가로선

House Prices : Advanced Regression Technique

**2020년 6월 17일**

**20141591 최기용**

**20141594 최승훈**

**20161584 민현홍**

# 개요

주어진 집의 정보를 통해 집의 가격을 예측하는 머신러닝 모델을 만들고, 그 정확도를 높이도록 학습시킨다. 이 때, 집의 정보가 담겨있는 Data Set은 Kaggle Knowledge Competition 항목인 House Prices에서 제공하는 Data Set을 사용한다.

# 목표

1. 주어진 Data set에서 적절한 Feature Engineering을 적용한다.
2. 적절한 Advanced Regression Techinques를 적용한다. (예: random forest)
3. 위 두 목표를 진행할 때 다양한 기법을 사용해보며 정확도를 높인다.

# 설명

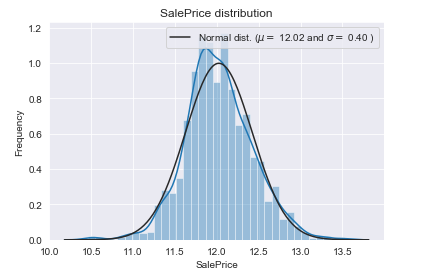
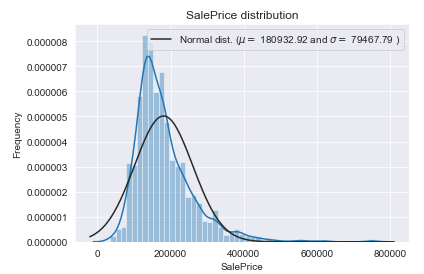
Kaggle Knowledge Competition 항목인 House Prices를 주제로 선정하였다. Data set은 해당 competition에서 제공하는 Data Set을 사용한다.

Test Data set : 1459가구 각각의 주택 관련 80개의 컬럼변수로 이루어져있다.

Train Data Set: 1460가구 각각의 주택 관련 81개(Price 추가)의 컬럼변수로 이루어져있다.

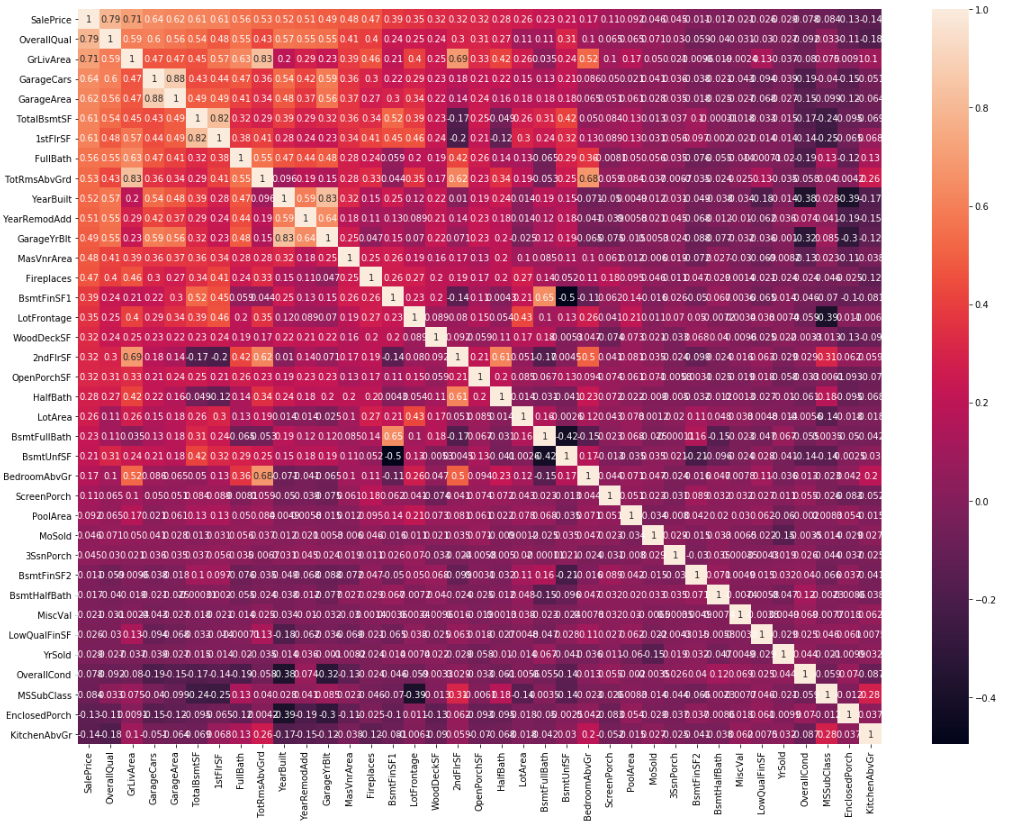
**데이터분석**

1. 타겟 변수 확인



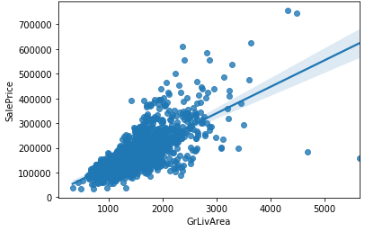
SalePrice의 값이 너무 크고, Right Skewed한 분포를 띄고 있어, 변수에 Log를 씌워 이를 정규분포에 가깝게 만들어 해결.

1. 변수 간 상관관계 확인



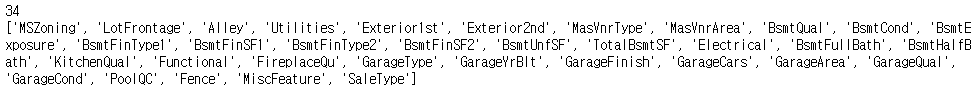
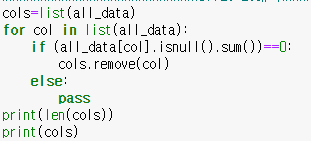
* Target Variable ‘SalePrice’와 상관관계 상위 40개 변수 Heatmap 그래프

1. Outlier 제거

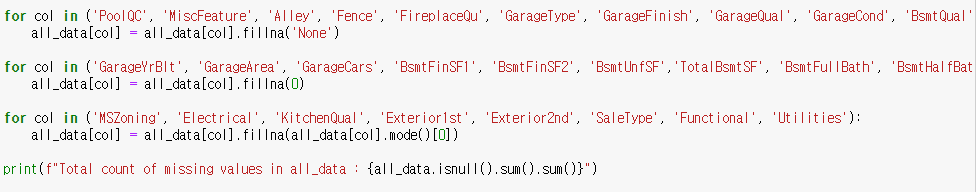


상관 관계와 전혀 다른 결과를 보여주고 있는 데이터(Outlier)들을 제거 하여 성능 향상

1. Missing Value



* 34개의 Missing Value 포함하는 Feature 확인.

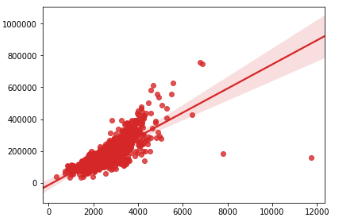




* Categorial Column의 경우 ‘None’, Numeric Comlumn의 경우 0, 해당 Column이 아예 없다고 보기 힘든 경우에는 임의의 값(최빈값 사용)으로 Missing Value를 채운다.

1. 예측에 도움이 되는 새로운 Feature 추가

* TotalBsmtSF, 1stFlrSF, 2ndFlrSF를 모두 합한 ‘TotalSF’ 변수 추가.



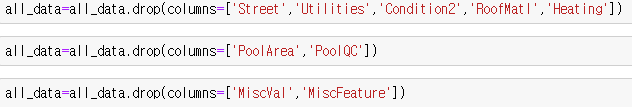
* 총 면적과 SalePrice 상관관계 표현 그래프

추가로, BsmtFullBath, FullBath, BsmtHalfBath, HalfBath를 모두 합한 TotalBath 변수 추가

1. 수치형 변수 및 비정상 변수 가공



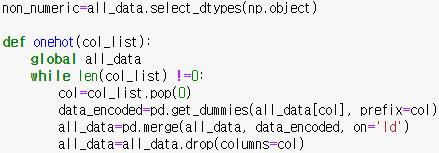
* 연도, 달 과 같은 Numeric 변수이지만 의미 상 Categorial 데이터 인 것들 변환



* 하나의 값의 비율이 너무 높은 변수들 삭제

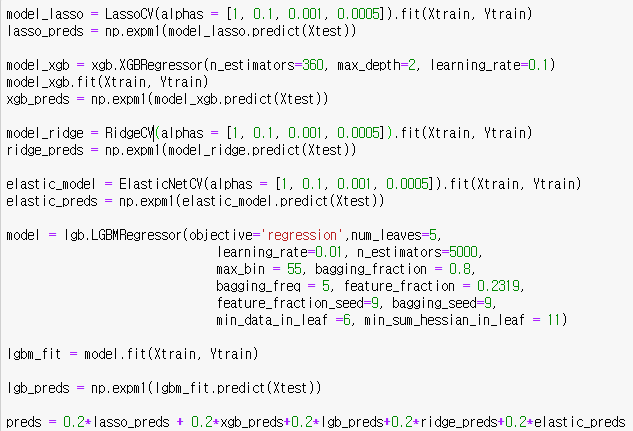
**전처리**

Categorial 변수들에 대해 One-hot Encoding



이 후, numeric 변수들에 대해 Right-skewed인 변수들 log통한 표준화

**모델 학습**



5가지 모델

Lasso, XGBoost, Ridge, Elastic, LGBM 사용하여 학습 후 각각에 같은 비율을 적용하여, 예측

**결과**



현재 Kaggle ‘House price’ Competition에 600/5386 순위에 rank되어있다.

기타 데이터 분석에서 Categorial 변수들에 SalePrice에 영향을 미치는 정도에 따라 가중치를 더 부여하는 방법, Test 데이터에 모델 적용 시, 모델 별 가중 치를 다르게 하는 방법 등을 통해 성능을 더 개선할 수 있을 것이라 예상된다.

**프로젝트 공헌도**

최기용 : 30%

최승훈 : 30%

민현홍 : 40%

가로선