지도 정형 2조

주택가격예측

심소정, 송민, 이인성, 조혜영, 최은혁

목차

주택 가격 예측? -> 가격 협상의 효율성!

(김철수, 고객)

집을 좀 사려고 하는데요..

평수는.. 차고는.. 방 개수는.. 담장은... 위치는... 화장실 개수는...
가격은 얼마정도 하나요??

어떤 집을 찾고 계신가요?

(김모씨, 중개업자)

아.. 잠시만요 비슷한 조건들을 보니 대략 얼마정도 입니다

-> 주관적, 부정확!

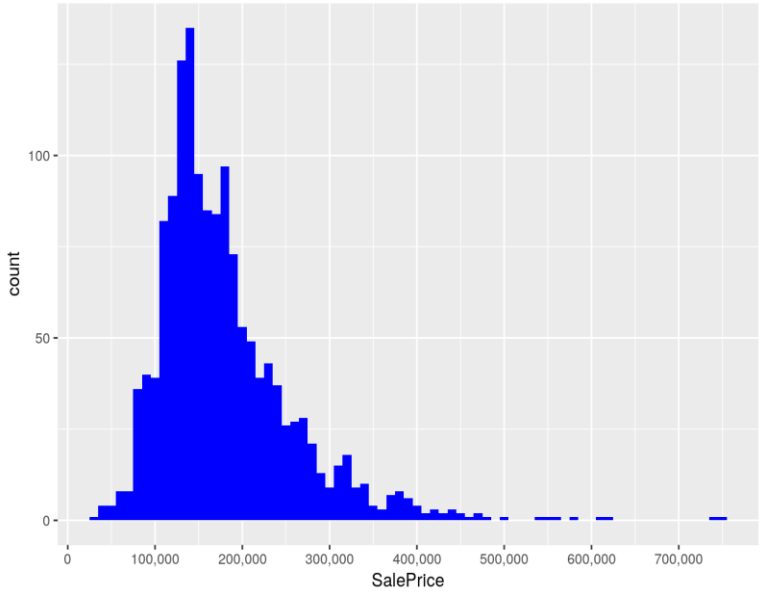
-> 주택 가격을 예측해주는 객관적, 정확한 모델을 만들어보자!

데이터 소개

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

4,389 teams · Ongoing

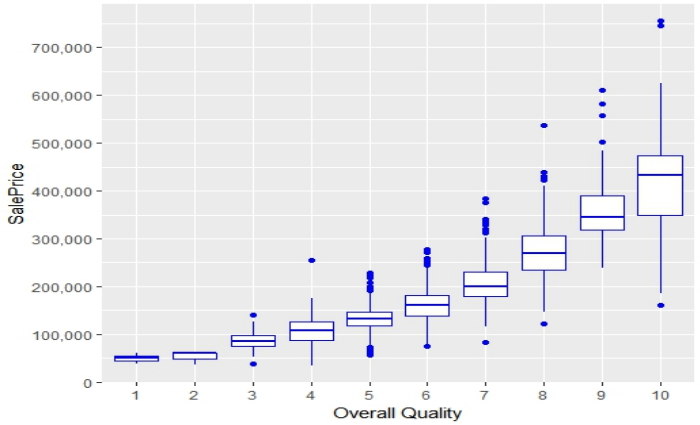


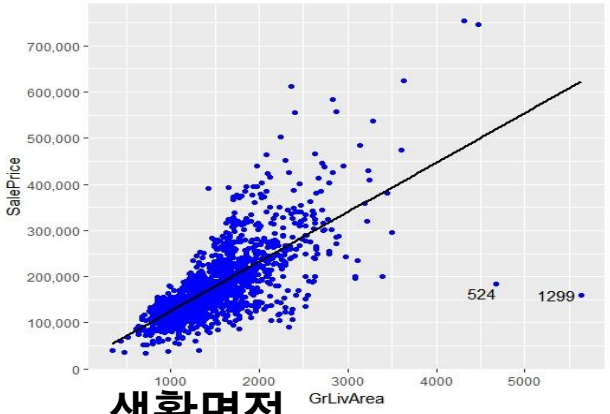
-목적변수: 집 판매 가격 (Sale Price)

-설명변수: 79개 (ex : 집의 용도, 바닥재, 풀장, 복도, 차고크기,..등등)

-총 2915개의 observations

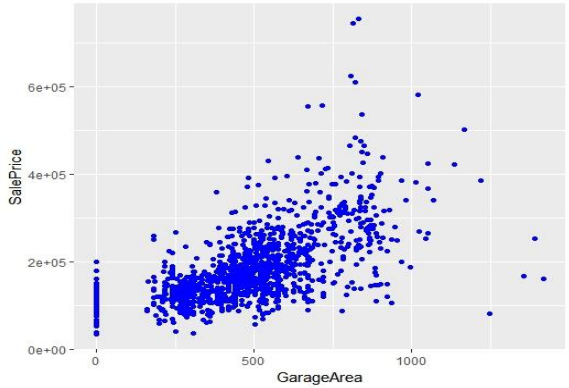
MSSubClas	MSZoning	otFrontag	LotArea	Street	Alley	LotShape	andContou	LotConfig	LandSlope	eighborho	Condition1	Condition2	BldgType	louseStyl
60	RL	65	8450	1	None	3	Lvl	Inside	2	CollgCr	Norm	Norm	1Fam	2Story
20	RL	80	9600	1	None	3	Lvl	FR2	2	Veenker	Feedr	Norm	1Fam	1Story
60	RL	68	11250	1	None	2	Lvl	Inside	2	CollgCr	Norm	Norm	1Fam	2Story
70	RL	60	9550	1	None	2	Lvl	Corner	2	Crawfor	Norm	Norm	1Fam	2Story
60	RL	84	14260	1	None	2	Lvl	FR2	2	NoRidge	Norm	Norm	1Fam	2Story
50	RL	85	14115	1	None	2	Lvl	Inside	2	Mitchel	Norm	Norm	1Fam	1.5Fin
20	RL	75	10084	1	None	3	Lvl	Inside	2	Somerst	Norm	Norm	1Fam	1Story

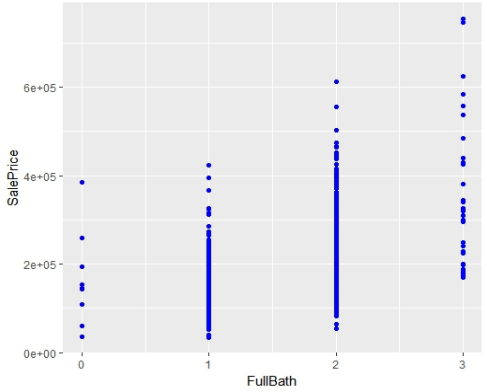




Overall Quality 가 높을 수록 집값 비싸짐

생활면적





주차 면적

화장실 개수

데이터 전처리 - 결측치 처리

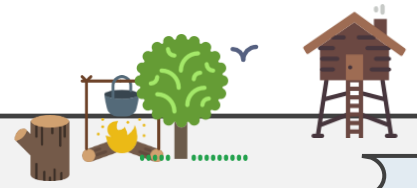
```
> NAcol <- which(colSums(is.na(all)) > 0)
> sort(colSums(sapply(all[NAcol], is.na)), decreasing = TRUE)
      PoolQC  MiscFeature        Alley        Fence    SalePrice  FireplaceQu
        2909         2814         2721         2348         1459         1420
 LotFrontage  GarageYrBlt GarageFinish   GarageQual   GarageCond   GarageType
         486          159          159          159          159          157
    BsmtCond BsmtExposure     BsmtQual BsmtFinType2 BsmtFinType1   MasVnrType
          82           82           81           80           79           24
  MasVnrArea     MSZoning    Utilities BsmtFullBath BsmtHalfBath   Functional
          23            4            2            2            2            2
 Exterior1st  Exterior2nd   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF
           1            1            1            1            1            1
  Electrical  KitchenQual   GarageCars   GarageArea     SaleType
           1            1            1            1            1
> length(NAcol)
[1] 35
```

34개 변수에서 결측치 발견

-> 0을 NA로 표기한 경우가 다수

ex) 풀장이 없다 -> NA

담장이 없다 -> NA



데이터 전처리-결측치 처리

변수	방법
Pool Quality	NA -> 0으로
MiscFeature	NA -> 0으로
Alley	NA -> 0으로
Fence	NA-> 0으로
Fireplace quality	NA->0
Garage	garageYrBlt->YrBlt로 Garage 와 관련된 대부분 변수들이 157개의 결측치를 보임. 다 같은 observation. NA->0 159개의 결측치를 보이는 변수는 2개의 ROW를 삭제 했음 Garage Quality, Garagefinish -> ordinal
Basement	Garage관련 변수들 전처리와 같은 과정으로 진행
그 외	Year sold, Month sold -> factor 로

-범주형 변수
-> 대부분이 quailty를 의미하는 데이터

ex>Pool Quality는
Excellent, Good,..,Fair의 범주를 가지고 있고,
결측치에 “No pool”범주를 지정,

-> quailty를 의미하는 변수간에 범주가 제각각인 문제 발생
-> 이후 분석의 편의를 위해 0,1,..값을 갖는 integer로 변환하였음

or 최빈값을 넣음

-수치형 변수
->median값을 넣음
-Scaling진행

변수 선택 과정
논문 참고_주택 가격 결정요인 & 모델링 공부

헤도닉가격모형을활용한주택가격 결정요인에관한연구

머신러닝 vs 시계열 부동산 가격 예측

서울부동산예측_시공간변수 이용

헤도닉 가격모형과 공간분석기법을 통한 주거지역 가격추정

〈표 3-1〉 주택가격 결정요인 유형1

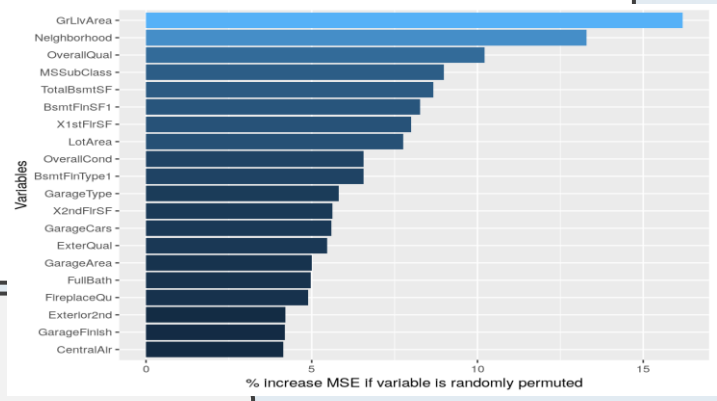
구분	변수명	단위	설명	비고
주택특성	공급면적(X_1)	㎡	전용면적과 주거공용면적의 합	
	전용면적(X_2)	㎡	주거공용면적과 공급면적으로 배기	
	방수(X_3)	개	주택 내 방의 개수	
	욕실수(X_4)	개	주택 내 욕실의 개수	
	주차대수(X_5)	대	가구당 평균주차대수	
	경과년수(X_6)	년	입주년도와 연구년도로 배기	
이웃특성	세대수(X_7)	세대	한 아파트단지의 총 세대수	
	최고층(X_8)	층	아파트단지 내의 최고층수	
	평균관리비(X_9)	원	각각 아파트 단지의 평균 관리비용	
	건설사 지명도(X_{10})	더미	건설사 순위 전10위 여부	더미 변수
	한강까지 거리(X_{11})	m	표본과 한강까지의 거리	
	국립중앙박물관과의 거리(X_{12})	m	표본과 국립중앙박물관 정문까지의 거리	
접근성	초등학교와의 거리(X_{13})	m	표본과 가장 가까운 초등학교 정문까지의 거리	
	중학교와의 거리(X_{14})	m	표본과 가장 가까운 중학교 정문까지의 거리	

변수 선택 과정 : Stepwise & Backward & Forward & RandomForest 비교

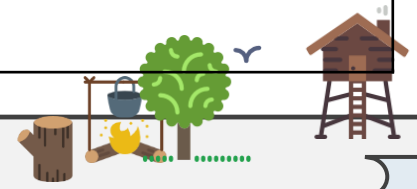
```
> summary(stepw)

Call:
lm(formula = SalePrice ~ LotArea + LandSlope + OverallQual +
    OverallCond + YearBuilt + MSSubClass + LandContour + Neighborhood +
    Condition2 + MasVnrType + BsmtCond + BsmtExposure + BsmtFinSF1 +
    BsmtFinSF2 + BsmtUnfSF + CentralAir + RoofMatl + Heating +
    X1stFlrSF + X2ndFlrSF + FullBath + KitchenAbvGr + KitchenQual +
    TotRmsAbvGrd + Functional + GarageYrBlt + GarageArea + PavedDrive +
    ScreenPorch + PoolArea + MiscFeature + MiscVal + MoSold +
    SaleType, data = data_s)
```

```
> summary(forward)

Call:
lm(formula = SalePrice ~ LotFrontage + LotArea + LotShape + LandSlope +
    OverallQual + OverallCond + YearBuilt + YearRemodAdd + MSSubClass +
    MSZoning + Street + Alley + LandContour + LotConfig + Neighborhood +
    Condition1 + Condition2 + BldgType + HouseStyle + RoofStyle +
    MasVnrType + MasVnrArea + ExterQual + ExterCond + BsmtQual +
    BsmtCond + BsmtExposure + BsmtFinType1 + BsmtFinSF1 + BsmtFinType2 +
    BsmtFinSF2 + BsmtUnfSF + TotalBsmtSF + HeatingQC + CentralAir +
    RoofMatl + Exterior1st + Exterior2nd + Foundation + Heating +
    Electrical + X1stFlrSF + X2ndFlrSF + LowQualFinSF + GrLivArea +
    BsmtFullBath + BsmtHalfBath + FullBath + HalfBath + BedroomAbvGr +
    KitchenAbvGr + KitchenQual + TotRmsAbvGrd + Functional +
    Fireplaces + FireplaceQu + GarageYrBlt + GarageFinish + GarageCars +
    GarageType + GarageArea + GarageQual + GarageCond + PavedDrive +
    WoodDeckSF + OpenPorchSF + EnclosedPorch + X3SsnPorch + ScreenPorch +
    PoolArea + PoolQC + Fence + MiscFeature + MiscVal + MoSold +
    YrSold + SaleType + SaleCondition, data = data_s
```

```
> summary(backward)

Call:
lm(formula = SalePrice ~ LotArea + LandSlope + OverallQual +
    OverallCond + YearBuilt + MSSubClass + LandContour +
Neighborhood +
    Condition2 + MasVnrType + BsmtCond + BsmtExposure +
BsmtFinSF1 +
    BsmtFinSF2 + BsmtUnfSF + CentralAir + RoofMatl + Heating +
    X1stFlrSF + X2ndFlrSF + FullBath + KitchenAbvGr + KitchenQual
+
    TotRmsAbvGrd + Functional + GarageYrBlt + GarageArea +
PavedDrive +
    ScreenPorch + PoolArea + MiscFeature + MiscVal + MoSold +
    SaleType, data = data_s)
```



범주	변수	설명	변수	설명
면적	LotArea	주차장 면적	X1stFlrSF	1층 면적
	GarageArea	창고 면적	X2ndFlrSF	2층 면적
	MSSubClass	몇층 집		
질	OverallQual	전반적 질	BsmtCond	지하실 질
	OverallCond	전반적 조건	BsmtExposure	집 외부 노출 정도
	KitchenQual	부엌 종류	BsmtFinSF1	지하실 인테리어 필요 TYPE1
	KitchenAbvGr	부엌 질 등급	BsmtFinSF2	지하실 인테리어 필요 TYPE1
	FullBath	변기, 샤워기, 욕조 있음	TotRmsAbvGrd	질 좋은 방 수
특징	MasVnrType	벽 구성	ScreenPorch	좋은 현관
	PoolArea	수영장 넓이	PavedDrive	도로 정비됨
	Heating	난방 종류	CentralAir	중앙 에어컨
	RoofMatl	지붕 종류	Functional	세금 감면 정도
	YearBuilt	건축 년도	LandSlope	집 경사
입지	Neighborhood	주요 도시 주변	Condition1/2	입지 특징1/2



변수 선택

모델링

Model 1 : Lasso Regression

Model 2: Ridge Regression

Model 3: Elastic Net

Model 4 : XgBoost

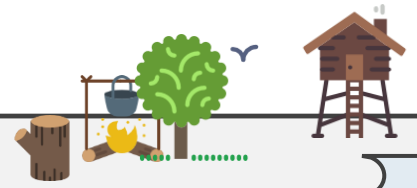
Lasso Regression

```
> lasso.coef ###
31 x 1 sparse Matrix of class "dgCMatrix"
                             1
(Intercept)      5.788152e+03
LotArea          .
LandSlope        .
OverallQual      .
OverallCond      .
YearBuilt        .
MSSubClass       .
Neighborhood     .
Condition1       .
Condition2       .
MasVnrType       .
BsmtCond         .
BsmtExposure     .
BsmtFinSF1       .
BsmtFinSF2       .
CentralAir       .
RoofMatl         .
Heating          .
X1stFlrSF        .
X2ndFlrSF        .
FullBath         .
KitchenAbvGr     .
KitchenQual      .
TotRmsAbvGrd     .
Functional       .
GarageArea       .
PavedDrive       .
ScreenPorch      .
PoolArea         .
```

-Lasso 회귀는 우리가 선택한 변수들을 모두 중요하지 않은 변수로 판단해 coefficient를 0으로 만듬.

-why? 다중공선성 문제

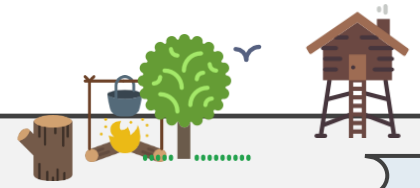
-Lasso는 적합하지 않은 모델



Ridge Regression

```
> coef(cv.ridge)
30 x 1 sparse Matrix of class "dgCMatrix"
                             1
(Intercept)          5.0546208601
LotArea              0.0066666058
LandSlope           -0.0023522482
OverallQual          0.0438178578
OverallCond          0.0141216917
YearBuilt            0.0255227211
MSSubClass          -0.0018184295
Neighborhood         0.0004769196
Condition1           0.0019400735
Condition2          -0.0151874800
MasVnrType           0.0025965888
BsmtCond             0.0053207772
BsmtExposure         0.0093253371
BsmtFinSF1           0.0124584147
BsmtFinSF2           0.0034619345
CentralAir           0.0090831127
RoofMatl             0.0042443907
Heating              0.0032342254
X1stFlrSF            0.0311077032
X2ndFlrSF            0.0255245700
FullBath             0.0137048000
KitchenAbvGr        -0.0058561758
KitchenQual          0.0177305564
TotRmsAbvGrd         0.0180198469
Functional           0.0073348838
GarageArea           0.0184525340
PavedDrive           0.0053306585
ScreenPorch          0.0075498147
PoolArea            -0.0044893092
```

Coefficient의 문제점
-Pool면적과 집값은 반비례?
-Condition이 나쁠 수록
집값이 높다? 등등



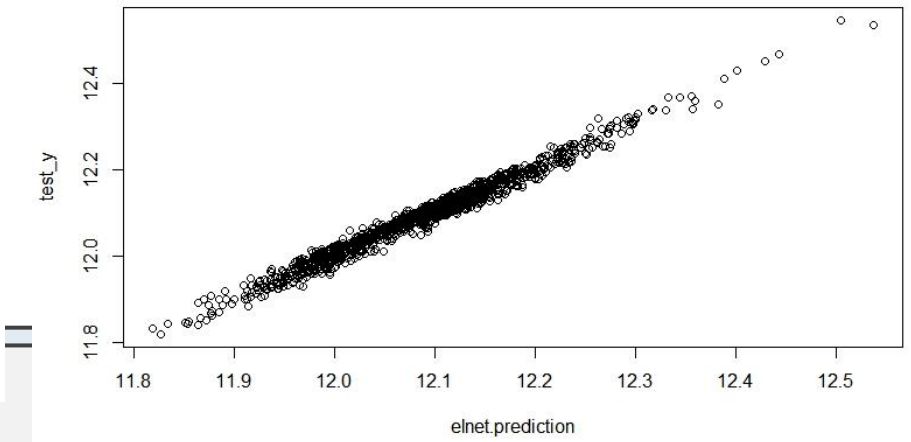
Elastic Net

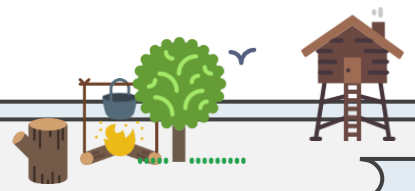
```
> coef(cv.elnet)
30 x 1 sparse Matrix of class "dgCMatrix"
                             1
(Intercept)     0.6883841577
LotArea         .
LandSlope       .
OverallQual     0.0100718569
OverallCond     .
YearBuilt       .
MSSubClass      .
Neighborhood    .
Condition1      .
Condition2      .
MasVnrType      .
BsmtCond        .
BsmtExposure    .
BsmtFinSF1      .
BsmtFinSF2      .
CentralAir      .
RoofMatl        .
Heating         .
X1stFlrSF       0.0013696874
X2ndFlrSF       .
FullBath        0.0002242323
KitchenAbvGr    .
KitchenQual     0.0009027097
TotRmsAbvGrd    0.0003518415
Functional      .
GarageArea      0.0019666008
PavedDrive      .
ScreenPorch     .
PoolArea        .
```

Lasso 와 Ridge를 합한 형태

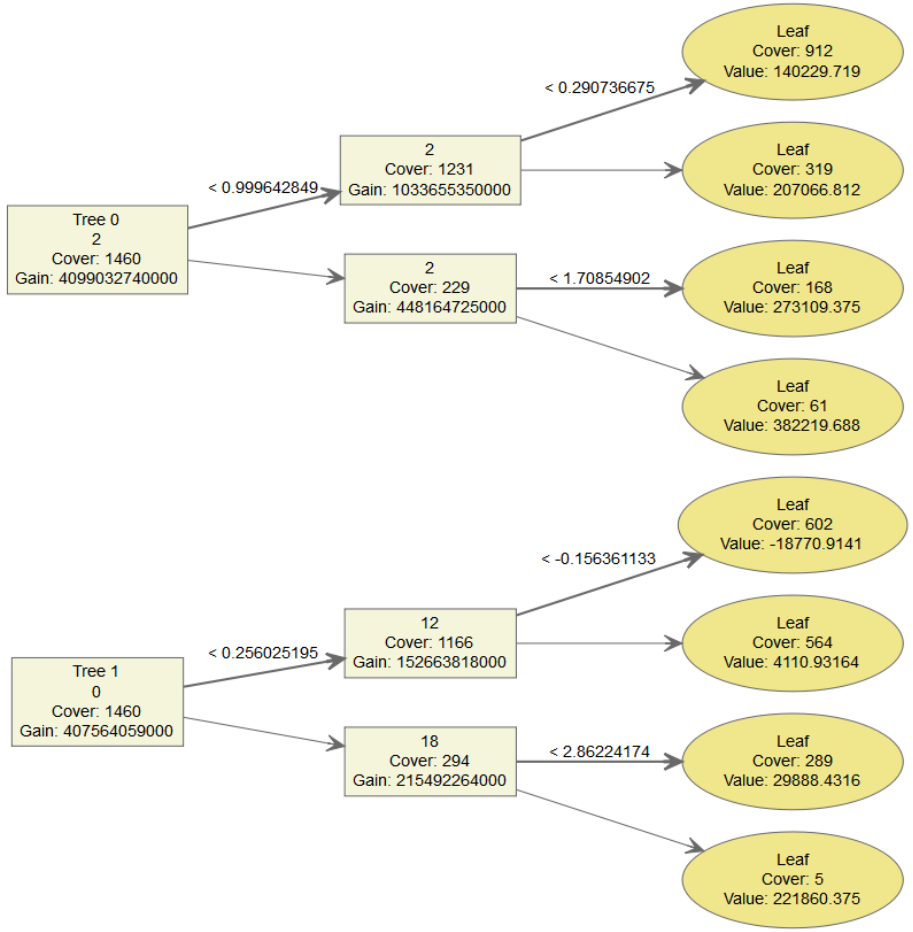
Elastic net은 설명변수 간의 강한 상관 관계 때문에 없어지는 feature는 냅두고 response랑 관계 없는 feature만 걸러준다!

-> 다중공선성이 발생하는 우리 데이터에 적합!





XG Boost



다중공선성에 영향을 받지 않는 모델링
->시도
->결과가 좋지 않았다

주택 값 추정 : 설명변수의 값들

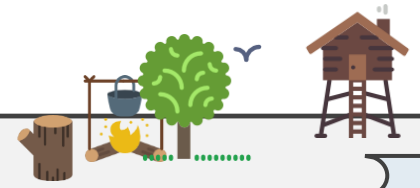
범주	변수	값	변수	값
면적	**LotArea**	8450	**X1stFlrSF**	856
	GarageArea	548	**X2ndFlrSF**	854
	MSSubClass	60		
질	**OverallQual**	7	**BsmtCond**	3
	OverallCond	5	**BsmtExposure**	1
	KitchenQual	4	**BsmtFinSF1**	706
	KitchenAbvGr	1	**BsmtFinSF2**	0
	FullBath	2	**TotRmsAbvGrd**	8
특징	**MasVnrType**	1	**ScreenPorch**	0
	PoolArea	0	**PavedDrive**	2
	Heating	GasA	**CentralAir**	1
	RoofMatl	CompShg	**Functional**	7
	YearBuilt	2003	**LandSlope**	2
입지	**Neighborhood**	CollgCr	**Condition1/2**	Norm / Norm

모델링 별 주택 값 추정 결과

	예측값	MSE
실제 y값	169277.1 (달러)	
Xgboost	135067.89	1.642789
Ridge	161943.9	0.02165925
Lasso	169649.6	0.00001283382
Elastic Net	169142.1	0.0001930559

-> 모든 변수의 coefficient를 0로 만든 lasso를 제외하고

-> Elastic Net 의 mse가 가장 낮은 것을 확인



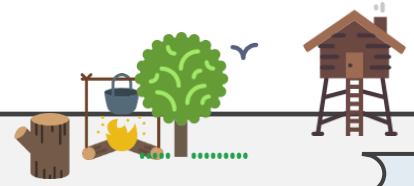
변수 선택 없이 바로 elastic net을 진행 했다면..?

기존 데이터 -> Elastic Net

(Intercept)	0.0001963949
BsmtFinType1	0.0101648927
BsmtFinSF1	0.9561006165
TotalBsmtSF	0.0011895698
BsmtFullBath	0.0054658679

변수 선택 후 -> Elastic Net

(Intercept)	0.6883841577
OverallQual	0.0100718569
X1stFlrSF	0.0013696874
FullBath	0.0002242323
KitchenQual	0.0009027097
TotRmsAbvGrd	0.0003518415
GarageArea	0.0019666008



프로젝트 한계

- 특정 시점의 가격이다. (시점이 변화하면 모델링을 새롭게 해야 함)
- 변수 선택이 완벽하지 않음.
- 입지/ 금리/ 정책에 관한 변수가 부족해 완벽한 모델이 아님.

감사합니다 ♥