# 보스턴 주택 가격(SalePrice) 예측

---

- 분석 데이터 : 보스턴 집값 데이터 (Boston House Prices Data)
- 데이터 형태 : 1460개의 행, 81개의 열
- 분석 목적 : 보스턴 주택 가격(SalePrice)를 예측하고자 함.

  이에 따라 해당 데이터에서 주어진 변수 사용함.
- 탐색적 데이터 분석 및 데이터 전처리 진행 : 결측치 및 이상치 처리 / 전체적인 변수 변환 /

  상관계수 분석 / 시각화 등 (코드 참조)
- (전처리 후)

  훈련 데이터 셋 : 80%

  테스트 데이터 셋 : 20%

  X_train : (1166, 270)

  y_train : (292, 270)

# 모델 설계 과정

## 1. LinearRegression (선형회귀모델)

모형에 대한 RMSE 값과 R squared는 다음과 같다.

| | LinearRegression |
|---|---|
| RMSE | 0.116 |
| Train R^2 score | 0.947 |
| Test R^2 score | 0.921 |

: 모델 평가 지표로써 RMSE를 활용하였으며,
선택된 최종 독립변수들이 목적변수인 'SalePrice'에 어느 정도의
설명력을 가지는지에 대해서는 R squared값으로 확인하고자 함.

---

## 2. Ridge, Lasso (규제화 된 선형회귀모델)

| | Ridge | Lasso |
|---|---|---|
| RMSE | 0.108 | 0.171 |
| Train R^2 score | 0.947 | 0.816 |
| Test R^2 score | 0.921 | 0.800 |

# 모델 설계 과정

Ridge, Lasso 두 개의 모델이 갖는 최적 평균 RMSE 값을 찾고자 하였으며, 이때의 최적 alpha 값을 하이퍼파라미터 튜닝을 통해 찾고자 함.

**Ridge_Params = [0.05, 0.1, 1, 5, 8, 10, 12, 15, 20]**

**Lasso_Params = [0.001, 0.005, 0.008, 0.05, 0.03, 0.1, 0.5, 1,5, 10]**

**Alpha 값**

**Ridge : 12**

**Lasso : 0.001**

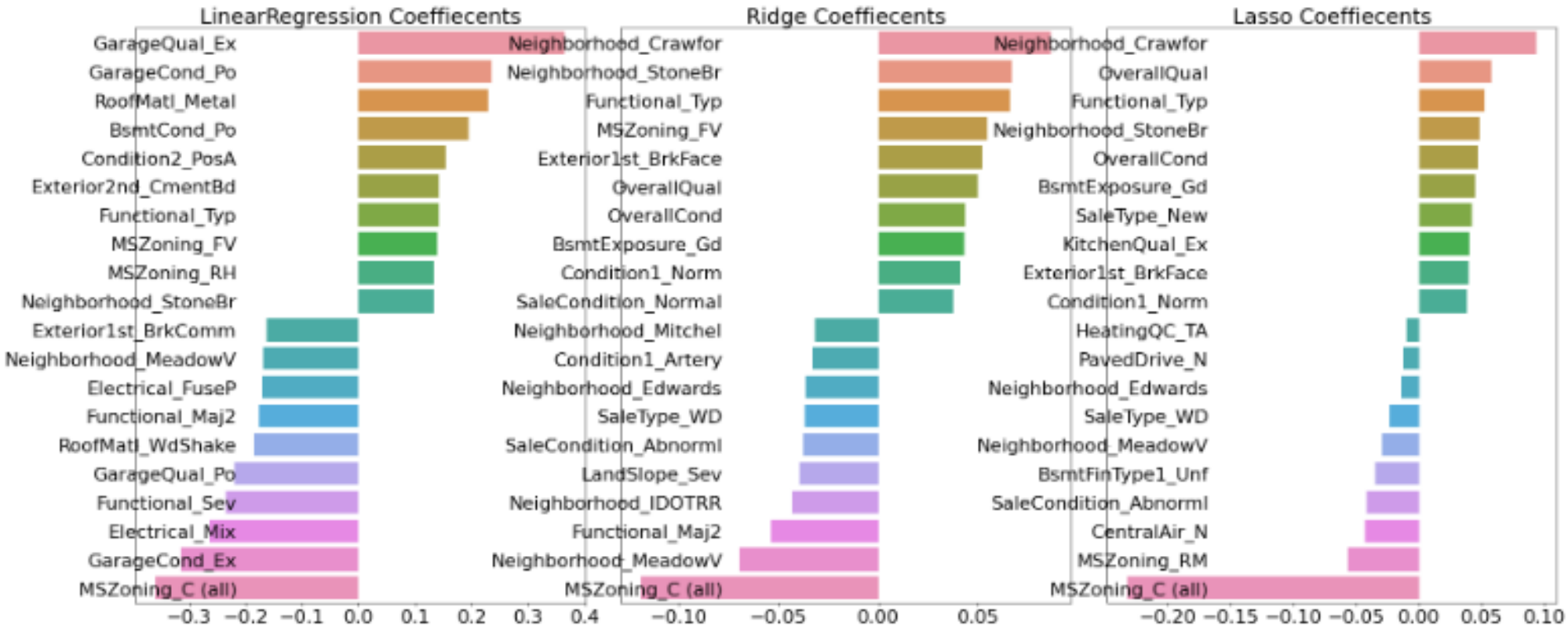
최종적으로 선택한 Ridge와 Lasso 모형의 평가지표 값 (RMSE)과 R squared는 다음과 같다.

| | LinearRegression | Ridge | Lasso |
|---|---|---|---|
| RMSE | 0.116 | 0.103 | 0.103 |
| Train R^2 score | 0.947 | 0.940 | 0.931 |
| Test R^2 score | 0.921 | 0.927 | 0.928 |

: 하이퍼파리미터 튜닝을 진행하기 이전에 비해 Ridge 모델에서는 별다른 차이를 확인할 수 없었으나,

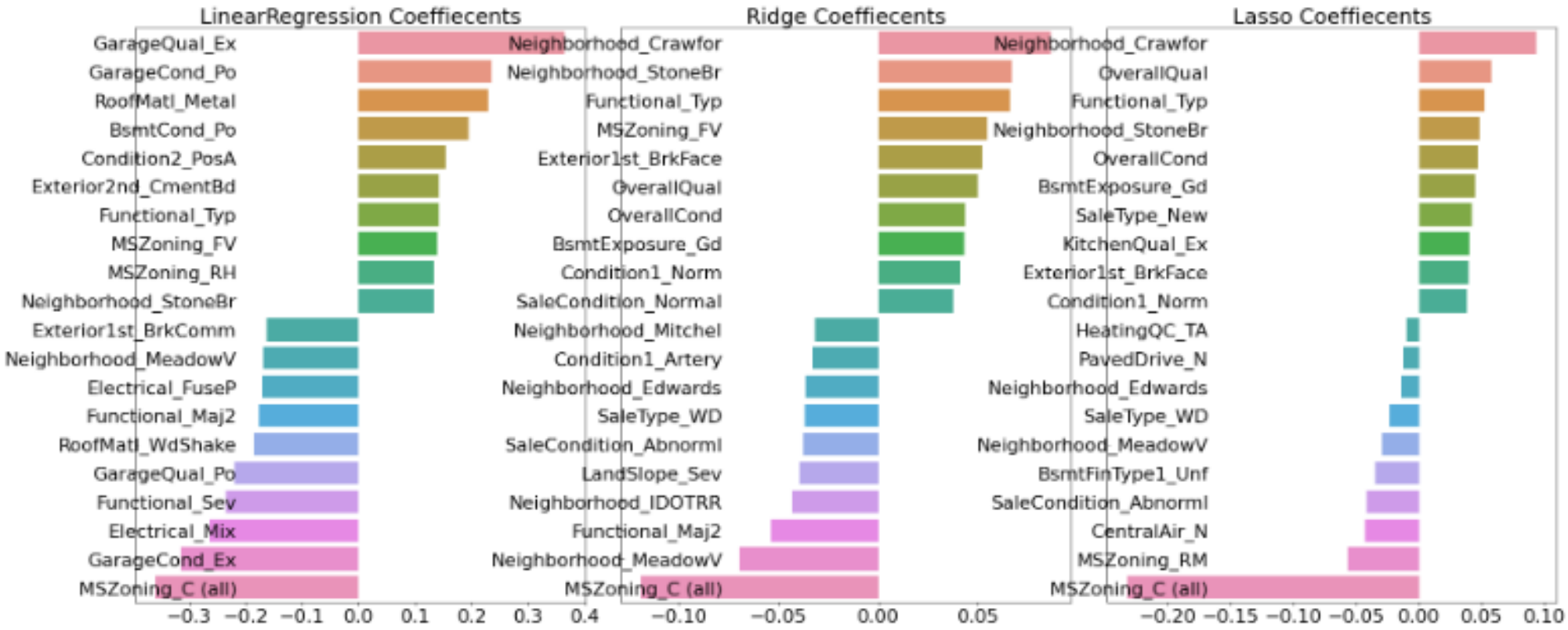
Lasso 모델의 RMSE 값이 낮아지고, R squared 값이 80%대에서 90%대로 높아지며 좋은 성능을 보임을 알 수 있음.

# LinearRegression, Ridge, Lasso 모델 변수의 설명력



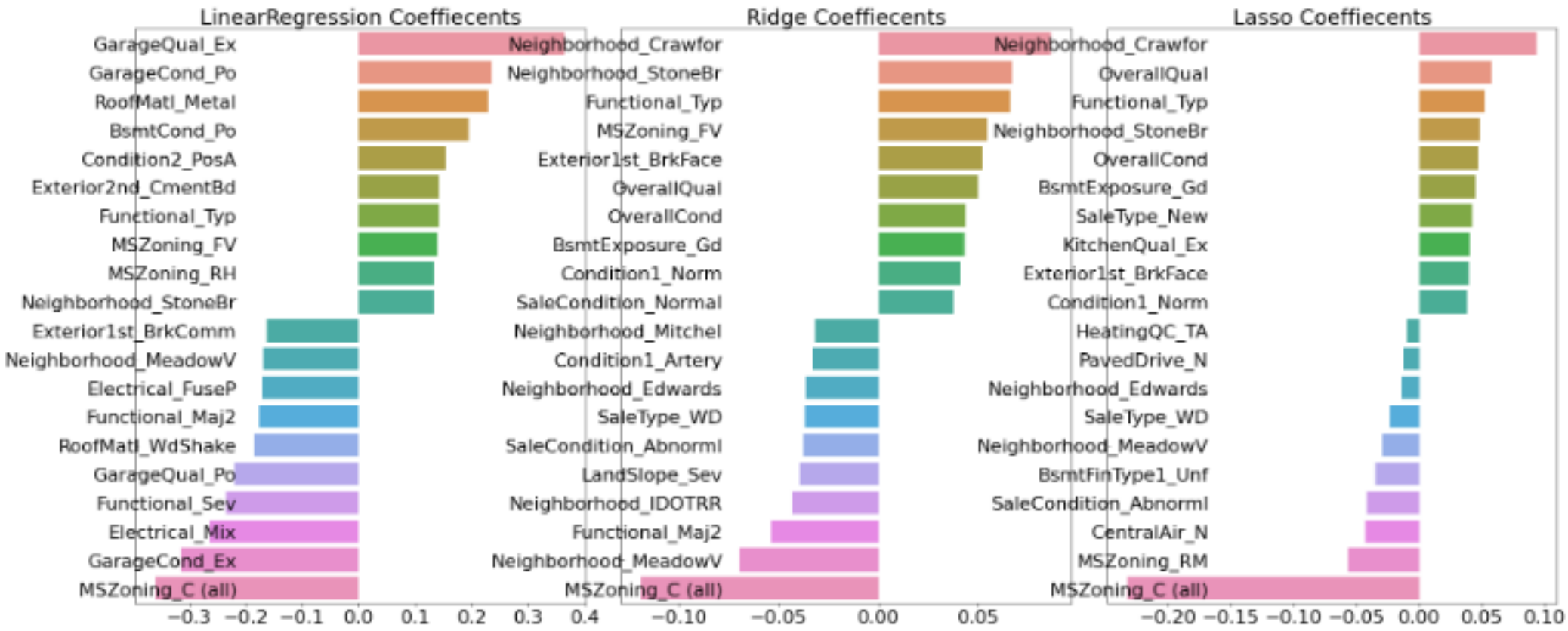

**LinearRegression** : 높은 설명력을 갖는 변수는 GarageQual_EX, GarageCond_Po, RoofMatl, BsmtCond 등의 변수로써,
'차고 품질 및 상태', '지붕 유형', ''지하실 상태' 등의 변수 값이 한 단위 증가할수록, 목적 변수인 'SalePrice'인 집 값이 높아짐을 알 수 있음.
이에 반해 ' 주택 외장재', ' 에임스 시 경계 내의 물리적인 위치', '전기 시스템' 등에 따른 변수는 한 단위 낮아질수록 집 값이 증가한다는 것을 알 수 있음.

# LinearRegression, Ridge, Lasso 모델 변수의 설명력




**Ridge** : 높은 설명력을 갖는 변수는 Neighborhood, Functional, MSZoning 등의 변수로써,
'에임스 시 경계 내의 물리적 위치', '집 기능', '매매의 일반적인 지역 분류' 의 등 양의 계수를 갖는 변수 값이 한 단위 증가할수록,
목적 변수인 'SalePrice'인 집 값이 높아짐을 알 수 있음.
이에 반해 ' 에임스 시 경계 내의 물리적인 위치' 중 Mitchel, '주요 도로 및 철도와의 접근성', '판매 유형 ' 등에 따른 변수는 한 단위 낮아질수록 집 값이 증가한다는 것을 알 수 있음.

# LinearRegression, Ridge, Lasso 모델 변수의 설명력




**Lasso** : 높은 설명력을 갖는 변수는 Neighborhood, OverallQual, Functional_Typ 등의 변수로써,

'에임스 시 경계 내의 물리적 위치', '전체적인 재료 및 마감 등급', '집 기능' 등의 양의 계수를 갖는 변수 값이 한 단위 증가할수록,

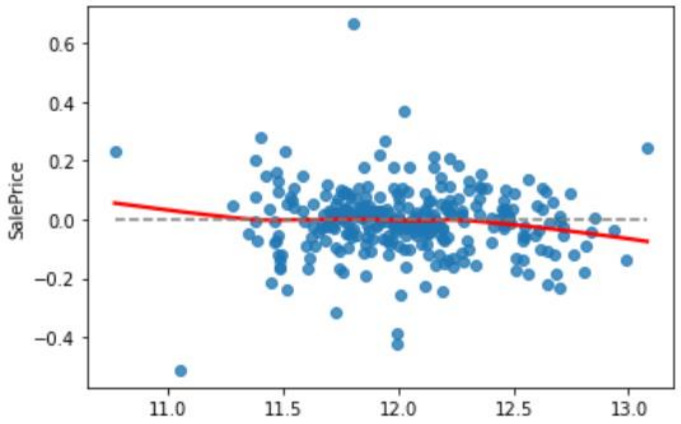
목적 변수인 'SalePrice'인 집 값이 높아짐을 알 수 있음.

이에 반해'MSZoning' 중 Commercial or RM, 'CentralAir_N', '판매 상태'등에 따른 변수는 한 단위 낮아질수록 집 값이 증가한다는 것을 알 수 있음.

# LinearRegression 진단

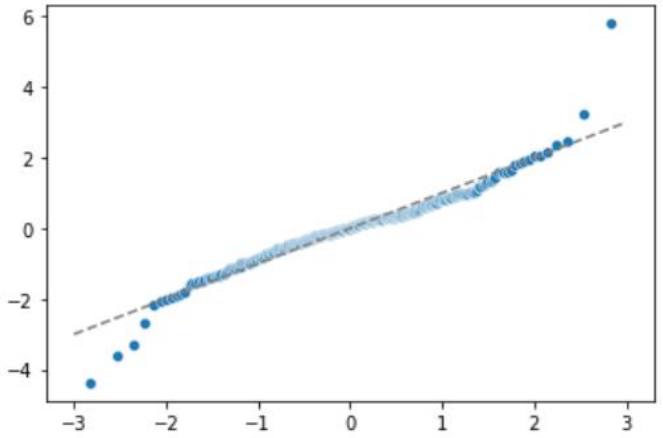
: 선형회귀분석의 기본 가정인 선형성, 독립성, 등분산성, 정규성에 대한 검토 진행

## 1. 선형성



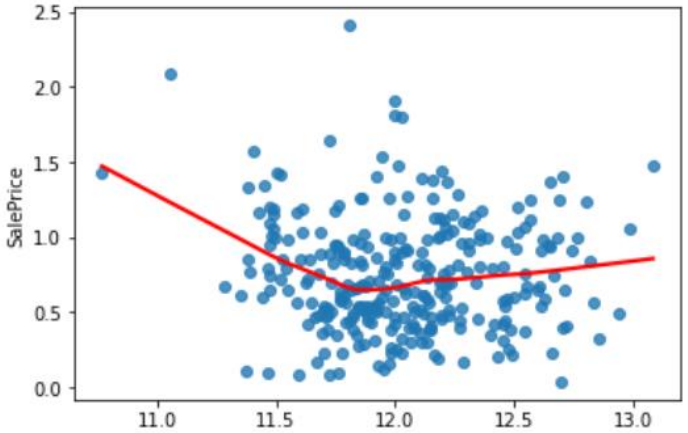

- 잔차의 산점도를 통해 시각적으로 모형이 선형의 관계성을 나타내고 있는지 알아보고자 함.
- 잔차가 해당 예측값에 따라 크게 변하지 않고, 패턴을 띠지 않으므로 선형성이 있음을 알 수 있음.

## 2. 정규성




- QQ plot에서 점들이 점선에 따라 배치되어 있음을 확인하였고, 이에 따라 정규성 가정에 만족함을 알 수 있음.
- Shapiro 검정 -> p-value=1.6146으로써, 귀무가설을 기각하지 못하므로 정규성을 따른다고 할 수 있음.

## 3. 등분산성




- 예측값에 따라 잔차의 형태가 랜덤적으로 분포되어 있으므로 등분산성 가정에 만족한다고 할 수 있음.
- 해당 fitting된 실선이 불규칙적이지 않음.

## 4. 독립성

```
Durbin-Watson: 2.112948966407264
```

- 통계적으로 검증하기 위해 Durbin-Watson 통계량을 확인하였으며, 확인 결과 2에 매우 가까운 값으로 나타남. 이에 따라 잔차의 자기상관이 있다고 보기 어려움.

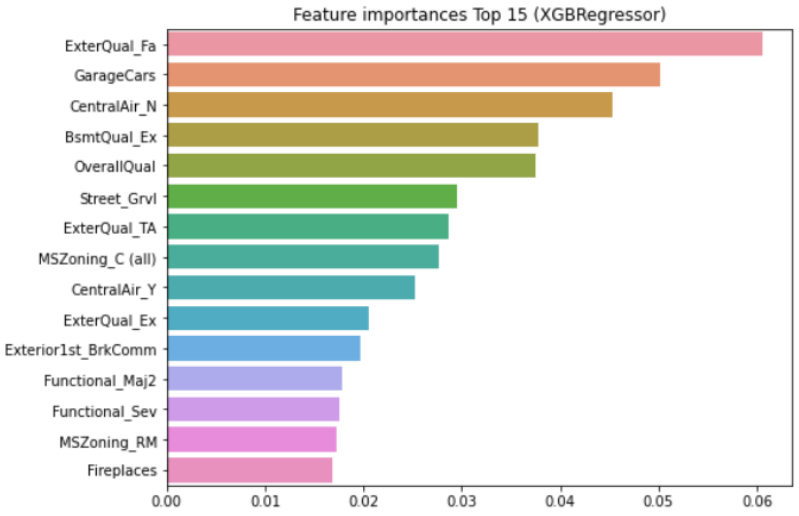
## 3. 트리 기반 회귀 모델

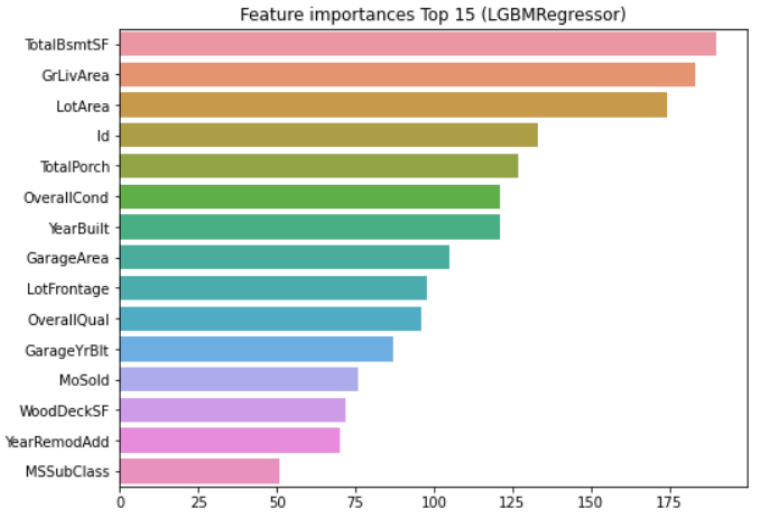
모형에 대한 RMSE 값과 R squared는 다음과 같다.

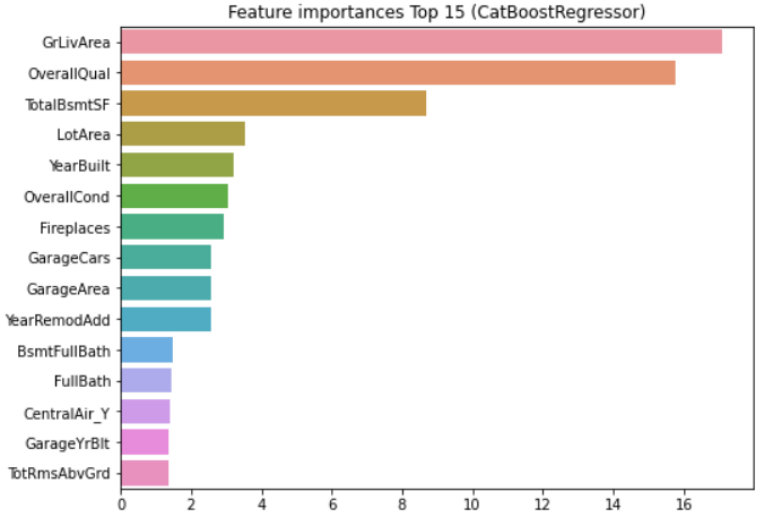
| | XGBRegressor | LGBMRegressor | CatBoostRegressor |
|---|---|---|---|
| RMSE | 0.106 | 0.108 | 0.104 |
| Train R^2 score | 0.99 | 0.96 | 0.99 |
| Test R^2 score | 0.93 | 0.92 | 0.92 |

: 3개의 모델 성능 차이가 적으나, CatBoostRegressor가 가장 좋은 성능을 보이고 있음을 알 수 있음.

---




**대표적으로 CatBoostRegressor의 변수 중요도 plot 해석** : 높은 설명력을 갖는 변수는 GrLivArea, OverallQual, TotalBsmtSF, LotArea, YearBuilt 등의 변수로써, '지상 거실 면적 평방 피트', '전체적인 재료 및 마감 등급', '지하 총 평방 피트', '공사일' 등의 변수 순으로 본 분석에서 예측하고자 하는 'SalePrice' 변수에 대한 설명력이 어느 정도 차지한다고 할 수 있음. 특히, GrLivArea, OverallQual 변수가 많은 비중을 차지함.

# 최종 예측 모델

**[CatboostRegressor]**

최종 예측 모델을 결정하는데 있어 '성능평가지수' 및 '변수 설명도'에 따라 결정함.
RMSE 값이 0.104, R squared 값이 train 0.99, test 0.92로 나타났으며, 이에 따라 CatboostRegressor를 'SalePrice'를 예측하는 최종 모델로써 결정.

보스턴 집값 예측에 있어 영향을 주는 변수 : GrLivArea, OverallQual, TotalBsmtSF, LotArea, YearBuilt, OverallCond, FirePlaces, Garage, YearBuilt, BsmtHalfBath ---

'지상 거실 면적 평방 피트', '전체적인 재료 및 마감 등급', '지하 총 평방 피트', '공사일', '집의 전체적인 상태 등급', '벽난로 수', '차고', '공사일', '지하 총 평방피트' 등의 요인이 보스턴 집값에 영향을 미치는 요인으로 판단할 수 있을 것임.
전반적으로 집의 세부적인 사항 및 등급 요인이 전체적으로 해당 예측에 대부분을 차지하는 것으로 보임.
앞서 시각화 한 TOP 15개의 몇 개의 변수는 다른 해당 모형에서도 어느정도 설명력 있는 변수로 채택되기도 함.

=> 해당 집 값 예측 모델을 활용하여 시세 예측 및 다양한 부분에서 사용할 수 있을 것이라 생각함.