Project 2 – Regression Analysis

Author – Udhai Pratap Singh

Project Overview

- **Project Overview**

- This project analyzes the house prices in King County, WA, USA. The analysis ails to guide a real estate agency is advising their clients on the right price for house sale or purchase, as well as impact of renovation on house prices.

Business Problem

Prime House Agency is a renowned real estate agency active in the King County and they want help in giving prospective clients advice on houses prices in the area.

They are looking for insights whether houses are overpriced/underpriced and what is the impact of factors like waterfront, view and renovations on a house price.

Features used

❖ We have used most of the features available from the king county sales dataset for the houses sold in 2014-15.

❖ We have created quite a few new features like renovation and age of the house to account for various factors influencing prices.

❖ We have identified the top features which can predict house prices

Power of prediction

Just demonstrating that our model has a good fit

Actual vs Predicted Sale Price(Validation Set)

Actual house prices($)

Predicated house prices($)
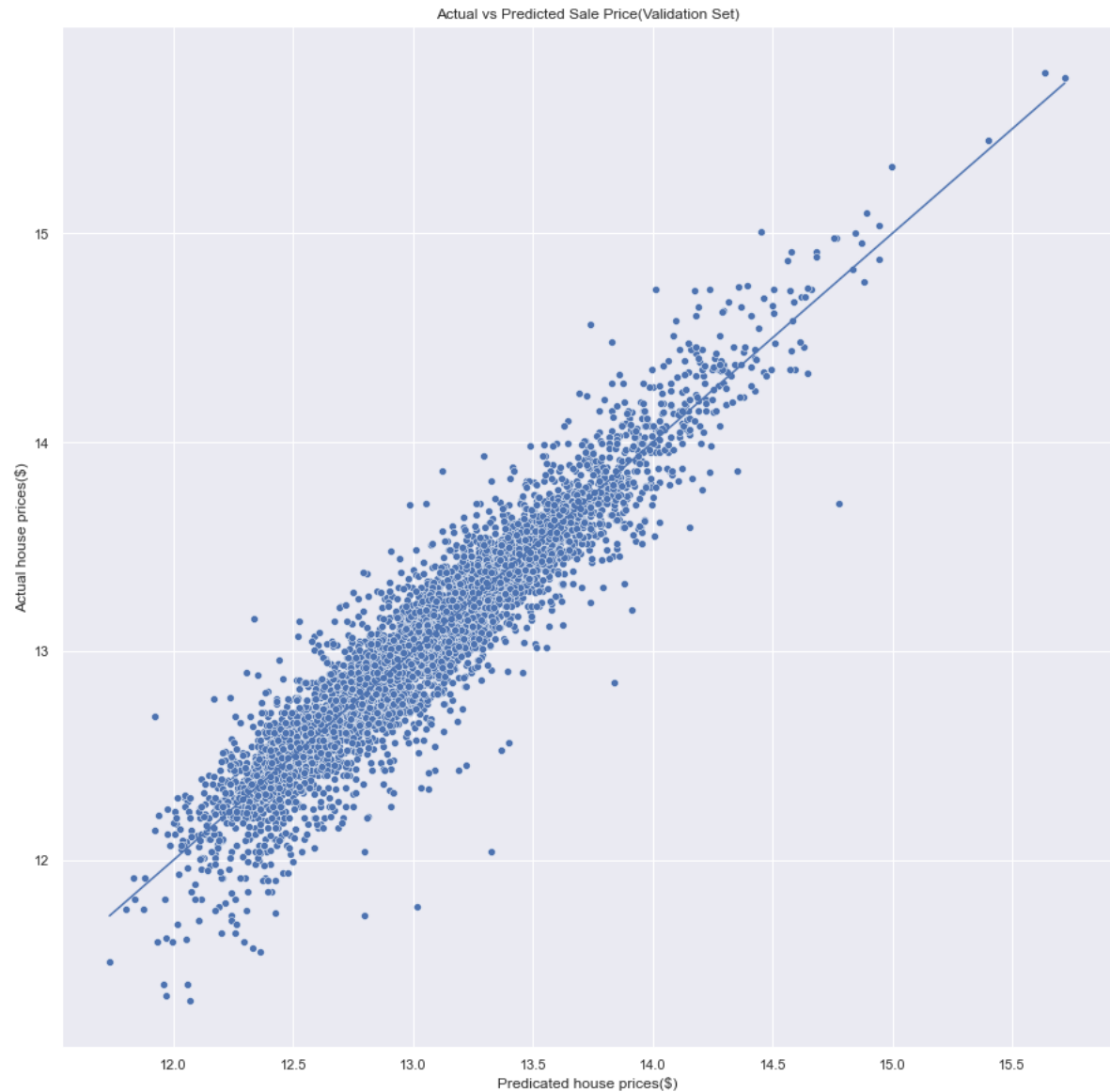
Various iterations

- We arrived at our final model via an iterative process

- Our final model is represented at row 10 in the image on the right

- Test_r2 simply represents the fit of our model and 0.89 value is considered pretty good

	fit_time	score_time	test_r2	train_r2	test_neg_mean_squared_error	train_neg_mean_squared_error	dataset	n_features
0	0.007580	0.006981	0.655989	0.657006	-0.095132	-0.095078	baseline	3
1	0.010173	0.006383	0.658585	0.663126	-0.094323	-0.093378	cont features pre scaling	6
2	0.008578	0.005386	0.633704	0.640273	-0.101215	-0.099715	cont(scaled) features	7
3	0.051861	0.009375	0.610047	0.619745	-0.107476	-0.105390	ohe cat features	75
4	0.007380	0.006183	0.199790	0.204785	-0.221141	-0.220429	discrete features	5
5	0.072406	0.009973	0.864410	0.866867	-0.037483	-0.036904	cont(scaled) + discrete + cat features	87
6	0.008776	0.006582	0.171338	0.177186	-0.228851	-0.228071	5 features via RFE	5
7	0.010173	0.006583	0.449899	0.453306	-0.152064	-0.151541	13 polynomial features	8
8	0.071608	0.009176	0.872836	0.875177	-0.035160	-0.034601	preprocessed + polynomial	95
9	0.009974	0.007379	0.619850	0.627552	-0.104954	-0.103237	7 dropped features	7
10	0.087765	0.013763	0.891607	0.893996	-0.029960	-0.029384	preprocessed + polynomials +7 dropped feat	102
11	0.010772	0.005784	0.396071	0.401331	-0.166612	-0.165933	top10(of 101) feats via RFE	10

index	top_10_features	coeffecients
1	Renovation_1.0	5.783980e+11
2	waterfront_1.0	5.212103e+10
3	zipcode_98039	9.465826e-01
4	year_sold	1.579698e+09
5	Renovation_1.0-s2_norm	-9.674879e+09
6	Renovation_1.0-Sq_norm	-9.576285e+10
7	waterfront_1.0-s2_norm	-8.910966e+11
8	waterfront_1.0-Sq_norm	8.865311e+11
9	sqft_living15	9.574054e-01
10	date	-1.579698e+09

Top Features

◈ The coefficient columns simply helps us interpret the extent of influence of these factors on house prices

◈ Warning: correlations are not necessarily causation

Top features(cont.)

◈ We can also see a lot of these features relates to renovation and waterfront. Hence, we can say that renovation and waterfront have strong predictive relationship with house prices. Since coefficient for both renovation_1.0 and waterfront_1.0 are positive hence having a renovation will likely increase house price.

◈ Caution: predictive accuracy of this model is not 100%. So keep this in mind, while advising or planning!!

Key takeaways

◈ If a house has waterfront then it likely to fetch a higher price than a house without it, in the same zip code

◈ If a house has renovation – then it is likely to fetch a higher price than a house without it

◈ If the prospective house is below the line of fit shown before, then your house has probably scope of price appreciation and one measure which can be taken is renovation

◈ We can fill in the required info on a house (not in this dataset) and see where it lies in our prediction graph and then gauge the extent of under/over pricing.