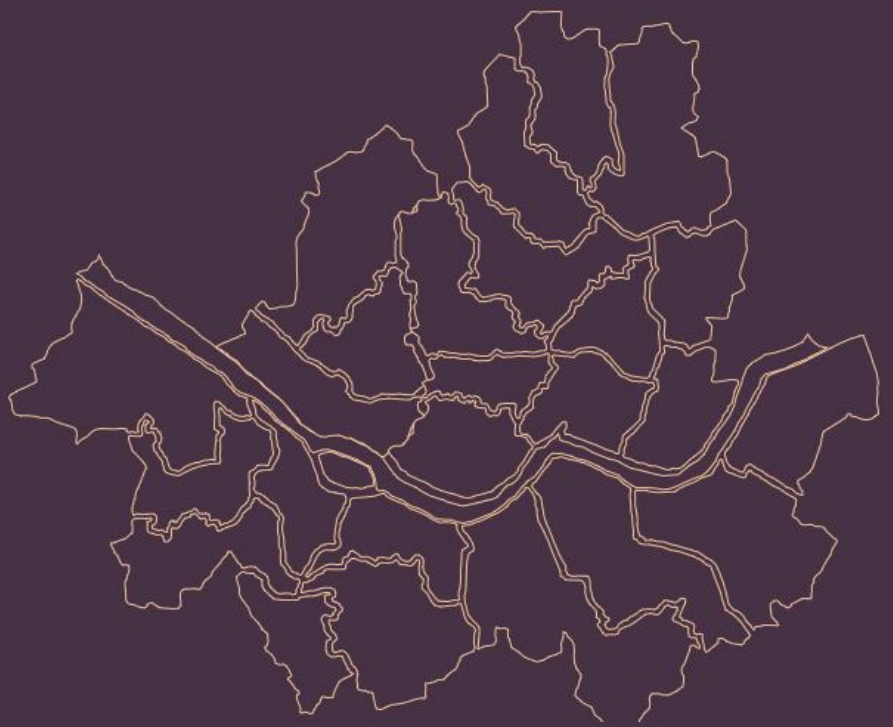
20-1 YONSEI ESC FINAL PROJECT

REGRSSION WITH
서울 주민 가구원 행복도 데이터

2020. 06. 04

2조 김윤환 백채빈 손지우 신혜연 이상완 조인식



목 차

1. 역할 분담 및 INTRO
2. Our Approach
3. Model Selection
4. 결과
5. 결과 해석 및 한계점
6. Q & A

역할 분담 및 INTRO
사람 좋은 2조~~!

역 할 분 담

1. 김윤환(조장님) : SVR 모델링
2. 백채빈 : 변수 정리 및 Lasso 모델링
3. 손지우 : FA 모델링
4. 신혜연, 조인식 : PPT 제작 및 발표
5. 이상완 : Random Forest, Xgboost, lightGBM 모델링 및 배경지식 제공

역할 분담 및 INTRO
지난 번 어디까지 했었지?

지난 1주차

- 범주형 변수가 남발하는 설문조사 데이터 전처리
- 변수 정리

지난 1주차 마지막 피피티

앞으로...

1) Feature Extraction
PCA < FAMD

2)Analysis
SVM
XGboost
Group Lasso
Random Forest
*stacking

지난 1주차의 계획

- PCA, FA, FAMD ?
- ANALYSIS
- FA ~ 행복의 3요소?

역할 분담 및 INTRO
지난 번 어디까지 했었지?

지난 1주차

- 범주형 변수가 남발하는 설문조사 데이터 전처리
- 변수 정리

지난 1주차 마지막 피피티

앞으로...

1) Feature Extraction
PCA < FAMD

2)Analysis
SVM
XGboost
Group Lasso
Random Forest
*stacking

지난 1주차의 계획 결과

- 변수 정리 및 더미변수 생성
- PCA, FA, FAMD ?
  + FA ~ 행복의 3요소?
    FA !

- ANALYSIS
  우리가 배운 것을
  적용해 해석해보자

Our Approach
앞으로 뭘 할까?

Our Approach ---- in Github

1. 가구원의 소득, 계층, 주거환경 등 다양한 변수가 서로 어떤 관련이 있는가?
2. 가구원의 소득, 계층, 주거환경 등 다양한 변수를 바탕으로 행복도를 예측할 수 있는가?
3. 가구원의 행복도를 제고하기 위해 어떻게 해야 하는가?

데이터 전처리 후

유의할 모델들 적용해보기

SVR, Lasso
PCA, FA
Random Forest,
Xgboost, lightGBM 등

⇄

모델 해석

RMSE 낮추기

RMSE를 감안하더라도 한 번 "잘" 해석해보자 !

Our Approach
앞으로 뭘 할까?

Our Approach ---- in Github

1. 가구원의 소득, 계층, 주거환경 등 다양한 변수가 서로 어떤 관련이 있는가?
2. 가구원의 소득, 계층, 주거환경 등 다양한 변수를 바탕으로 행복도를 예측할 수 있는가?
3. 가구원의 행복도를 제고하기 위해 어떻게 해야 하는가?

데이터 전처리 후

유의할 모델들 적용해보기

SVR, Lasso
PCA, FA
Random Forest,
Xgboost, lightGBM 등

설문조사의 특성, 행복도 고려

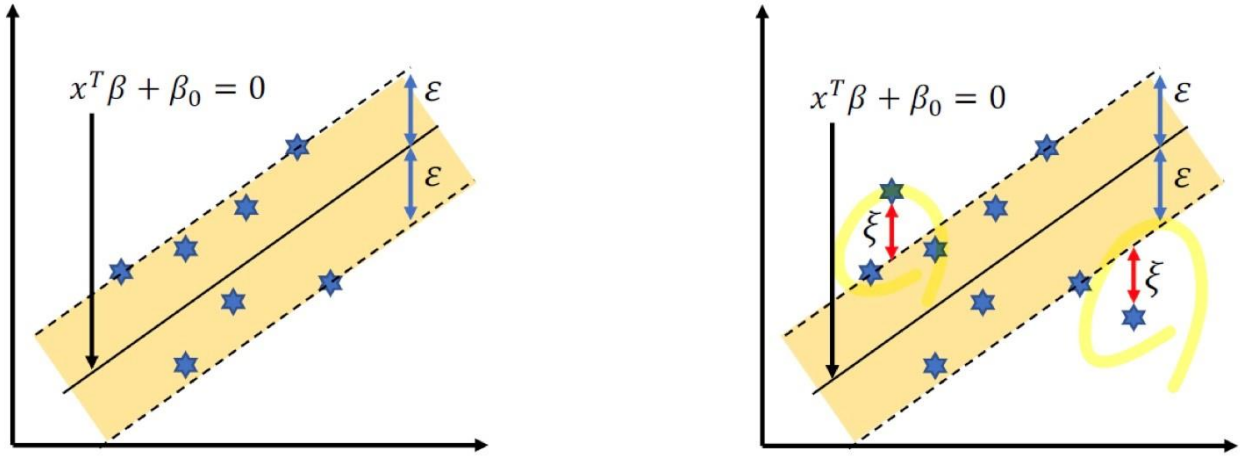
모델 해석

RMSE 낮추기

RMSE를 감안하더라도 한 번 "잘" 해석해보자 !

SVR이란 : SVM의 Regression version

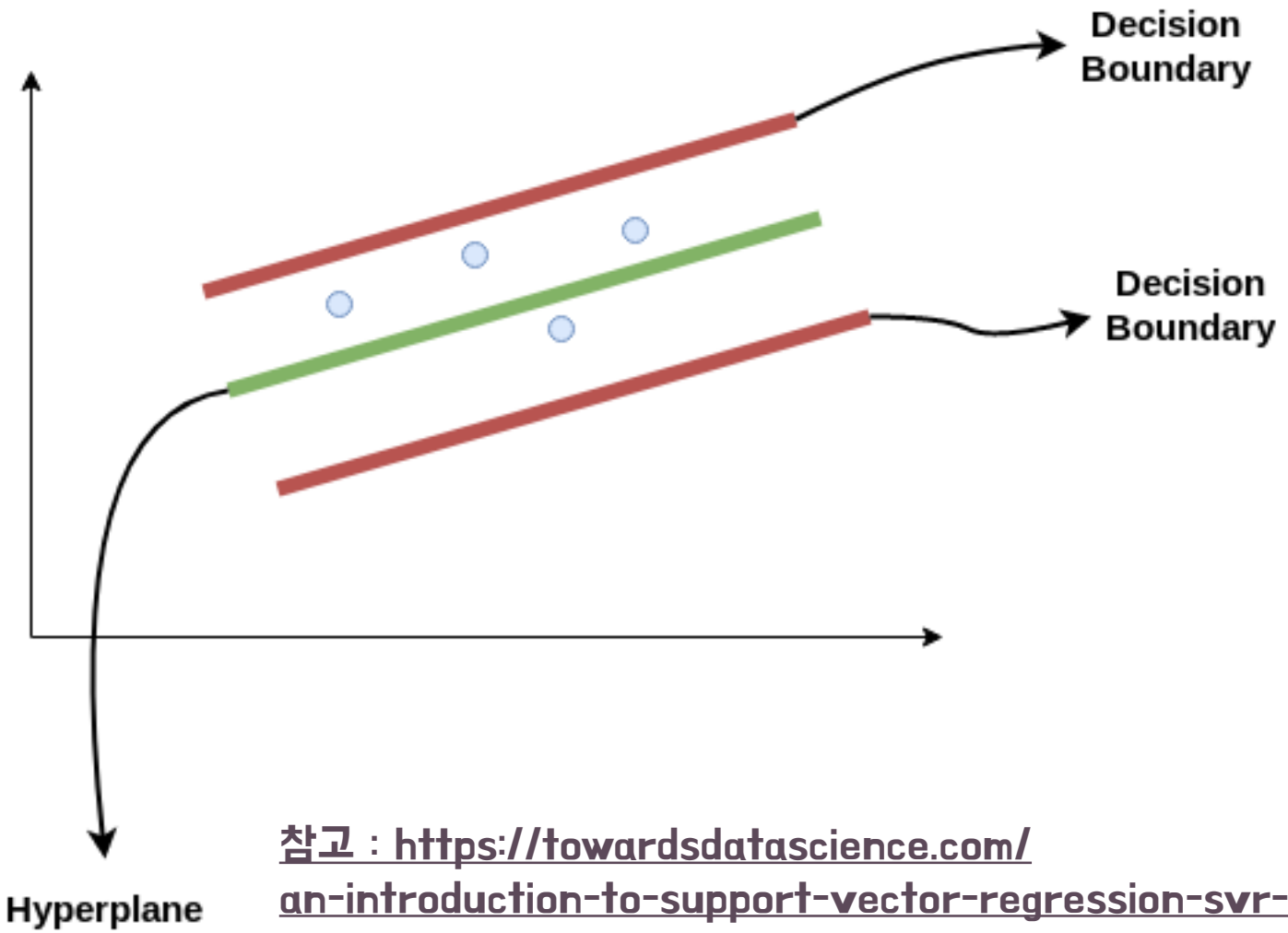
SVR(support vector regression)



Q. 종속변수가 연속형 자료라면 ?
차이점은, 에러의 정의

Model Selection
SVR

SVR이란 : **SVM의 Regression version**



참고 : https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

Hyperparameter 임의로 지정

```
from sklearn.svm import SVR
regressor = SVR(kernel = 'linear',C=100,gamma='auto')
regressor.fit(X_train, y_train)
```

```
from sklearn.metrics import mean_squared_error
mean_squared_error(y_test, y_pred)
```

```
96.77509533338213
```

Hyperparameter 임의로 지정

```
from sklearn.svm import SVR
regressor = SVR(kernel = 'linear',C=100,gamma='auto')
regressor.fit(X_train, y_train)
```

```
from sklearn.metrics import mean_squared_error
mean_squared_error(y_test, y_pred)
```

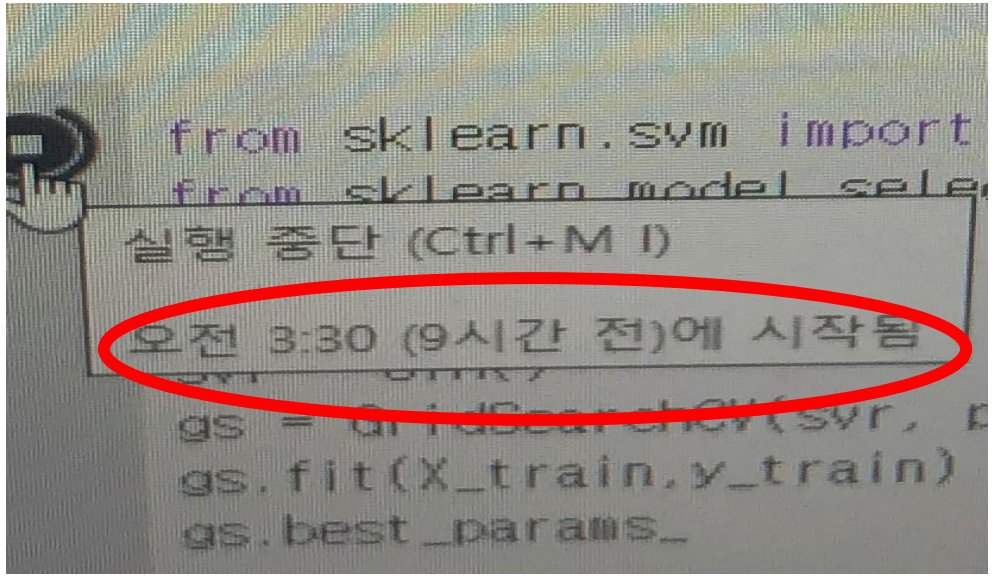
```
96.77509533338213
```

Grid search로 최적의 hyperparameter 찾자!

Grid Search 시도

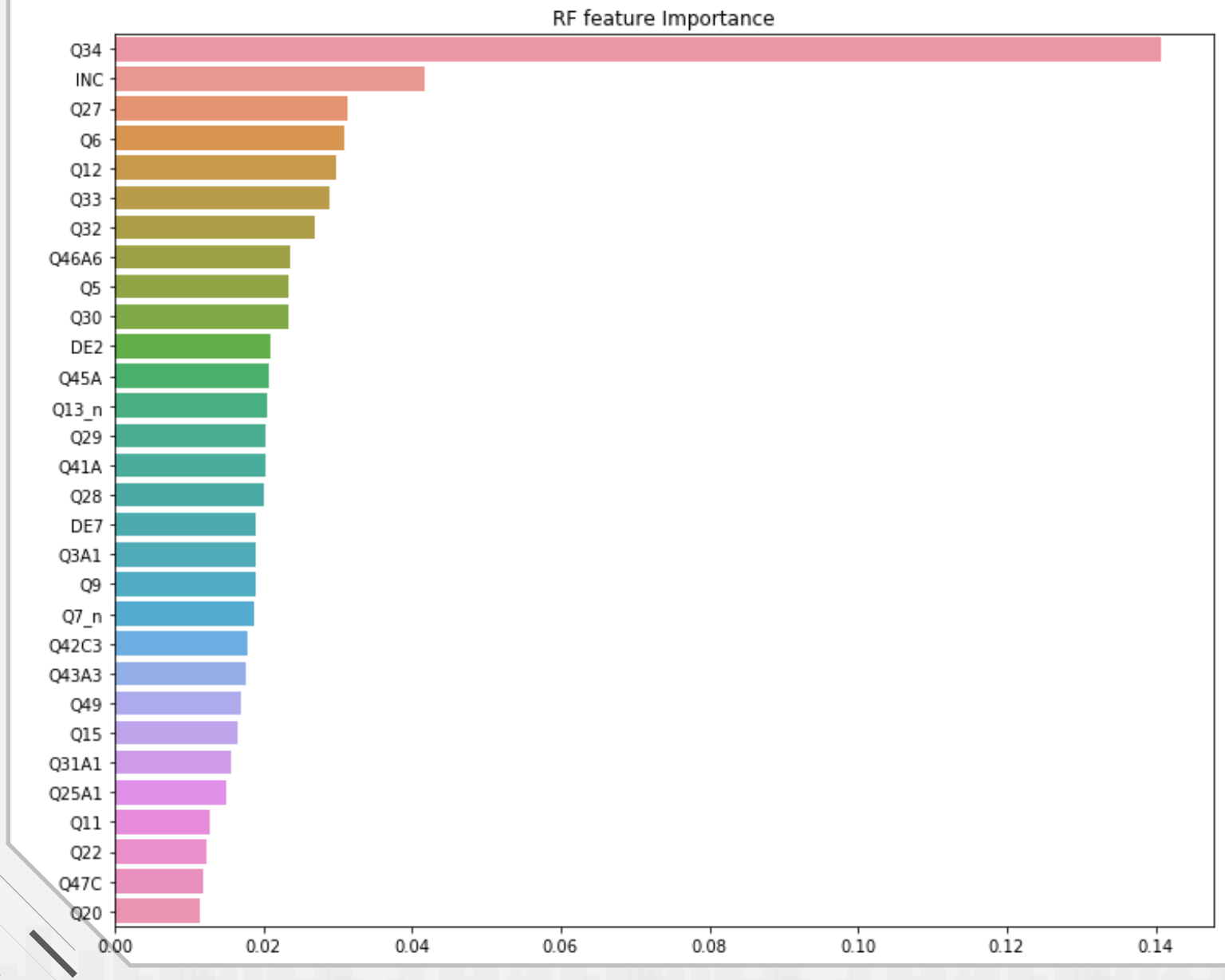
```
[ ]  from sklearn.svm import SVR
     from sklearn.model_selection import GridSearchCV

     parameters = {'kernel': ('linear', 'rbf'), 'C':[1,5,10],'gamma': [1e-7,0.1,1.0,10.0,100.0],'epsilon':[0,0.1,1,2,4]}
     svr = SVR()
     gs = GridSearchCV(svr, parameters)
     gs.fit(X_train,y_train)
     gs.best_params_
```



포기!

Model Selection
Random Forest, XGBoost, lightGBM

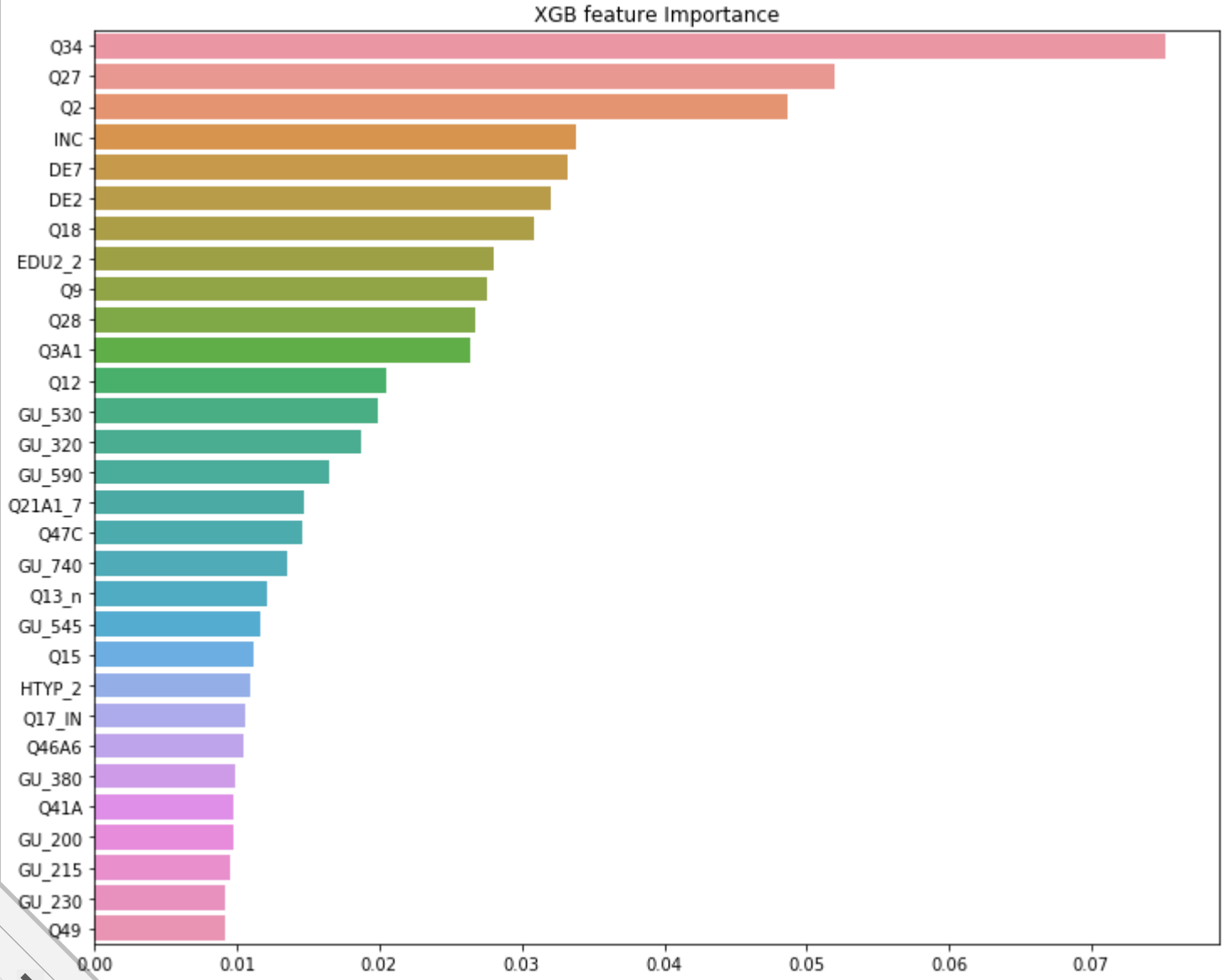


Random Forest

- RandomForestRegressor
- GridSearch
- Feature importance
- RMSE :

```
print(rf2_rmse)
8.440446955359622
```

Model Selection
Random Forest, XGBoost, lightGBM



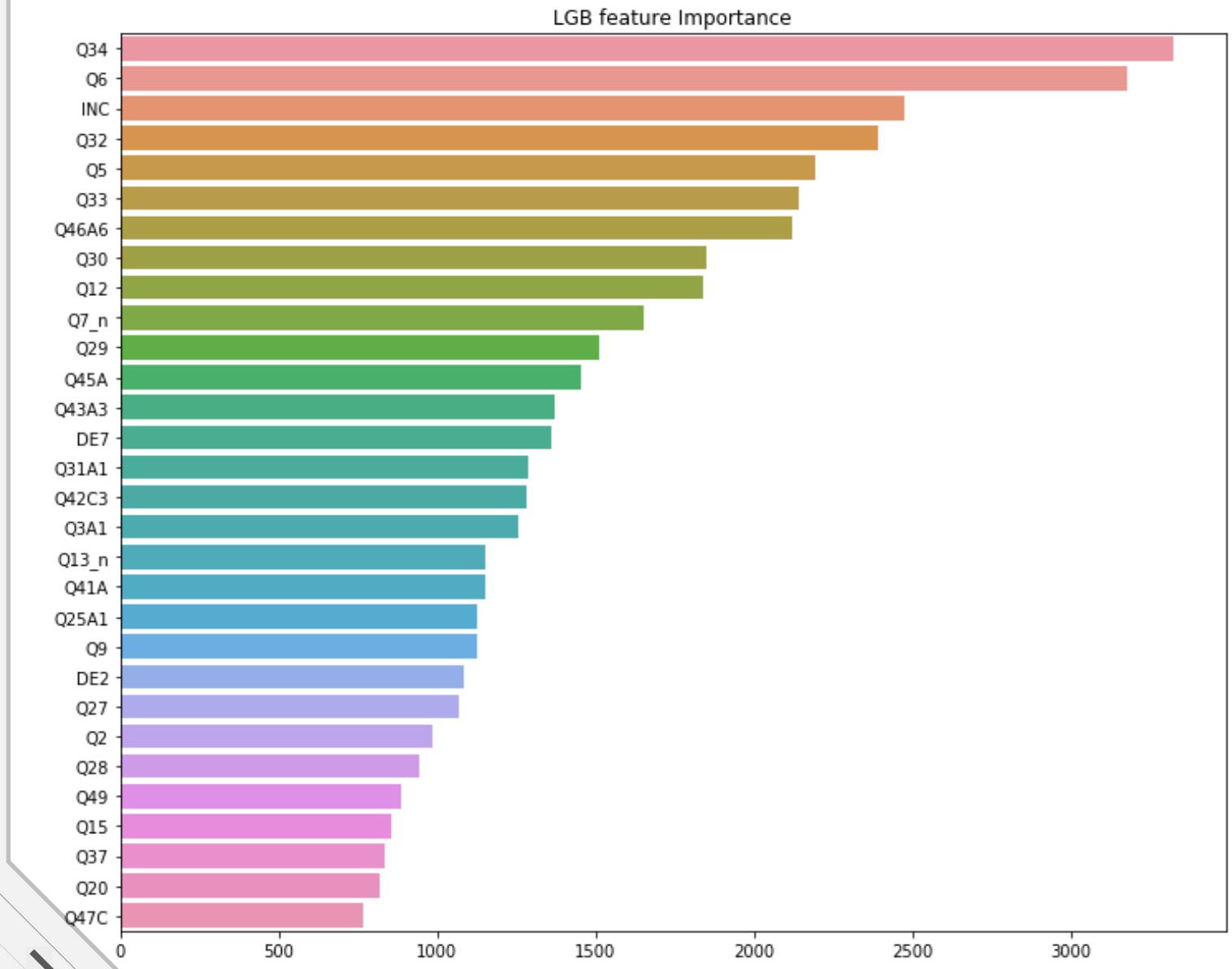
XGBoost

- XGBRegressor
- GridSearch
- Feature importance
- RMSE :

```
xgb_rmse
```

8.753194818536665

Model Selection
Random Forest, XGBoost, lightGBM



lightGBM

- LGBMRegressor
- GridSearch
- Feature importance
- RMSE

```
lgb_rmse
```

8.753194818536665

Model Selection
FA : LASSO 후 FA

1. 표준화 후 LASSO

```
clf_std = linear_model.Lasso(alpha=0.1, fit_intercept=True)
clf_std.fit(X_std_train, y_std_train)
```

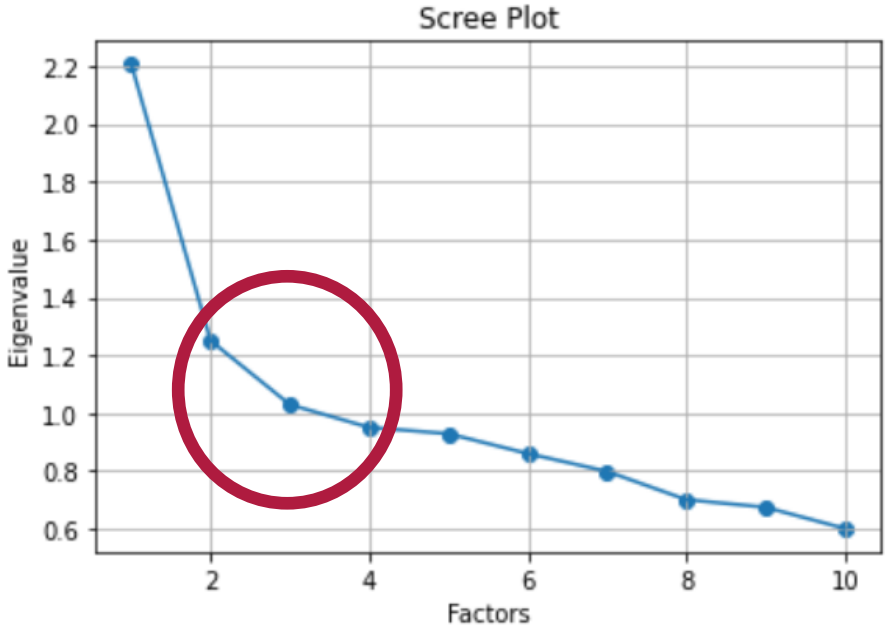
```
len(clf_std.coef_[clf_std.coef_ != 0])
```

```
11
```

```
lasso_std_var = X_train.columns[np.where(clf_std.coef_ != 0)]
lasso_std_var
```

```
Index(['INC', 'Q9', 'Q15', 'Q18', 'Q27', 'Q28', 'Q34', 'DE2', 'Q13_n',
       'GU_740', 'EDU2_2'],
      dtype='object')
```

2. Factor number 결정



3. FA with 3 Factor

- FactorAnalysis
- OLS 실시 ----------->
- RMSE :

```
#표준화 복원 후 RMSE
print('표준화 복원 후 RMSE: %f' %np.sq
```

표준화 복원 후 RMSE: 9.899882

OLS Regression Results

Dep. Variable:	Q4B	**R-squared (uncentered):**	0.185
Model:	OLS	**Adj. R-squared (uncentered):**	0.185
Method:	Least Squares	**F-statistic:**	1686.
Date:	Tue, 02 Jun 2020	**Prob (F-statistic):**	0.00
Time:	15:08:00	**Log-Likelihood:**	-30107.
No. Observations:	22293	**AIC:**	6.022e+04
Df Residuals:	22290	**BIC:**	6.024e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Factor 1	-0.5141	0.008	-68.390	0.000	-0.529	-0.499
Factor 2	0.1838	0.010	18.523	0.000	0.164	0.203
Factor 3	-0.0647	0.011	-6.126	0.000	-0.085	-0.044

Omnibus:	3932.153	**Durbin-Watson:**	1.444
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	11564.584
Skew:	-0.929	**Prob(JB):**	0.00
Kurtosis:	6.000	**Cond. No.**	1.40

- 행복도의 3 요소

A formula for wellbeing

SWB = LS + FPA - FNA

SWB = Subjective Wellbeing

LS = Life Satisfaction

FPA = Frequent Positive Affect

FNA = Frequent Negative Affect

결과
LASSO 묻고 FA로 가~!

1. Factor 별 coefficient

	coef
Factor 1	-0.5141
Factor 2	0.1838
Factor 3	-0.0647

```
#INC: 임금수준
#STR: 스트레스(STRESS)
#CUL: 일상생활에서 문화예술의 비중
#WIT: 함께 여가생활하는 사람
#SES: SocioEconomic Status
#CMP: 계층이동가능성(Class movement possibility)
#PRD: 서울 시민으로서의 자부심
#AGE: 연령(10단위)
#ESA: 문화환경만족도(Envinronment Satisfaction)
#GD: 강동구
#EDU: 학력(고졸 or NOT)
```

2. 11개 변수별 Factor 에 대한 weight

	INC	STR	CUL	WIT	SES	CMP	PRD	AGE	ESA	GD	EDU
0	-0.482514	0.038056	-0.528027	0.253000	0.554757	-0.324104	-0.259185	0.395685	-0.345583	-0.164507	-0.449886
1	0.047825	0.231324	-0.094062	0.031549	-0.290380	0.326311	0.086959	0.359734	0.032187	0.141333	-0.283642
2	0.296181	-0.066858	-0.364506	0.012909	-0.163148	-0.085713	-0.109765	-0.037807	-0.250989	0.015407	0.167932

결과 해석 및 한계점
결과를 잘 해석해보자

"FACTOR의 coef"

	coef
Factor 1	-0.5141
Factor 2	0.1838
Factor 3	-0.0647

+

"변수별 의미"

```
#INC: 임금수준
#STR: 스트레스(STRESS)
#CUL: 일상생활에서 문화예술의 비중
#WIT: 함께 여가생활하는 사람
#SES: SocioEconomic Status
#CMP: 계층이동가능성(Class movement possibility)
#PRD: 서울 시민으로서의 자부심
#AGE: 연령(10단위)
#ESA: 문화환경만족도(Envinronment Satisfaction)
#GD: 강동구
#EDU: 학력(고졸 or NOT)
```

INC	STR	CUL	WIT	SES	CMP	PRD	AGE	ESA	GD	EDU
-0.482514	0.038056	-0.528027	0.253000	0.554757	-0.324104	-0.259185	0.395685	-0.345583	-0.164507	-0.449886
0.047825	0.231324	-0.094062	0.031549	-0.290380	0.326311	0.086959	0.359734	0.032187	0.141333	-0.283642
0.296181	-0.066858	-0.364506	0.012909	-0.163148	-0.085713	-0.109765	-0.037807	-0.250989	0.015407	0.167932

결과 해석 및 한계점
결과를 잘 해석해보자

"FACTOR의 coef"

	coef
Factor 1	-0.5141
Factor 2	0.1838
Factor 3	-0.0647

+

"변수별 의미"

```
#INC: 임금수준
#STR: 스트레스(STRESS)
#CUL: 일상생활에서 문화예술의 비중
#WIT: 함께 여가생활하는 사람
#SES: SocioEconomic Status
#CMP: 계층이동가능성(Class movement possibility)
#PRD: 서울 시민으로서의 자부심
#AGE: 연령(10단위)
#ESA: 문화환경만족도(Envinronment Satisfaction)
#GD: 강동구
#EDU: 학력(고졸 or NOT)
```

INC	STR	CUL	WIT	SES	CMP	PRD	AGE	ESA	GD	EDU
-0.482514	0.038056	-0.528027	0.253000	0.554757	-0.324104	-0.259185	0.395685	-0.345583	-0.164507	-0.449886
0.047825	0.231324	-0.094062	0.031549	-0.290380	0.326311	0.086959	0.359734	0.032187	0.141333	-0.283642
0.296181	-0.066858	-0.364506	0.012909	-0.163148	-0.085713	-0.109765	-0.037807	-0.250989	0.015407	0.167932

결과 해석 및 한계점
결과를 잘 해석해보자

* Factor 별 coef

	coef
Factor 1	-0.5141
Factor 2	0.1838
Factor 3	-0.0647

- 음수인 경우, **파랑**
- 양수인 경우, **빨강**

* Factor Loading table

INC	STR	CUL	WIT	SES	CMP	PRD	AGE	ESA	GD	EDU
-0.482514	0.038056	-0.528027	0.253000	0.554757	-0.324104	-0.259185	0.395685	-0.345583	-0.164507	-0.449886
0.047825	0.231324	-0.094062	0.031549	-0.290380	0.326311	0.086959	0.359734	0.032187	0.141333	-0.283642
0.296181	-0.066858	-0.364506	0.012909	-0.163148	-0.085713	-0.109765	-0.037807	-0.250989	0.015407	0.167932

* 변수 설명

```
#INC: 임금수준
#STR: 스트레스(STRESS)
#CUL: 일상생활에서 문화예술의 비중
#WIT: 함께 여가생활하는 사람
#SES: SocioEconomic Status
```

```
#CMP: 계층이동가능성(Class movement possibility)
#PRD: 서울 시민으로서의 자부심
#AGE: 연령(10단위)
#ESA: 문화환경만족도(Envinronment Satisfaction)
#GD: 강동구
#EDU: 학력(고졸 or NOT)
```

결과 해석 및 한계점
결과를 잘 해석해보자

* Factor 별 coef

	coef
Factor 1	-0.5141
Factor 2	0.1838
Factor 3	-0.0647

- 음수인 경우, 파랑
- 양수인 경우, 빨강

* Factor Loading table

INC	STR	CUL	WIT	SES	CMP	PRD	AGE	ESA	GD	EDU
-0.482514	0.038056	-0.528027	0.253000	0.554757	-0.324104	-0.259185	0.395685	-0.345583	-0.164507	-0.449886
0.047825	0.231324	-0.094062	0.031549	-0.290380	0.326311	0.086959	0.359734	0.032187	0.141333	-0.283642
0.296181	-0.066858	-0.364506	0.012909	-0.163148	-0.085713	-0.109765	-0.037807	-0.250989	0.015407	0.167932

* 변수 설명

```
#INC: 임금수준
#STR: 스트레스(STRESS)
#CUL: 일상생활에서 문화예술의 비중
#WIT: 함께 여가생활하는 사람
#SES: SocioEconomic Status
```

```
#CMP: 계층이동가능성(Class movement possibility)
#PRD: 서울 시민으로서의 자부심
#AGE: 연령(10단위)
#ESA: 문화환경만족도(Envinronment Satisfaction)
#GD: 강동구
#EDU: 학력(고졸 or NOT)
```

결과 해석 및 한계점
행복도와 3요소?!?

결과 해석

A formula for wellbeing

SWB = LS + FPA - FNA

SWB = Subjective Wellbeing

LS = Life Satisfaction

FPA = Frequent Positive Affect

FNA = Frequent Negative Affect

Factor 1와 Factor 2

한 계 점

A formula for wellbeing

SWB = LS + FPA - FNA

SWB = Subjective Wellbeing

LS = Life Satisfaction

FPA = Frequent Positive Affect

FNA = Frequent Negative Affect

Factor 3

2조 발표 끝 !

감사합니다.

Q & A

2020. 06. 04

2조

조장 : 김윤환

조원 : 백채빈 손지우 신혜연 이상완 조인식