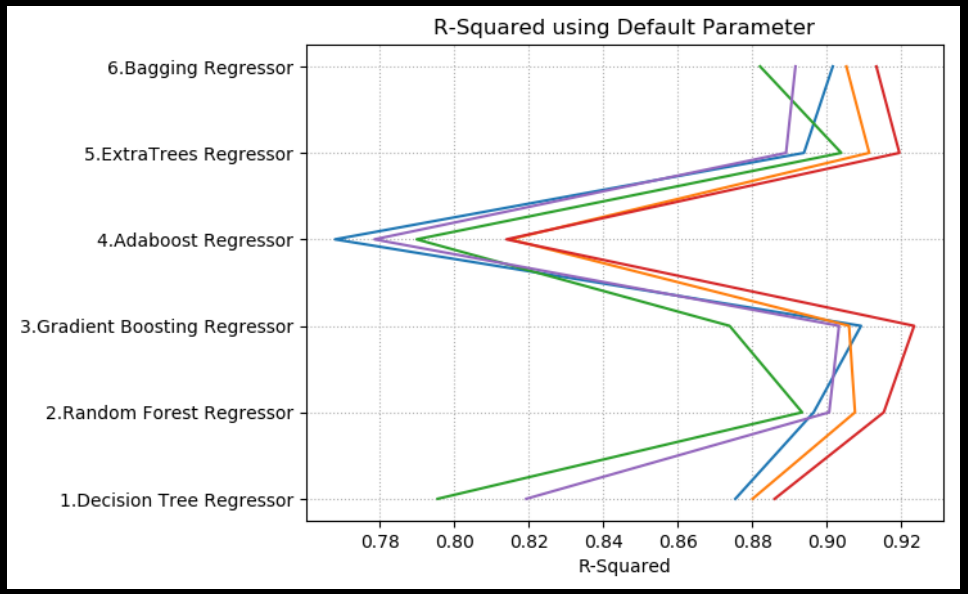
Non-Tree Evaluation

Non-Tree_Def	Model	RMSE	R-Squared
0	Linear Regression	[7.833048466164469, 7.7034619173636765, 7.2328...	[0.7632404213460613, 0.8033544613742243, 0.804...
1	Ridge Regression	[7.849927673761409, 7.695750183663445, 7.23337...	[0.7622189493441326, 0.8037479776894312, 0.804...
2	Lasso Regression	[8.77000230658119, 8.526379712910698, 8.305550...	[0.7032126772206928, 0.7590973612560209, 0.742...
3	Elastic-Net	[11.31883354593852, 11.352356046363132, 10.578...	[0.5056332225208076, 0.5729445928626462, 0.581...
4	K-NN Regressor	[8.208662370657805, 8.376599327686563, 8.20790...	[0.7399896012799085, 0.7674867518374878, 0.748...
5	SVR	[9.571194700505467, 9.612394156799946, 9.61676...	[0.6465091043546312, 0.6938210287523329, 0.654...
6	PLS	[8.536678069202077, 7.968551057904823, 8.46575...	[0.7187945465896879, 0.7895877902485265, 0.732...

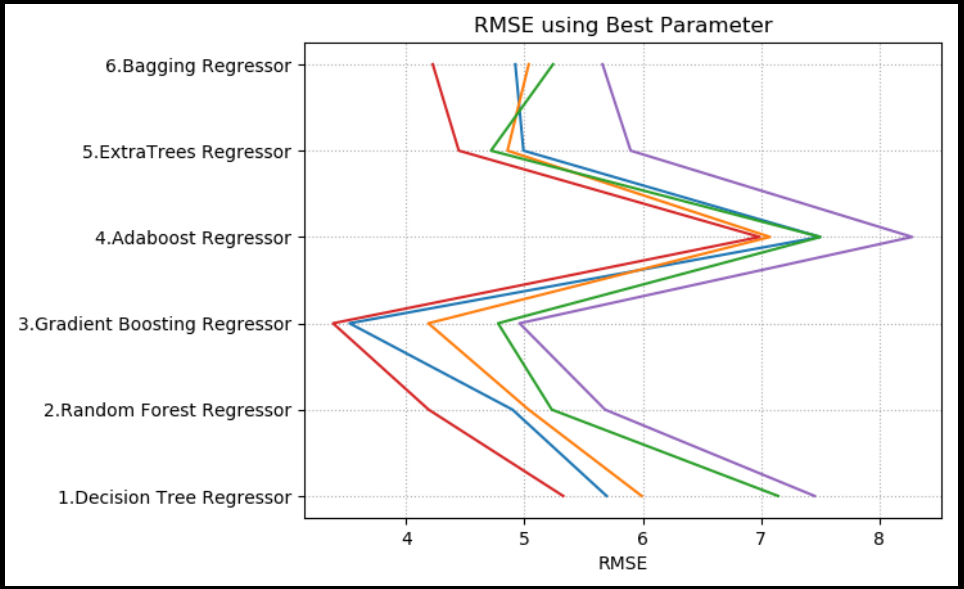
Non-Tree_Best	Model	RMSE	R-Squared
0	Linear Regression	[7.833048466164469, 7.7034619173636765, 7.2328...	[0.7632404213460613, 0.8033544613742243, 0.804...
1	Ridge Regression	[7.849927673761409, 7.695750183663445, 7.23337...	[0.7622189493441326, 0.8037479776894312, 0.804...
2	Lasso Regression	[7.849309409127169, 7.695588501054089, 7.23646...	[0.7622564034025812, 0.8037562238525686, 0.804...
3	Elastic-Net	[7.839379044483202, 7.70032070208997, 7.233100...	[0.7628575740476407, 0.8035147996834413, 0.804...
4	K-NN Regressor	[8.191920833028613, 7.86177006111093, 7.702494...	[0.7410491003165243, 0.7951891815580846, 0.778...
5	SVR	[6.693846615768109, 6.1623439883622595, 6.5911...	[0.8270990242230281, 0.8741642352557155, 0.837...
6	PLS	[8.204011422651021, 7.89620462548936, 7.340302...	[0.7402841565201537, 0.7933911090134961, 0.798..

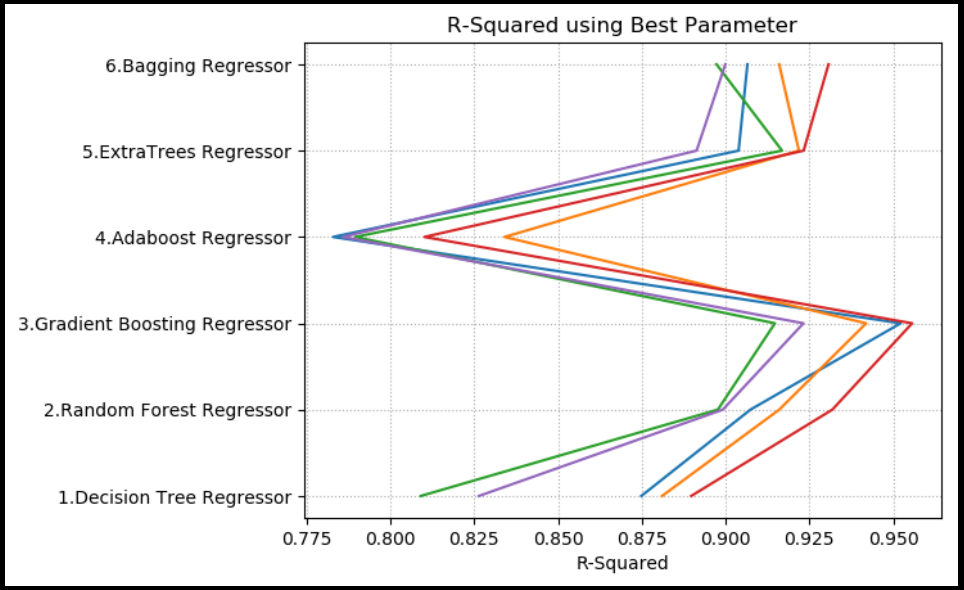
Tree_Default





Tree_Best





Gradient Boosting Regressor

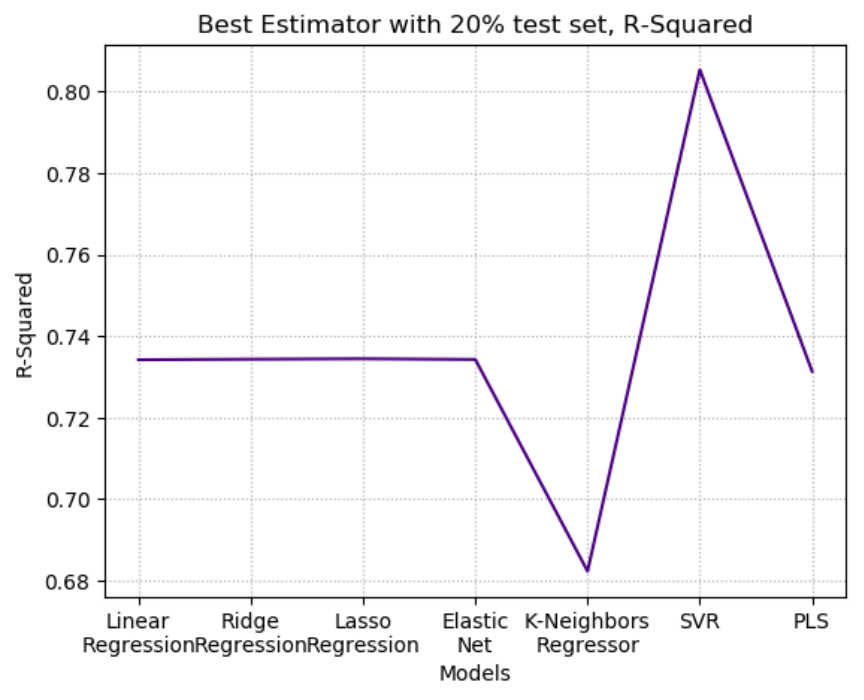
Gradient Boosting Regressor은 앙상블된 이전 모델의 Error에 대한 Gradient Descent을 활용해 예측 성능이 눈에 띄게 개선된다. 또한, Boosting 즉 iteration 전에 다른 샘플로 다시 분기하므로 더 많은 정보를 활용할 수 있다.

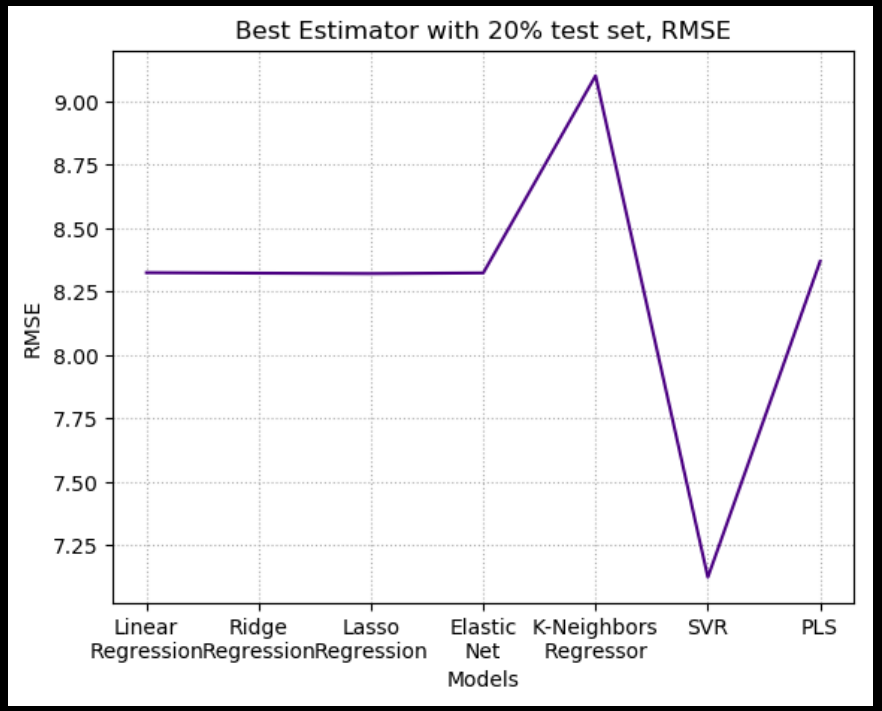
Tree Evaluation

Tree_Default	Model	RMSE	R-Squared
0	Decision Tree Regressor	[6.19900177692,5.9018652775, 7.18610561705, ...	[0.870884198583, 0.88419079835, 0.796053409931...
1	Random Forest Regressor	[5.06152334947,5.0176656627, 5.65803981202, ...	[0.901151315409, 0.909246488177, 0.89184317674...
2	Gradient Boosting Regressor	[4.86371983218,5.32443868902, 5.7966921346, 4...	[0.908884568584, 0.906131257399, 0.87399969906...
3	Adaboost Regressor	[7.73567289118,7.6489303221, 7.50327460029, ...	[0.774897330752, 0.81908220848, 0.787790945786...
4	ExtraTrees Regressor	[5.01400769149,5.1533198814, 5.33360427228, 5...	[0.896933498594, 0.91258340774, 0.895044610641...
5	Bagging Regressor	[5.49019724484,5.4718871751, 5.75808774804, ...	[0.883426672105, 0.908496588798, 0.87245505354...

Tree_Best	Model	RMSE	R-Squared
0	Decision Tree Regressor	[5.69466218808,5.9912308986, 7.14522161638, ...	[0.874865889115, 0.881056895317, 0.80911373625...
1	Random Forest Regressor	[4.89857484488,5.0352104152, 5.23111719475, 4...	[0.907406761263, 0.915988974377, 0.89768668169...
2	Gradient Boosting Regressor	[3.52037058636,4.1868523172, 4.77637529888, 3...	[0.952179232001, 0.941913350919, 0.91470174275...
3	Adaboost Regressor	[7.50017478396,7.0761762232, 7.50356630126, ...	[0.782938625981, 0.834083377616, 0.78948709749...
4	ExtraTrees Regressor	[4.99134590119,4.8564203977, 4.71667876843, ...	[0.903866420283, 0.921849440221, 0.91682058349...
5	Bagging Regressor	[4.92234925216,5.0357112805, 5.24156409422, ...	[0.906505808893, 0.915970596403, 0.89727762023...

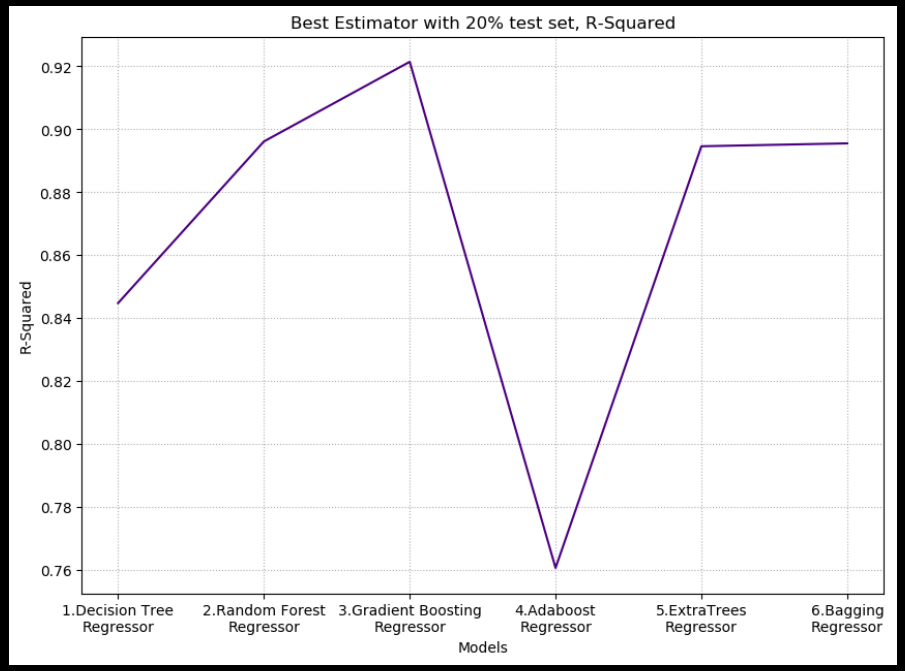
Non-Tree Test

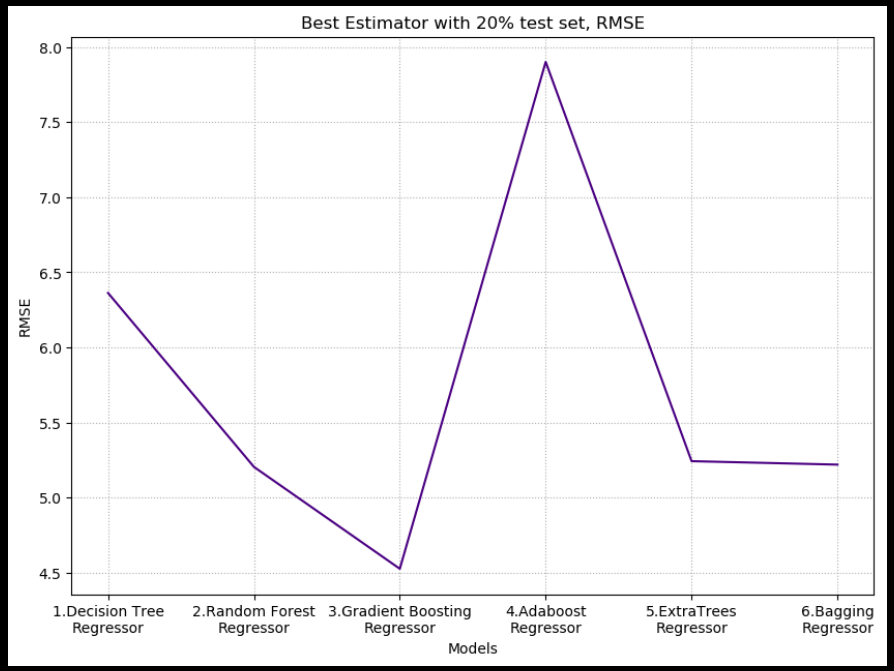




Non-Tree Test	RMSE	R-Squared
Linear Regression	8.324818	0.734163
Ridge Regression	8.322738	0.734296
Lasso Regression	8.321047	0.734404
Elastic-Net	8.323653	0.734238
K-NN Regressor	9.102043	0.682208
SVR	7.122552	0.805403
PLS	8.369277	0.731316

Tree Test





Tree Test	RMSE	R-Squared
Decision Tree Regressor	6.361956	0.844744
Random Forest Regressor	5.204073	0.896115
Gradient Boosting Regressor	4.525716	0.921433
Adaboost Regressor	7.900304	0.760583
Extra-Trees Regressor	5.242133	0.894589
Bagging Regressor	5.219202	0.89551

Key Takeaway

Domain

낯선 분야, 그러나 한국콘크리트학회 등 자료 활용해 체계적 변수 검증 및 설정

Approach

8가지 데이터셋, 2가지 Regression 대분류 후 13개의 Regression 알고리즘 실행해 성능 향상뿐 아니라 더미화/정규화 등 알고리즘의 특성에 대한 실질적 이해

Training

1000여 개의 데이터셋으로는 iteration의 한계, 과적합 발생 가능성이 농후
그러나 R^2 유의미하게 향상

Afterwork

1. Correlation 추가 분석 통해 또 다른 변수 조합 및 차원 축소를 통한 추정 성능 향상
2. 과제 요구사항이었던 R^2, RMSE 외에 다른 score 방식(예컨대 Timing)을 통해 콘크리트 관련 현장에 보다 적합한 추정 알고리즘 순위 부여
3. Rule ensembles (Cubist model)과 neural networks 모델도 추후 모델링하고 튜닝하여 추가 연구 진행 필요

참고문헌

- 김경환 외 (2014).[감수율에 따른 압축강도와 물–시멘트비 관계에 관한 연구], 한국콘크리트학회 : 콘크리트학회 논문집 26권 5호, 1–2
- Max Kuhn et.al. Applied Predictive Modeling.pdf, Chap10.
- Ben Gorman (2017).*A Kaggle Master Explains Gradient Boosting*, GormAnalysis.

감사합니다

5조

노진현 문명진 박지연 이한얼