Advanced Regression Techniques To Predict Housing Prices

By

Daniel Bradley, Nate Berman and Peter Shapiro

The Business Problem

Buying the right house is a difficult decision and important investment

What factors have the greatest impact on the price

We wanted to create a tool to help home buyers identify undervalued homes in the marketplace

Current State

- People rely on the following for information for guidance:
  - Real Estate Agents- Might not have your best interest in mind
  - Homes In Close Proximity- This will have some effect, but might not be the difference making factor
  - Information Online- Can sometime be misleading

Home Equity

- “The difference between how much you owe on your mortgage and the market price or value of your home”- JIM PROBASCO
- Being able to find a home that is underpriced can help to improve the rate at which you are able to build equity
- Minimum Performance Measure Needed, How to measure performance

https://www.investopedia.com/articles/mortgages-real-estate/08/home-ownership.asp

Frame The Problem

01

SUPERVISED LEARNING TASK-

WE KNOW THE OUTCOME (EACH INSTANCE COMES WITH HOUSE PRICES)

02

MULTIPLE REGRESSION-

MULTIPLE ATTRIBUTES/FEATURES- USES MULTIPLE FEATURES TO MAKE THE PREDICTION

03

UNIVARIATE REGRESSION-

WE ARE ONLY TRYING TO PREDICT ONE OUTCOME

04

BATCH LEARNING

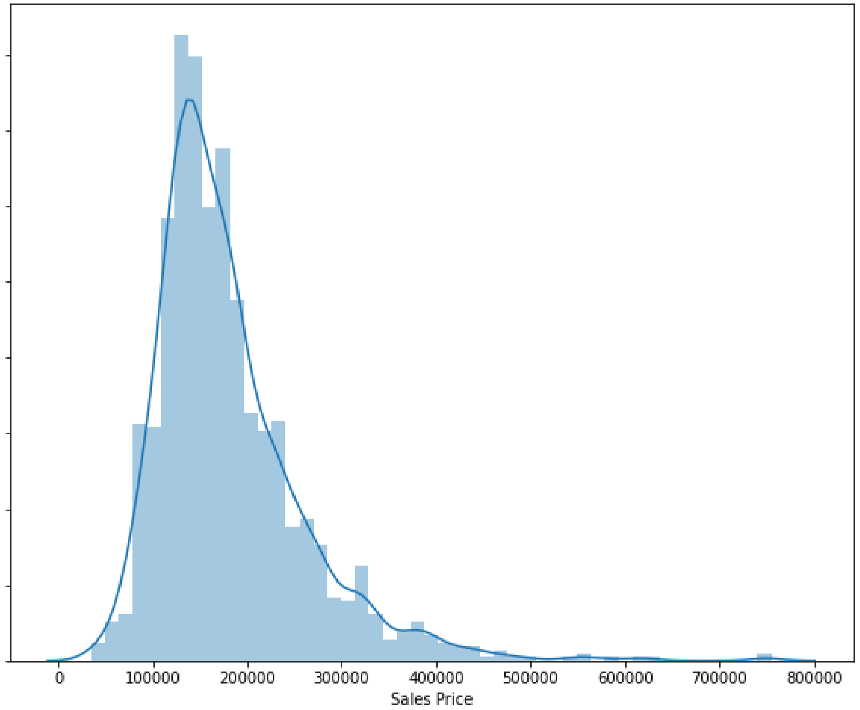
BECAUSE IT IS SMALL ENOUGH TO FIT INTO LOCAL MEMORY

The Data

- Kaggle DataSet(https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data)
- 1,460 homes in Ames, Iowa
- 79 explanatory variables for almost every home
- Categorical and numerical data features

Exploring The Data

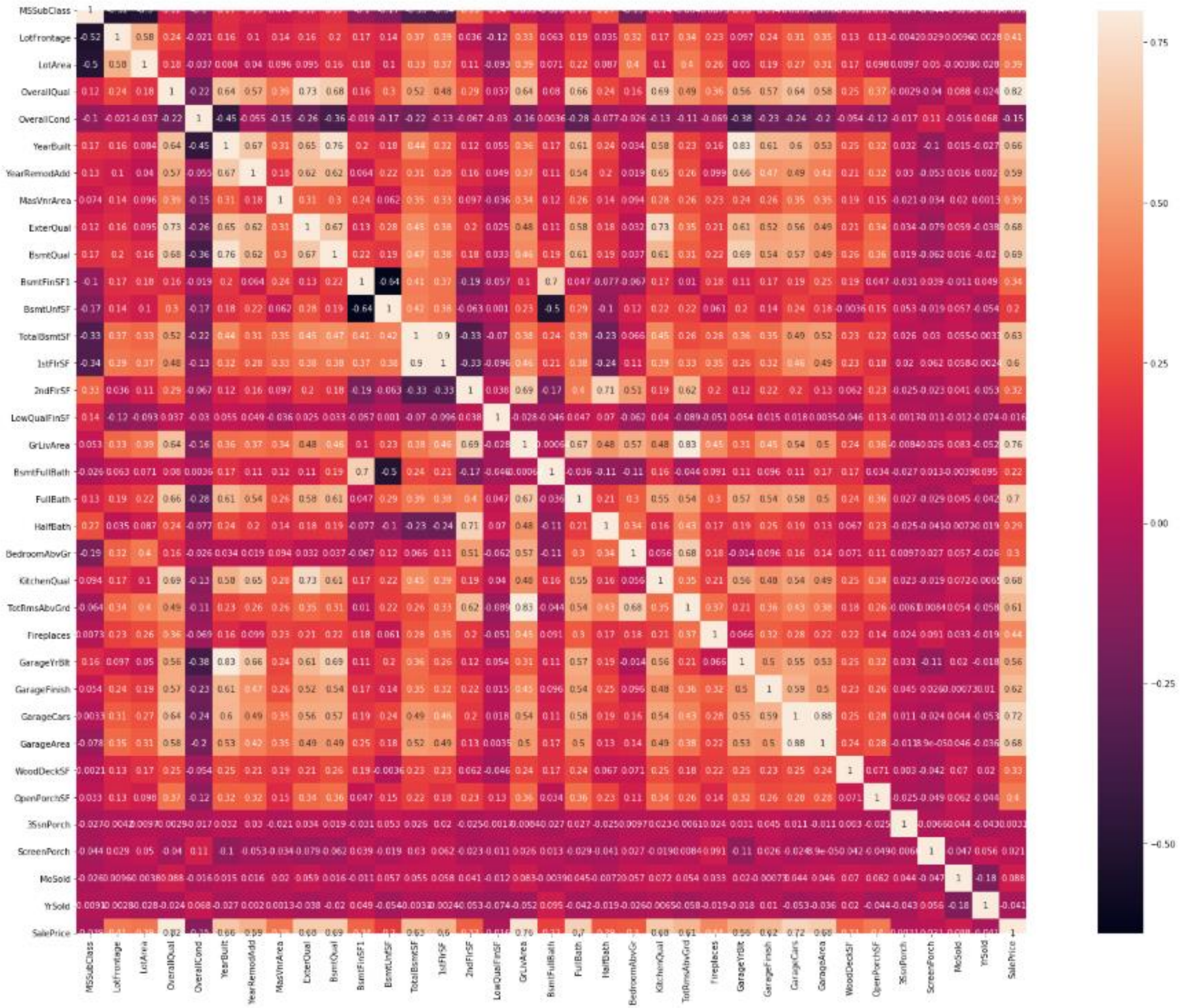
```
count      1458.000000
mean     180932.919067
std       79495.055285
min       34900.000000
25%      129925.000000
50%      163000.000000
75%      214000.000000
max      755000.000000
Name: SalePrice, dtype: float64
```



Null Values

PoolQC	1453
MiscFeature	1406
Alley	1369
Fence	1179
FireplaceQu	690
LotFrontage	259
GarageQual	81
GarageCond	81
GarageYrBlt	81
GarageFinish	81
BsmtExposure	38
BsmtFinType2	38
BsmtCond	37
BsmtFinType1	37
BsmtQual	37
MasVnrArea	8

Exploring The Data



Standard Linear Regression

- Able to get 92% accuracy out of the model
- However this required a large number of features

```
from sklearn.feature_selection import RFECV
from sklearn.svm import SVR
# X, y = make_friedman1(n_samples=50, n_features=10, random_state=0)
estimator = SVR(kernel="linear")
selector = RFECV(estimator, step=5, cv=2)
selector = selector.fit(X, y)
selector.support_
```

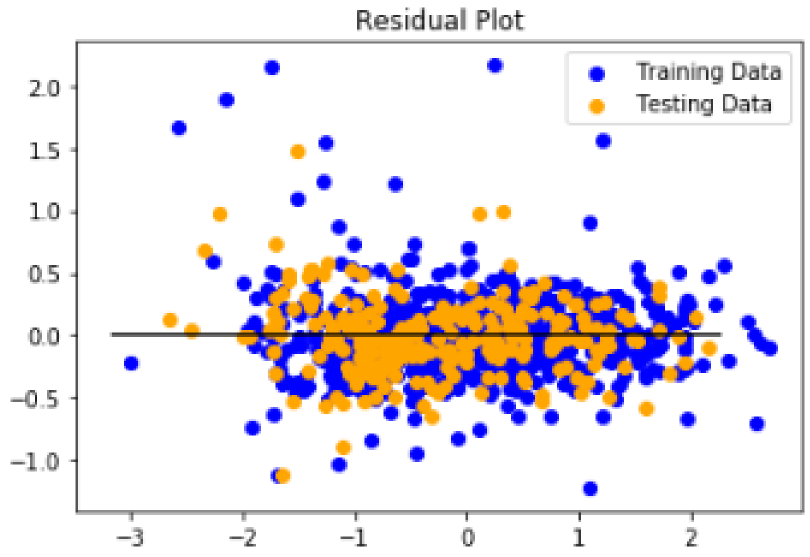
```
array([False, False, False,  True,  True, False, False, False,  True,
        True, False, False, False, False, False, False,  True,  True,
        True,  True,  True,  True,  True, False, False,  True, False,
       False, False, False, False, False,  True])
```

```
selector.ranking_
```

```
array([5, 2, 5, 1, 1, 3, 4, 3, 1, 1, 4, 4, 4, 5, 2, 5, 1, 1, 1, 1, 1, 1,
       1, 3, 2, 1, 4, 5, 3, 2, 3, 2, 1])
```

```
selector.n_features_
```

```
13
```




```
MSE: 0.09446364816511292, R2: 0.9127580961132866

0.8995727811388458
0.9127580961132866
```

Standard Linear Regression

Recursive feature elimination revealed that this model may have had too many features included in it for its prediction

```
from sklearn.feature_selection import RFECV
from sklearn.svm import SVR
# X, y = make_friedman1(n_samples=50, n_features=10, random_state=0)
estimator = SVR(kernel="linear")
selector = RFECV(estimator, step=5, cv=2)
selector = selector.fit(X, y)
selector.support_
```

```
array([False, False, False,  True,  True, False, False, False,  True,
        True, False, False, False, False, False, False,  True,  True,
        True,  True,  True,  True,  True, False, False,  True, False,
       False, False, False, False, False,  True])
```

```
selector.ranking_
```

```
array([5, 2, 5, 1, 1, 3, 4, 3, 1, 1, 4, 4, 4, 5, 2, 5, 1, 1, 1, 1, 1, 1,
       1, 3, 2, 1, 4, 5, 3, 2, 3, 2, 1])
```

```
selector.n_features_
```

```
13
```

Nate

Daniel