Machine Learning mini projest

Katayoon Sadeghi mehr - March 2020

Project: Housing Price

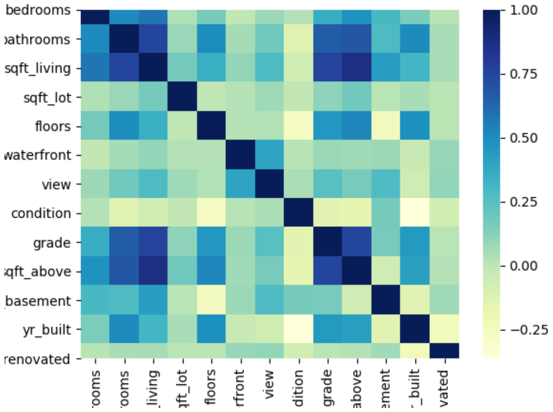
Abstract:

There are some features affected the Housing price. It can be the size of the house such as number of rooms and bathrooms, location and so on. The purpose of this study is to find the best model, using Linear regression, KNN, Adaboost, Random Forest and SVR, to find out what are the features affecting the housing price.

Data Explanation:

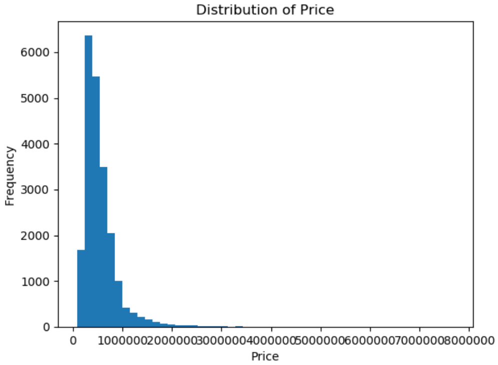
The data contains 21597 rows. The explanatory variables are #Bedrooms, #Bathrooms, sqft_living, sqft_lot, sqft_above, sqft_basement, floors, waterfront, view, condition, grade, yr_built and yr_renovated. The dependent variable is price which is a continuous variable.

Findings:



There is correlation between #of bedrooms, #of bathrooms and size of the house for living and upper floor (in sqft) and grade.

The mean and std of price are 540296.57 and 367368.14, while the median is: 450000. This shows that the data is very right skewed. Most of the houses are of cheaper price while there are some with very high prices. As most of the houses don't have renovation year; this X is removed.

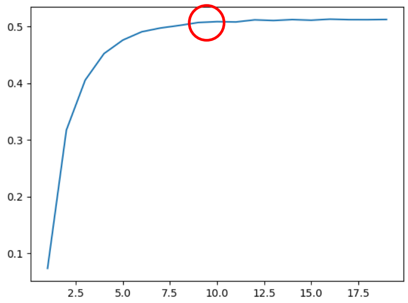


Linear Regression: The R-Square (0.209) and Adj-R-Square (0.208) are pretty low.
As most of the houses don't have renovation year and the p-value is greater than 0.05 this X is removed. Also, 'floors' is not significant. The rest of the X variables have p-values close to zero.

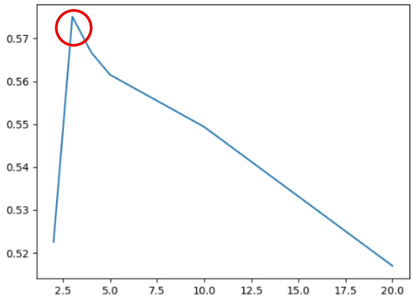
Price = 542725.23 - -36985.27 #Bedrooms + 40252.95 #Bathrooms + 82908.11 sqft_living - 14343.32 sqft_lot + 73207.05 sqft_above + 35247.39 sqft_basement + 48147.13 waterfront + 37298.16 view + 10912.13 condition + 147602.54 grade - 104508.32 yr_built

Clearly the size living area, basement and above floor with large slopes have positive impact on increasing the price. Also, if the house is waterfront with good view has significantly higher price. But year of built with negative slope shows that older homes have more value.
Mean square Error = 39827750819.19

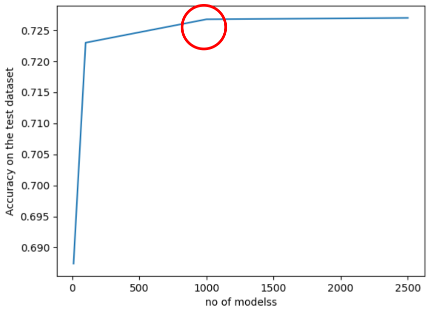
KN Neighbors: Applying KNN Regressor and looking at the accuracy for different Ks, K=10 gives the highest accuracy close to 51%.



Adaboost: Applying Adaboost Regressor and looking at the accuracy for different number of estimators, n=3 gives the highest accuracy of about 58%.



Random Forest: Applying RF Regressor, using different number of estimators we see that after n_estimators= 1000 there is no gain in the accuracy.
The highest possible accuracy is about 0.73.



SVR: Applying SVR, rbf method gives negative value. Linear has values close to zero. Testing some values for the parameters, method ‘poly’ with gamma=5 and c=1 has the best accuracy of 0.57

Conclusion: Linear Regression is not a good model to predict the dependent variable y because the R-Square is very low. Perhaps there are more important features in the housing price which we have not seen in this model. Between KNN, Adaboost, SVR and Random Forest, the latter one by using 1000 estimators has the highest value for accuracy with 73%.