House Price Prediction
Linear vs Gradient Boosting

AIDI 1002 – Section 01

Behnam Khazaei – 100782437

Alex Zhou - 100807843

Bayron Jose Guevara Calderon - 100813642

Sherif ELHennawy – 100799217

Problem Definition

- House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house.

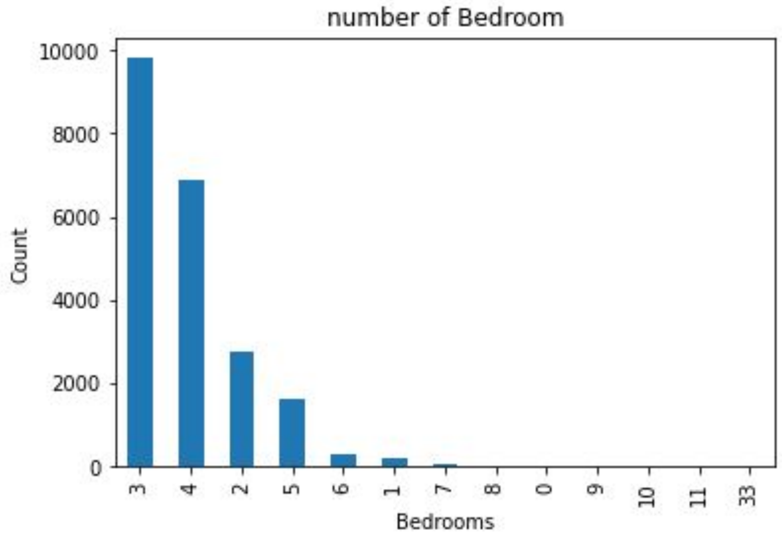
- Traditional house price prediction is based on cost and sale price comparison lacking of an accepted standard and other affecting parameters. Therefore, the availability of a house price prediction model helps fill up an important information gap and improve the efficiency of the real estate market.

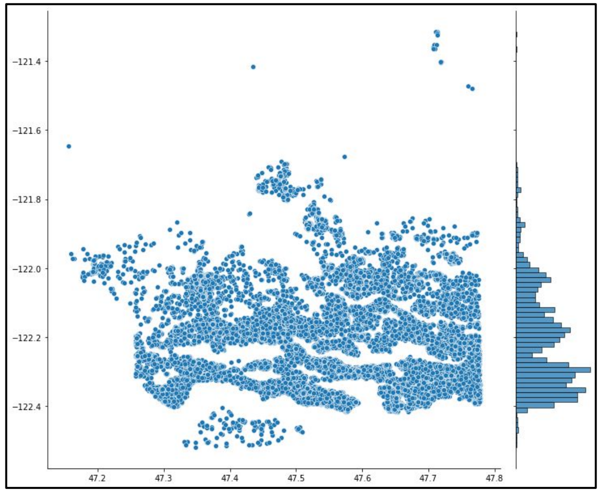
Data Input

- Test and validation dataset
  - Requirement: A dataset contains as many rows of tweets as possible
  - Purpose: This works as a test and validation for the model, so that it can be improved and tweaked to reach expectations

Dataset

- Source: Kaggle
- Shape: 21,613 rows
- Contains tweet information
  - ID number, date, housing price, the number of bedrooms and bathrooms, space for living and lot, floors, space above the ground and basement, years built and renovated, zip code, waterfront and view, latitude and longitude

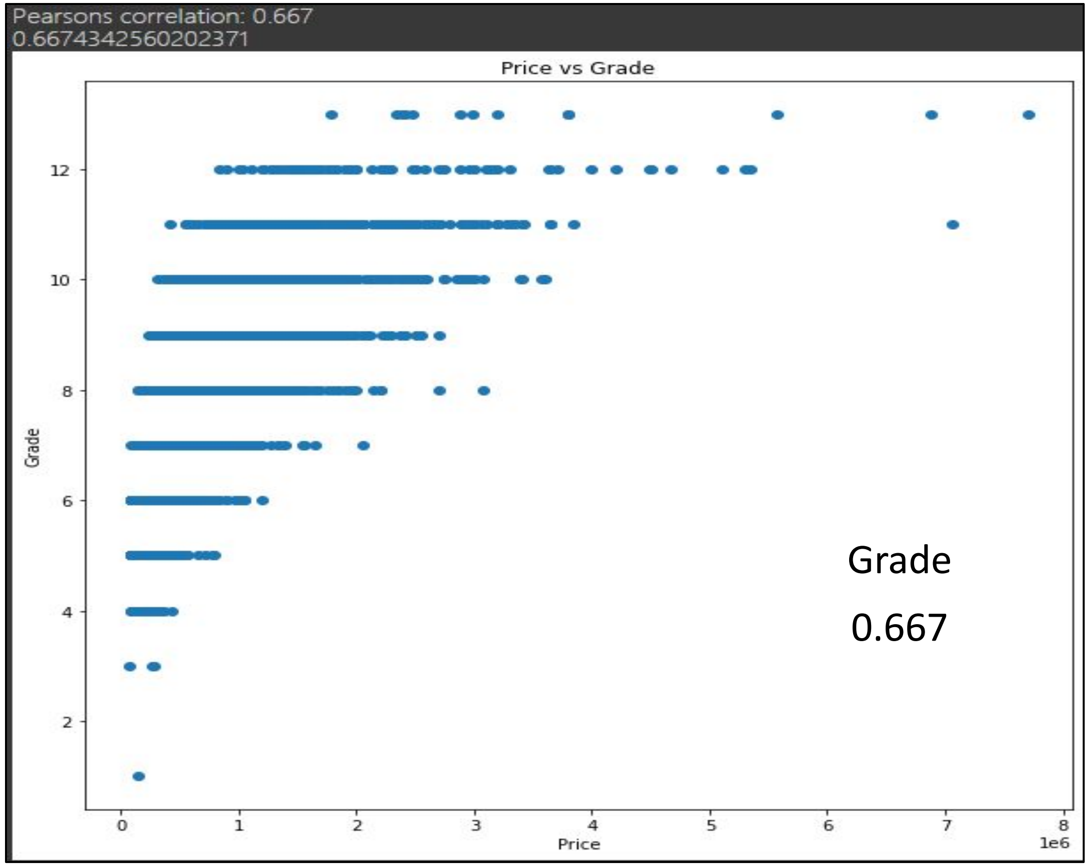


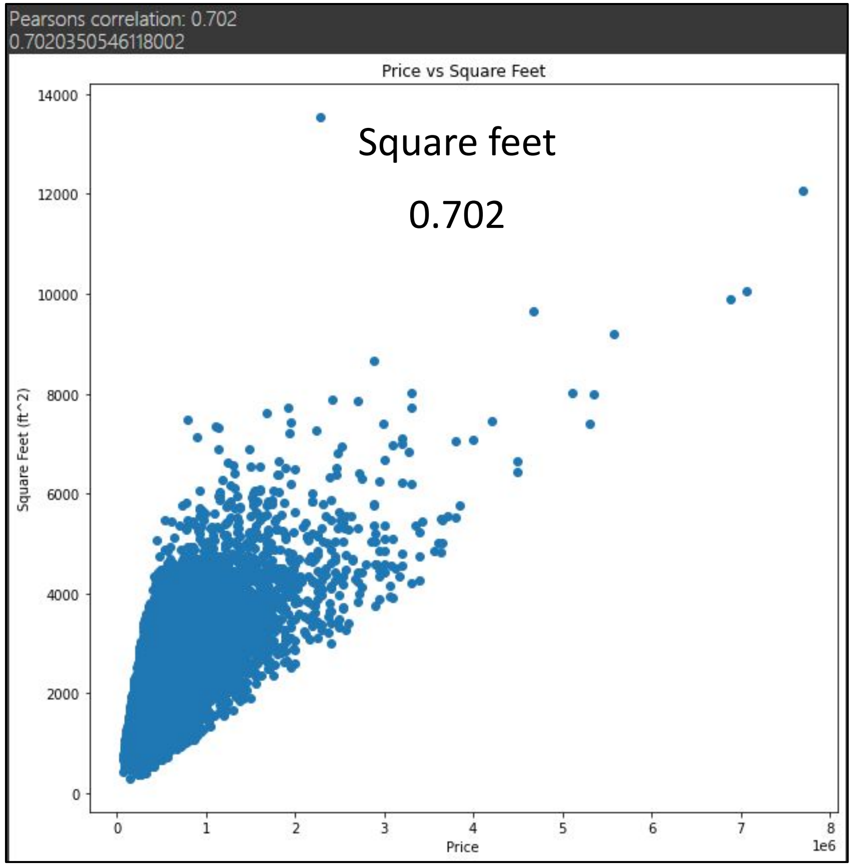



```
1   data.head()
```

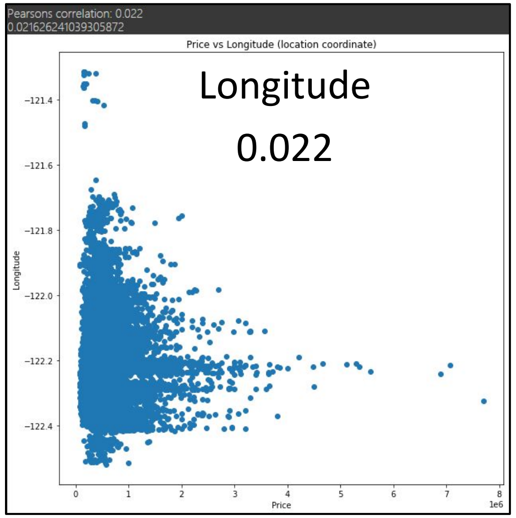
	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	3	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	3	7	2170	400	1951	1991	98125	47.7210	-122.319	1690	7639
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	3	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	5	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	3	8	1680	0	1987	0	98074	47.6168	-122.045	1800	7503

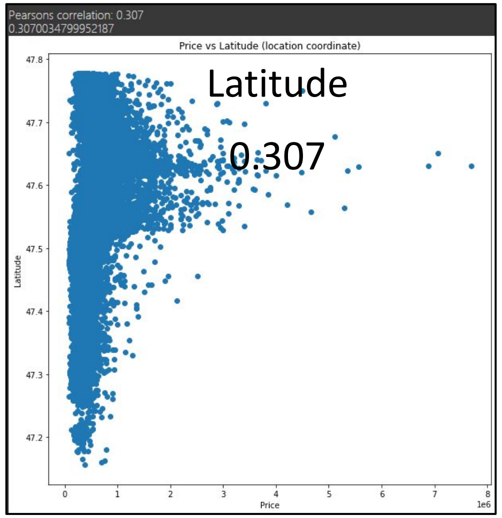
Factors have high correlation with price

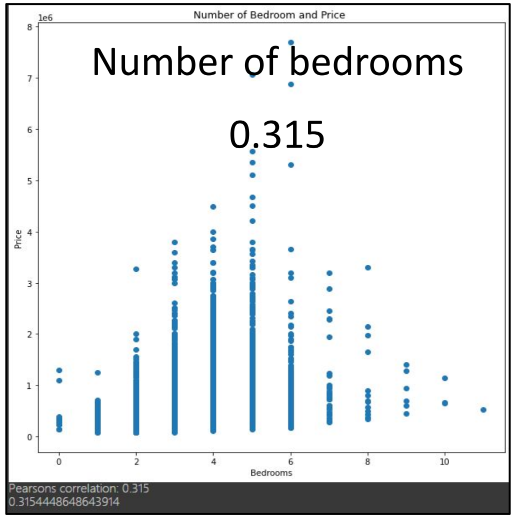


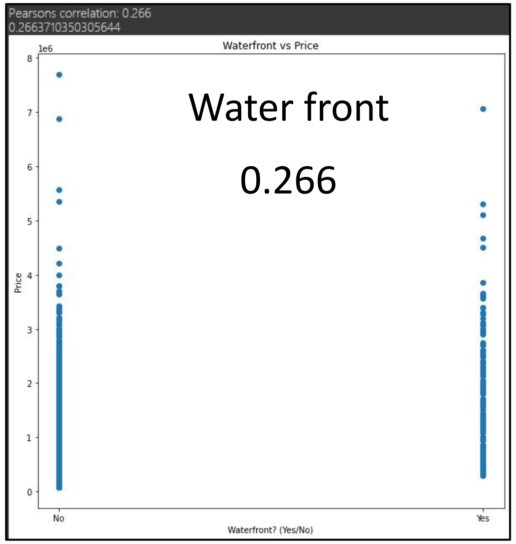


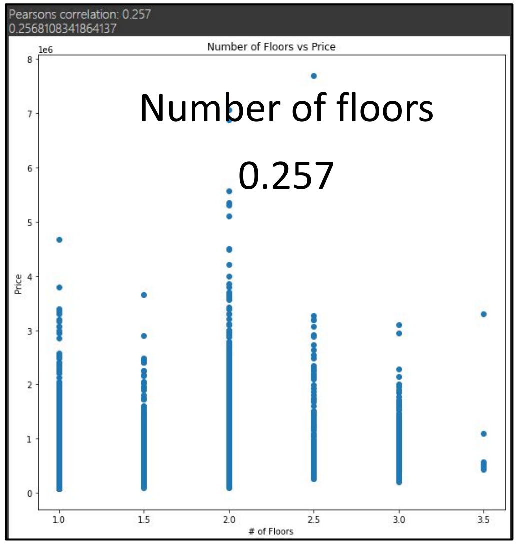
Factors have low correlation with price

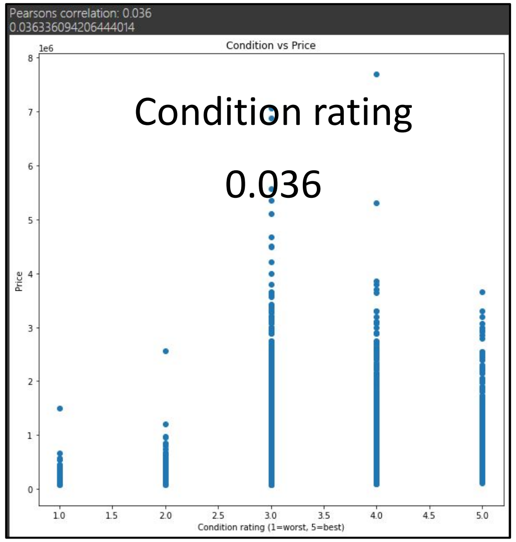


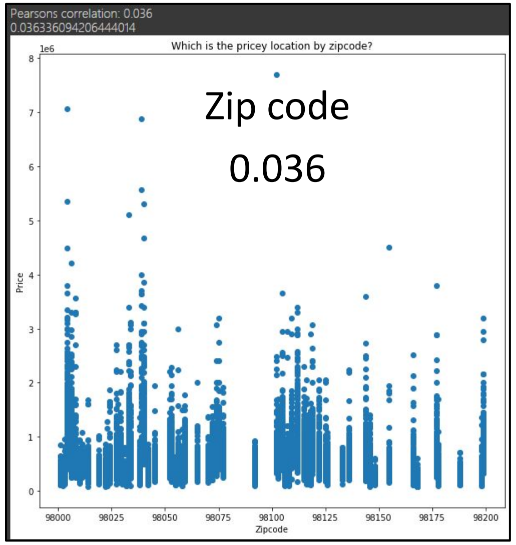








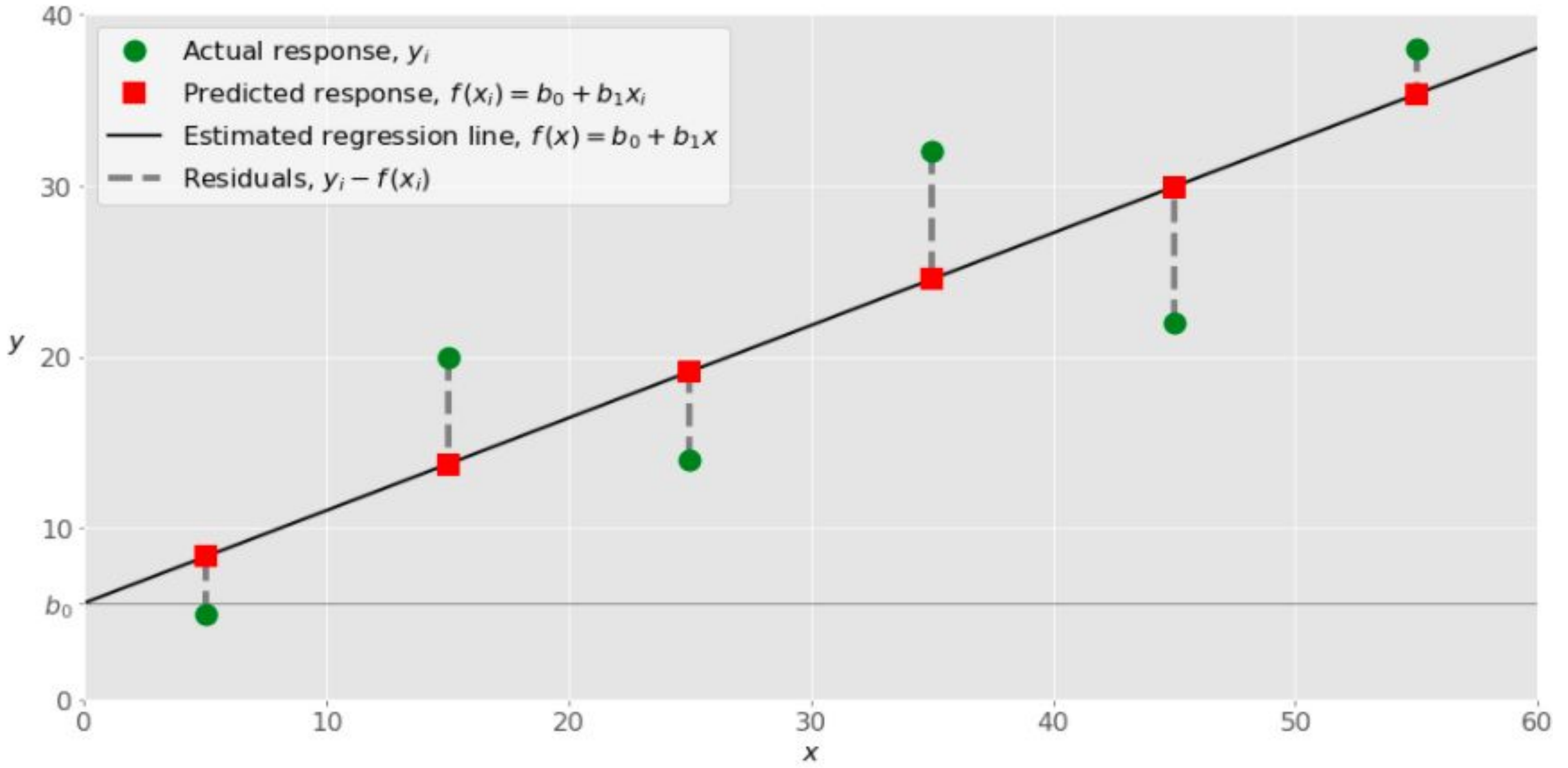




Linear Regression

The liner regression model tries to find a set of parameters for a liner equation that will describe the relation between the two variables

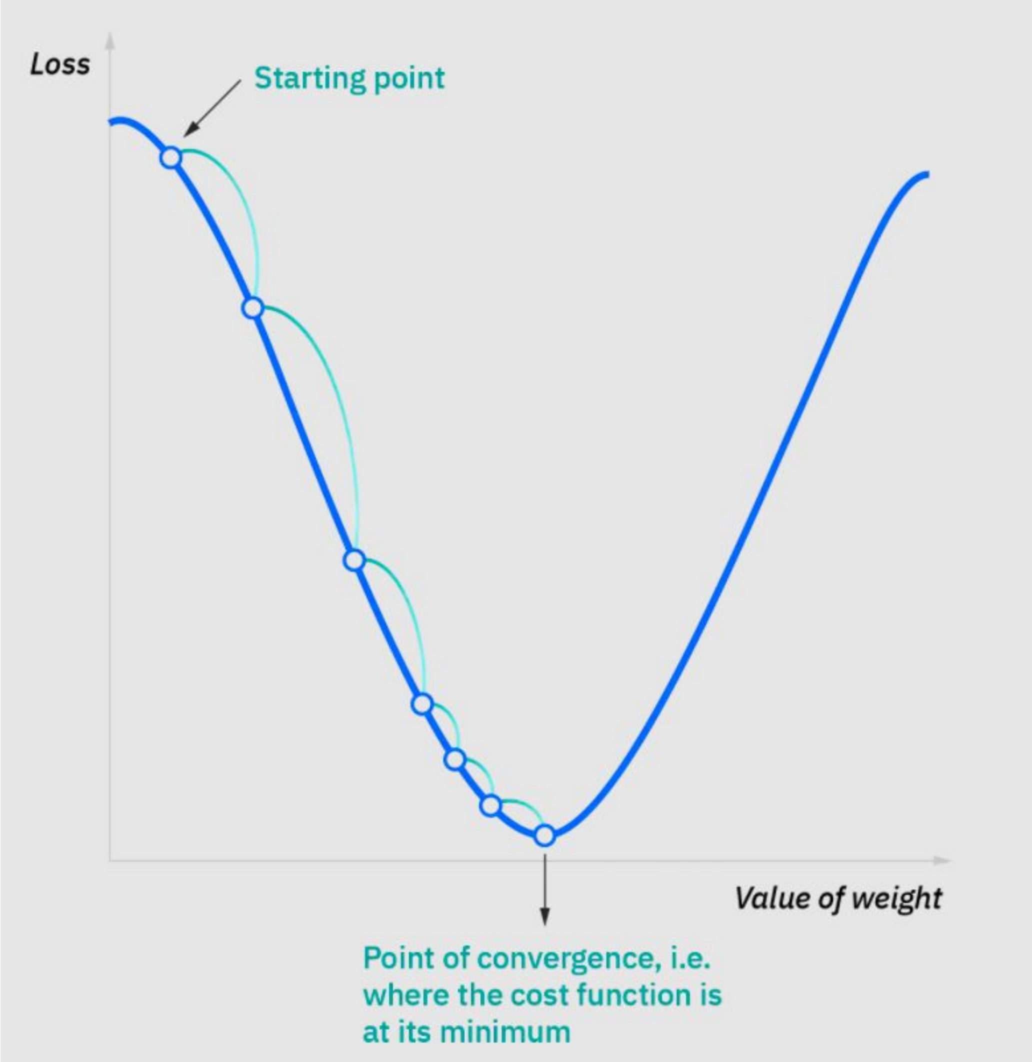
$$Y = a + bX$$



Example of simple linear regression

Gradient Boosting Regression

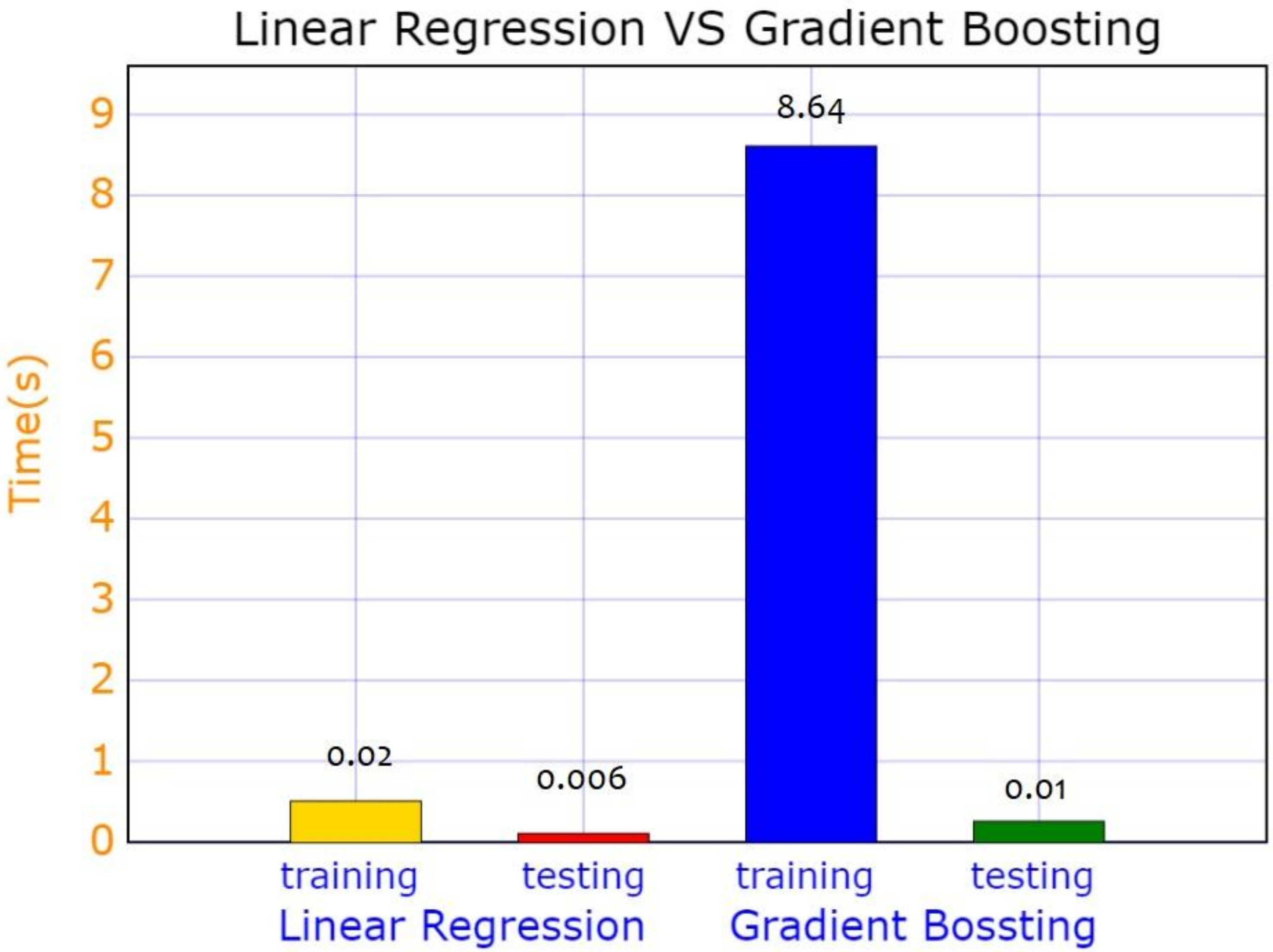
Gradient descent is an optimization algorithm used to minimize cost function by iteratively moving in the direction of **steepest descent** as defined by the negative of the **gradient**



Performance

- **Time**

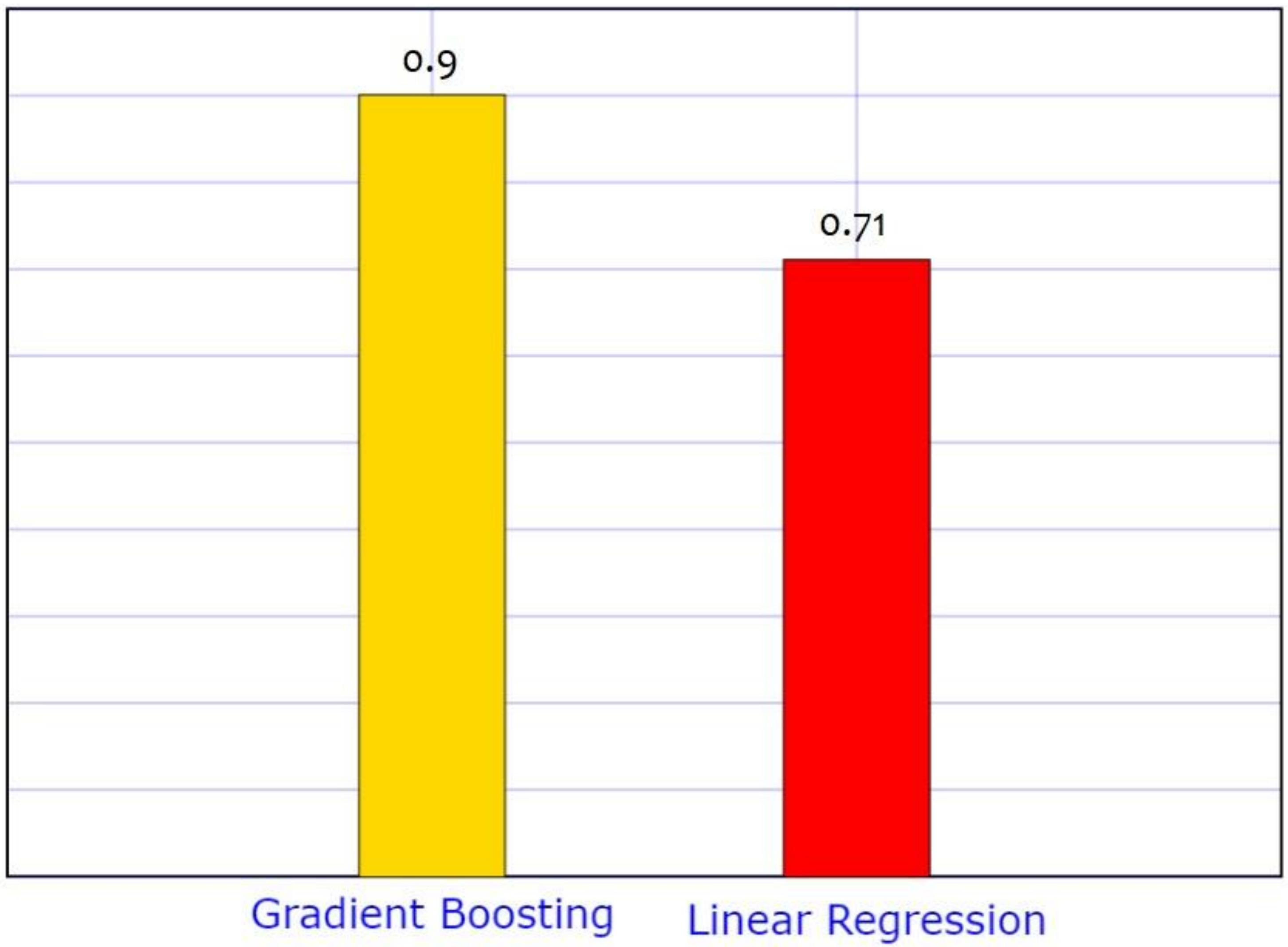
- **Coefficient of Correlation:**



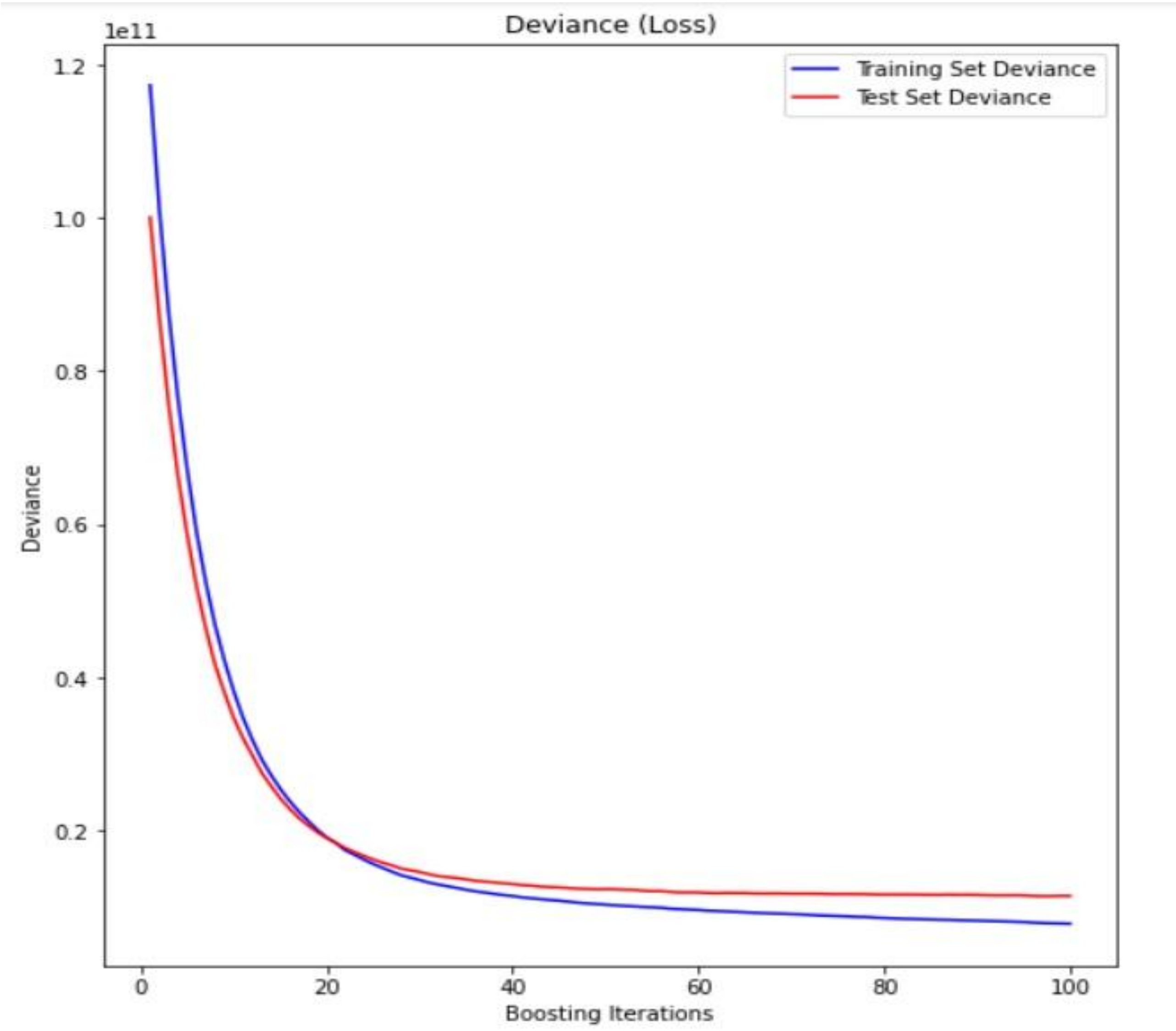
Specific Algorithm: Gradient Boosting

- **Number of boosting stages**
- **Learning Rate**
- **Maximum depth of the individual regression estimators**
- **Minimum sample split Deviance graph**

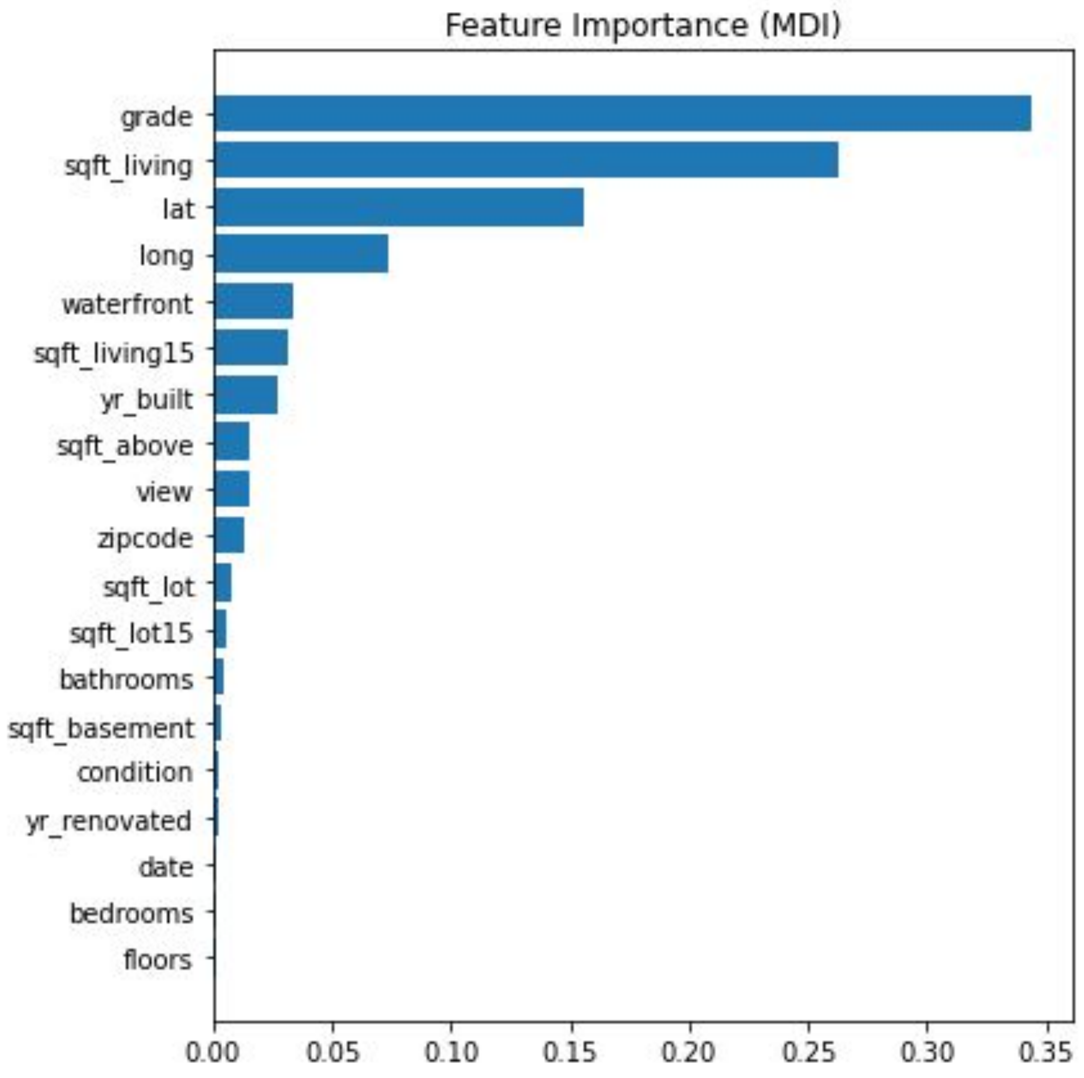
Coefficient of Correlation



Deviance Graph



Relative Importance of Features



Predictions

Portion of test dataset: **15 samples**

Multi Linear Regressor

	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15	Actual Price	Predicted Price	Diff (%)
1	1917	0	98107	47.6608	-122.359	1620	4400	$710,500.00	$768,740.56	8.20%
2	1985	0	98006	47.5617	-122.158	3760	9450	$1,505,000.00	$1,296,427.54	-13.86%
3	1921	0	98146	47.5031	-122.348	1170	7676	$425,000.00	$504,423.12	18.69%
4	1972	0	98022	47.1808	-122.023	1700	181708	$350,000.00	$235,971.24	-32.58%
5	2014	0	98003	47.3413	-122.180	2156	3920	$333,490.00	$377,051.78	13.06%
6	1990	0	98033	47.6533	-122.183	3310	11651	$980,000.00	$1,108,673.75	13.13%
7	2005	0	98019	47.7456	-121.984	1970	2952	$299,950.00	$408,357.28	36.14%
8	1974	0	98034	47.7174	-122.236	1650	9794	$446,000.00	$438,830.93	-1.61%
9	1998	0	98038	47.3832	-122.057	2880	26023	$448,000.00	$746,749.22	66.69%
10	2014	0	98006	47.5380	-122.114	5790	13928	$1,750,000.00	$1,500,457.78	-14.26%
11	1962	0	98118	47.5362	-122.290	1160	8906	$262,500.00	$259,054.07	-1.31%
12	2000	0	98075	47.5965	-122.038	2590	6530	$672,500.00	$675,685.93	0.47%
13	2008	0	98199	47.6374	-122.388	2010	3175	$465,000.00	$342,853.18	-26.27%
14	1940	2015	98133	47.7412	-122.355	1760	10505	$285,000.00	$939,614.12	229.69%
15	2014	0	98034	47.7323	-122.165	3080	11067	$960,000.00	$1,158,156.19	20.64%

Gradient Boosting Regressor

	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15	Actual Price	Predicted Price	Diff (%)
1	1917	0	98107	47.6608	-122.359	1620	4400	$710,500.00	$756,304.62	6.45%
2	1985	0	98006	47.5617	-122.158	3760	9450	$1,505,000.00	$1,393,142.55	-7.43%
3	1921	0	98146	47.5031	-122.348	1170	7676	$425,000.00	$312,340.07	-26.51%
4	1972	0	98022	47.1808	-122.023	1700	181708	$350,000.00	$400,244.77	14.36%
5	2014	0	98003	47.3413	-122.180	2156	3920	$333,490.00	$350,903.83	5.22%
6	1990	0	98033	47.6533	-122.183	3310	11651	$980,000.00	$947,999.79	-3.27%
7	2005	0	98019	47.7456	-121.984	1970	2952	$299,950.00	$339,973.99	13.34%
8	1974	0	98034	47.7174	-122.236	1650	9794	$446,000.00	$420,514.97	-5.71%
9	1998	0	98038	47.3832	-122.057	2880	26023	$448,000.00	$544,975.81	21.65%
10	2014	0	98006	47.5380	-122.114	5790	13928	$1,750,000.00	$1,821,261.41	4.07%
11	1962	0	98118	47.5362	-122.290	1160	8906	$262,500.00	$302,993.11	15.43%
12	2000	0	98075	47.5965	-122.038	2590	6530	$672,500.00	$701,060.06	4.25%
13	2008	0	98199	47.6374	-122.388	2010	3175	$465,000.00	$502,034.57	7.96%
14	1940	2015	98133	47.7412	-122.355	1760	10505	$285,000.00	$580,060.30	103.53%
15	2014	0	98034	47.7323	-122.165	3080	11067	$960,000.00	$894,364.06	-6.84%

	Gradient Boosting Regressor	Linear Regressor	Difference
Absolute error	$56,919.15	$155,020.66	$98,101 (decrease)
R^2	0.97	0.75	0.22 (increase)