ESC 2019 - FALL FINAL EDA

2조
서민희
손동규
안재형
유아현
정여진

Index

역할 분담

EDA 발표 전날에 조원 모두가 시험을 본 조가 있다?!

정여진(조장_선형통계 시험 봄) : 조원 진두지휘, Variable selection

유아현(수통2 시험봄) : 전처리 참여 및 PPT 제작

안재형(수통1 시험봄) : Corrplot 및 결측치 분석 코딩

서민희(무려 선형통계와 수통2를 같이 봄) : 전처리 참여

손동규(수통2 시험봄) : 전처리 참여 및 EDA 발표

Data overview

```
2215 obs. of  147 variables
```

모두 147개 변수, 2215개의 관측치

Numeric, factor variables

1. 변수가 너무 많음
2. Correlation이 너무 높음
3. 우리가 예측하고자 하는 Y 사이에도 correlation이 존재함.
4. 변수들의 단위가 상이함
5. Skewed

결측치 처리

먼저 ? 로 되어 있는 character들을 NA로 바꿈

countyCode	1221
communityCode	1224
LemasSwornFT	1872
LemasSwFTPerPop	1872
LemasSwFTFieldOps	1872
LemasSwFTFieldPer…	1872
LemasTotalReq	1872
LemasTotReqPerPop	1872
PolicReqPerOffic	1872
PolicPerPop	1872
RacialMatchCommPol	1872
PctPolicWhite	1872
PctPolicBlack	1872
PctPolicHisp	1872
PctPolicAsian	1872
PctPolicMinor	1872
OfficAssgnDrugUnits	1872
NumKindsDrugsSeiz	1872
PolicAveOTWorked	1872
PolicCars	1872
PolicOperBudg	1872
LemasPctPolicOnPatr	1872
LemasGangUnitDeploy	1872
PolicBudgPerPop	1872
rapes	208
rapesPerPop	208
robberies	1
robbbPerPop	1
assaults	13
assaultPerPop	13
burglaries	3
burglPerPop	3
larcenies	3
larcPerPop	3
autoTheft	3
autoTheftPerPop	3

결측치 처리

1. NA의 비율이 50% 이상인 변수를 구함 : 변수 147개 → 125개

	n	naratio	nacatg
LemasSwornFT		0.845	Bad
LemasSwFTPerPop		0.845	Bad
LemasSwFTFieldOps		0.845	Bad
LemasSwFTFieldPerPop		0.845	Bad
LemasTotalReq		0.845	Bad
LemasTotReqPerPop		0.845	Bad
PolicReqPerOffic		0.845	Bad
PolicPerPop		0.845	Bad
RacialMatchCommPol		0.845	Bad
PctPolicWhite		0.845	Bad
PctPolicBlack		0.845	Bad
PctPolicHisp		0.845	Bad
PctPolicAsian		0.845	Bad
PctPolicMinor		0.845	Bad
OfficAssgnDrugUnits		0.845	Bad
NumKindsDrugsSeiz		0.845	Bad
PolicAveOTWorked		0.845	Bad
PolicCars		0.845	Bad
PolicOperBudg		0.845	Bad
LemasPctPolicOnPatr		0.845	Bad
LemasGangUnitDeploy		0.845	Bad
PolicBudgPerPop		0.845	Bad
communityCode		0.553	Bad
countyCode		0.551	Bad

→ CommunityCode와 countyCode는 지역 코드로, 필요하게 될 수도 있으니 삭제하지 않고 이를 제외한 모든 변수를 삭제함

Deleting variables

2. Intuition & corrplot를 참고해 변수 제거 : 변수 125개 → 101개

agePct12t29
agePct16t24
numbUrban
NumUnderPov
TotalPctDiv
NumKidsBornNeverMar
PctImmigRecent

PctImmigRec5
PctRecImmig8
PctRecentImmig
PctSpeakEnglOnly
OwnOccLowQuart
OwnOccHiQuart

OwnOccQrange
PctImmigRec8
PctRecImmig5
RentLowQ
RentHighQ
RentQrange

Delete variables

Corrplot 일부

	fold	population	householdsize	racepctblack	racePctWhite	racePctAsian	racePctHisp	agePct12t21	agePct12t29
fold	1	-0.0443376	0.0159732	-0.0400639	0.0229727	0.00443861	0.0356202	-0.015715	-0.0237936
population	-0.0443376	1	-0.0188408	0.135641	-0.184685	0.0883603	0.0940484	-0.00762359	0.0462473
householdsize	0.0159732	-0.0188408	1	-0.0474552	-0.230117	0.186779	0.485178	0.491481	0.384322
racepctblack	-0.0400639	0.135641	-0.0474552	1	-0.820605	-0.0893003	-0.0639113	0.0970533	0.119994
racePctWhite	0.0229727	-0.184685	-0.230117	-0.820605	1	-0.276474	-0.408489	-0.142252	-0.208286
racePctAsian	0.00443861	0.0883603	0.186779	-0.0893003	-0.276474	1	0.198439	-0.00860939	0.0703418
racePctHisp	0.0356202	0.0940484	0.485178	-0.0639113	-0.408489	0.198439	1	0.109992	0.157164
agePct12t21	-0.015715	-0.00762359	0.491481	0.0970533	-0.142252	-0.00860939	0.109992	1	0.872338
agePct12t29	-0.0237936	0.0462473	0.384322	0.119994	-0.208286	0.0703418	0.157164	0.872338	1
agePct16t24	-0.022966	0.0176999	0.317889	0.0896884	-0.124263	0.037181	0.0540845	0.934932	0.946245
agePct65up	-0.00509701	-0.0438191	-0.578092	0.0375028	0.114381	-0.205774	-0.18211	-0.36025	-0.476703
numbUrban	-0.0444606	0.999052	-0.0192211	0.135202	-0.185192	0.095836	0.0941379	-0.017016	0.0406219
pctUrban	0.00119779	0.114299	-0.0114676	0.0158792	-0.0595047	0.22515	0.0276786	-0.225339	-0.101083
medIncome	0.029295	-0.0532395	0.179445	-0.348644	0.305698	0.311156	-0.151881	-0.270557	-0.329834
pctWWage	0.0113244	-0.0127349	0.416197	-0.240813	0.145876	0.253748	-0.0015341	0.104831	0.235642
pctWFarmSelf	0.00808507	-0.0671774	0.16211	-0.154586	0.103766	-0.0748389	0.0782012	0.215821	0.114461
pctWInvInc	0.009008	-0.0859295	-0.158514	-0.504866	0.599697	0.132789	-0.412157	-0.206551	-0.291238
pctWSocSec	-0.00620505	-0.0542476	-0.438959	0.111614	0.0422358	-0.300208	-0.11878	-0.232179	-0.389108
pctWPubAsst	-0.0154497	0.107935	0.113723	0.484314	-0.598154	-0.0476518	0.399804	0.160753	0.151114
pctWRetire	-0.0020426	-0.0593528	-0.327937	-0.0518068	0.191051	-0.16027	-0.232337	-0.276583	-0.393896
medFamInc	0.0310801	-0.0573532	0.0929846	-0.35038	0.337388	0.294576	-0.209634	-0.221128	-0.279797

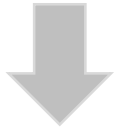
Percentile로 바꾸기

NumImmig	0.030862678	0.293805569
0.106594324	0.046308432	0.042764398
0.083034208	0.014336627	0.031275331
0.050027263	0.050787458	0.024111909
0.020353026	0.097076432	0.083689676
0.017429969	0.061860258	0.017135024
0.014883198	0.006617361	0.065581527
0.091881533	...	0.188967554
0.008695067	0.006680422	0.022447147
0.019889085	0.134784105	0.073543748
0.046268945	0.034249303	0.022809478
0.029682605	0.204652220	0.164890304

‘NumImmg’을 ‘population’으로 나누어 count data를 percentile로 바꿈

Zero값에 대한 판단

'NumInShelter'을 'NumStreet'과 merge한 후 'population'으로 나누려던 중 homeless=0인 도시가 1118개 임을 발견함



의미 있는 0인지 아닌지에 대한 판단 필요

```
zero_rat[zero_rat > 0.1]
```

pctUrban	0.269977
NumInShelters	0.557111
NumStreet	0.739052
LemasPctOfficDrugUn	0.849661
murders	0.463205
murdPerPop	0.463205

pctUrban은 대부분 0 혹은 100의 값을 가짐

Murder 또한 0이 46%지만, 'murder'= 1인 값이 많아 의미 있는 0이라고 판단함

Variable selection

'Violent Crimes'와 'Non-Violent Crimes'는 각 4개의 범죄와 **Sum의 관계**를 가짐

=> 이 둘과 관련이 있는 변수는 나머지 'Crimes'와도 관련이 있을 것

Variable selection

이 두 변수를 Y변수로 설정하여 각 Y에 대해 Variable Selection 진행

선택된 변수들의 Set을 비교하여 최종 변수 설정

Divide into train and test set

train set

결측치가 존재하지 않는 row

→ train, validation으로 나눌 예정

Test set

Y 변수 중 적어도 하나의 NA를 포함하고 있는 row (318개)

Variable selection – Boruta (random forest를 기반으로)

```
# Do a tentative rough fix
roughFixMod <- TentativeRoughFix(boruta_output)
boruta_signif <- getSelectedAttributes(roughFixMod)
print(boruta_signif)
```

```
 [1] "householdsize"       "racepctblack"        "racePctWhite"
 [4] "racePctAsian"        "racePctHisp"         "agePct12t21"
 [7] "agePct65up"          "medIncome"           "pctWWage"
[10] "pctWFarmSelf"        "pctWInvInc"          "pctWSocSec"
[13] "pctWPubAsst"         "medFamInc"           "perCapInc"
[16] "whitePerCap"         "blackPerCap"         "indianPerCap"
[19] "OtherPerCap"         "HispPerCap"          "PctPopUnderPov"
[22] "PctLess9thGrade"     "PctNotHSGrad"        "PctBSorMore"
[25] "PctUnemployed"       "PctEmploy"           "PctOccupManu"
[28] "PctOccupMgmtProf"    "MalePctDivorce"      "MalePctNevMarr"
[31] "FemalePctDiv"        "PersPerFam"          "PctFam2Par"
[34] "PctKids2Par"         "PctYoungKids2Par"    "PctTeen2Par"
[37] "PctWorkMomYoungKids" "PctWorkMom"          "PctKidsBornNeverMar"
[40] "NumImmig"            "PctImmigRec10"       "PctRecImmig10"
[43] "PctNotSpeakEnglWell" "PctLargHouseFam"     "PctLargHouseOccup"
[46] "PersPerOccupHous"    "PersPerOwnOccHous"   "PersPerRentOccHous"
[49] "PctPersDenseHous"    "PctHousLess3BR"      "HousVacant"
[52] "PctVacMore6Mos"      "MedYrHousBuilt"      "PctHousNoPhone"
[55] "OwnOccMedVal"        "RentMedian"          "MedRent"
[58] "MedRentPctHousInc"   "MedOwnCostPctInc"    "PctForeignBorn"
[61] "PctBornSameState"    "PctSameHouse85"      "PctSameCity85"
[64] "PctSameState85"      "LandArea"            "PopDens"
```

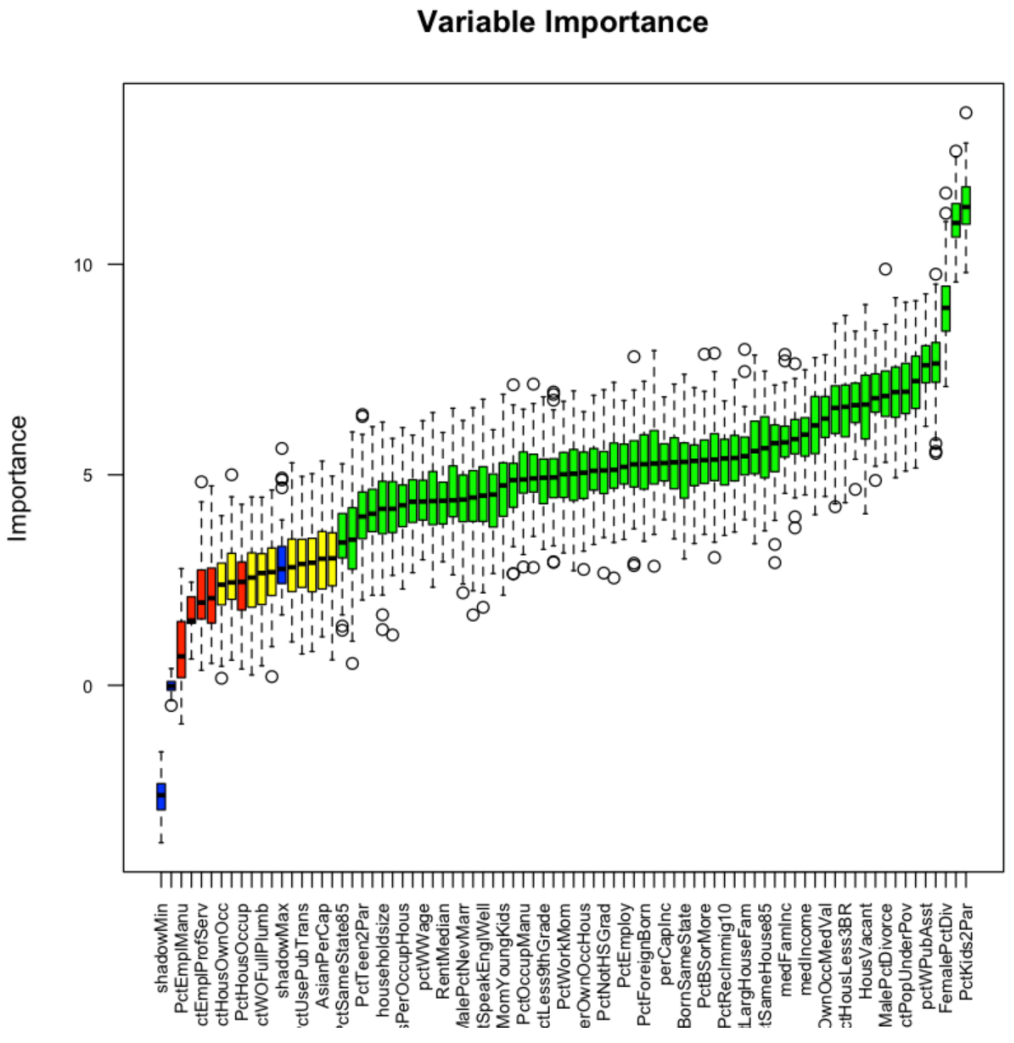
→ 실행한 결과
66개의 변수가 선택됨

Variable selection – Boruta

각 변수들의 Importance를 알 수 있다는 것이
이 방법의 가장 큰 장점

```
# Variable Importance Scores
imps <- attStats(roughFixMod)
imps2 = imps[imps$decision != 'Rejected', c('meanImp', 'decision')]
head(imps2[order(-imps2$meanImp), ],30)  # descending sort
```

	meanImp	decision
PctKids2Par	11.348909	Confirmed
PctFam2Par	11.034860	Confirmed
FemalePctDiv	9.057023	Confirmed
racePctWhite	7.657073	Confirmed
pctWPubAsst	7.600004	Confirmed
PctPersDenseHous	7.106061	Confirmed
PctPopUnderPov	7.053748	Confirmed
LandArea	6.993515	Confirmed
MalePctDivorce	6.914379	Confirmed
whitePerCap	6.871163	Confirmed
HispPerCap	6.684065	Confirmed
HousVacant	6.613172	Confirmed
racePctHisp	6.505338	Confirmed
PctHousLess3BR	6.486072	Confirmed
OwnOccMedVal	6.369884	Confirmed
agePct65up	6.146521	Confirmed
medIncome	5.923528	Confirmed
agePct12t21	5.855716	Confirmed



Variable selection – forward stepwise selection

✓ **Stepwise selection?**

: 처음 모델에서 변수를 하나씩 빼거나 더하며 가장 이상적인 모델을 고르는 방법

1) 단순한 모델에서 시작하여 유의한 변수를 늘리는 Forward selection
2) 모든 변수를 포함한 모델에서 시작, 변수를 제거하는 Backward selection
3) Forward와 Backward를 결합한 Iterative selection이 존재함

: 마지막 방법이 이상적이나, 변수가 너무 많아 가장 단순한 Forward selection으로 진행

Variable selection – forward stepwise selection

```
# Step 1: Define base intercept only model
base.mod <- lm(nonViolPerPop ~ 1 , data=train.set.lm1)
```

```
# Step 2: Full model with all predictors
all.mod <- lm(nonViolPerPop ~ . , data=train.set.lm1)
```

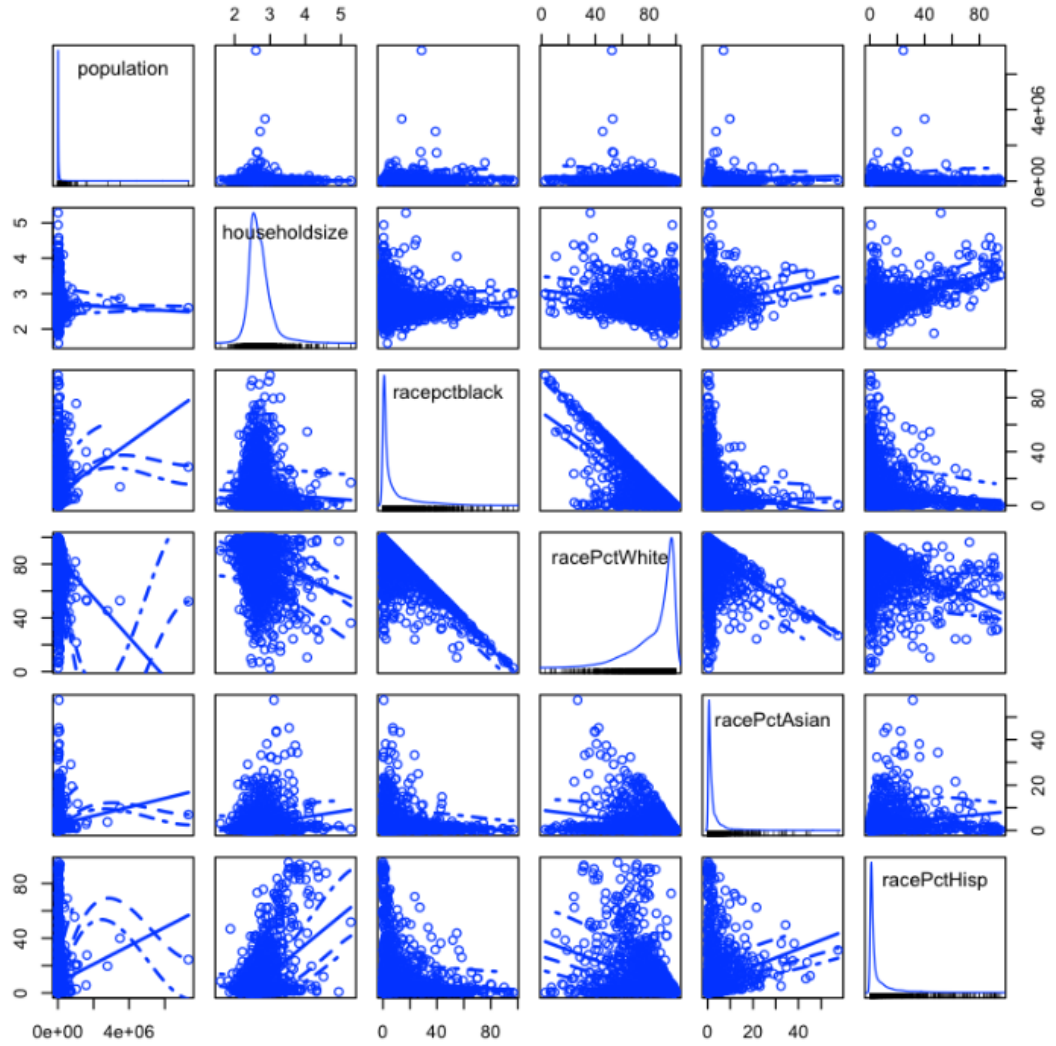
```
ep 3: Perform step-wise algorithm. direction='both' implies both forward and backward stepwise
Mod <- step(base.mod, scope = list(lower = base.mod, upper = all.mod), direction = "forward", trace = 0, steps = 1000)
```

```
# Step 4: Get the shortlisted variable.
shortlistedVars <- names(unlist(stepMod[[1]]))
shortlistedVars <- shortlistedVars[!shortlistedVars %in% "(Intercept)"] # remove intercept

# Show
print(shortlistedVars)
```

```
 [1] "FemalePctDiv"        "PctPopUnderPov"      "medIncome"
 [4] "racePctAsian"        "HousVacant"          "OtherPerCap"
 [7] "PctWOFullPlumb"      "racePctHisp"         "PctPersDenseHous"
[10] "racePctWhite"        "PctKidsBornNeverMar" "PctKids2Par"
[13] "PctImmigRec10"       "PctHousOccup"        "LandArea"
[16] "PctTeen2Par"         "householdsize"       "PctLargHouseOccup"
[19] "PctSameHouse85"      "OwnOccMedVal"        "PctUnemployed"
[22] "perCapInc"           "blackPerCap"         "PctBSorMore"
[25] "PctYoungKids2Par"    "whitePerCap"         "PctSameCity85"
```

skewness



왼쪽 그림에서 볼 수 있듯이
그래프들이 많이 skew되어있음

→ Log / Square 변환

Variable Selection / 앞으로 보완할 점들

1) Y = '강력범죄'에 대해서도 같은 방법으로 변수 선택

2) 앞의 두 가지 변수 선택 후, 겹치는 변수 확인

3) 다른 Selection 방법은 없을까? (Lasso, Caret 등)

Q&A