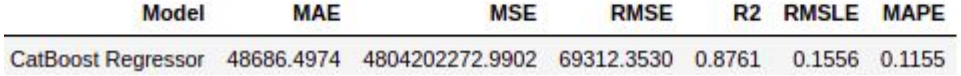
Stone house hunting case analysis

By Mark Brakhas

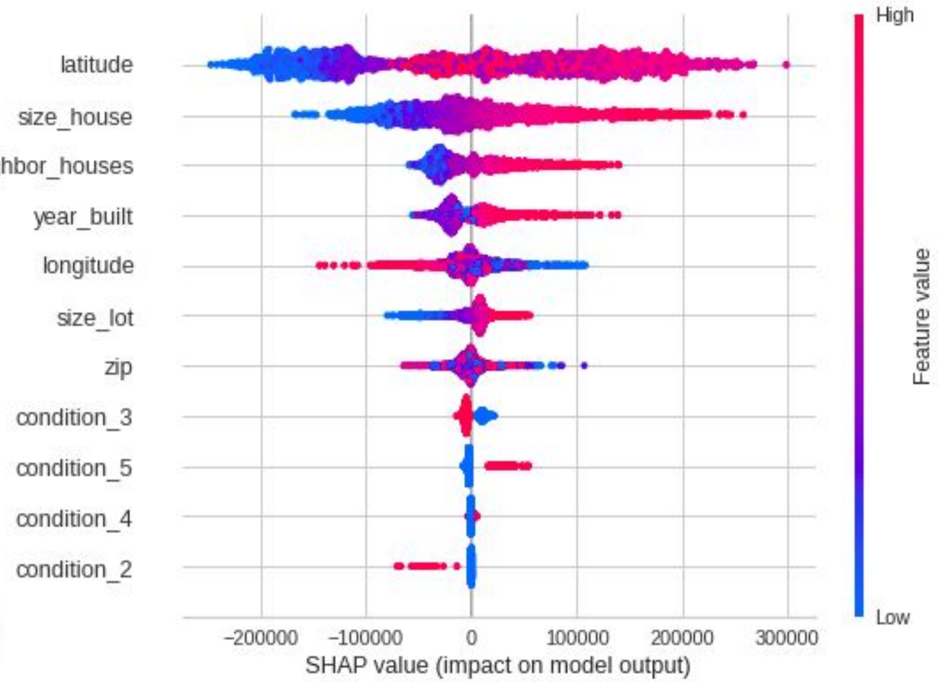
1. Prediction model and interpretation

Modeling Approach:

After removing unnessaccary columns and doing initial descriptive analysis, I removed price outliers after I checked their unpleasant effect on model performance. I selected the most perfect model by checking most of relevant models by pycaret and finally Catboost Regressor is my selected one. I didn't select tuned model as tuning reduced the model performance. Then I used shap to do the feature importance analysis where I used these results for the next step house hunting selection engine.



Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
CatBoost Regressor	48686.4974	4804202272.9902	69312.3530	0.8761	0.1556	0.1155



Conclusions from model interpretation:

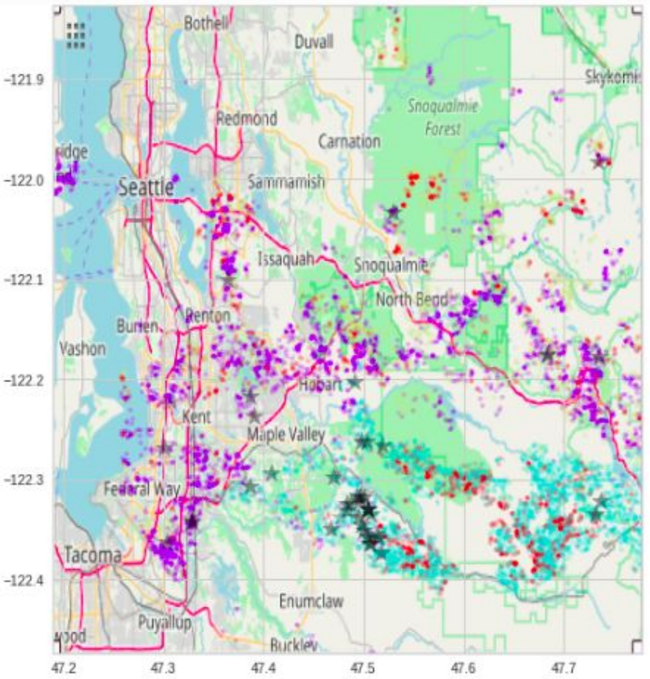
The lower the latitude is the lower the price, longitude has opposite behavior, showing that going to the East and south of the houses area, the house price increases which shows better neighbourhoods. The latitude and size_house features are by far the most important features that affect the model, so I decided to use them with price to define required coefficients for house hunting and selection..

2. House selection methodology

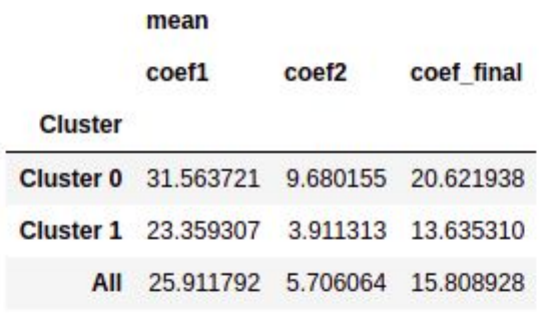
Selection Approach:

Using top most important features from model interpretation, I defined two coefficients after standardizing the data as below :
Coef1 : Latitude / Price: the more this coef is higher the more the house is in better neighborhood with good price.
Coef2 : Size_house / Price : the more this coef is higher the more the house is bigger in size with good price.

Then I considered two selection scenarios :
Scenario 1: Half budget for houses with highest Coef1 and half budget for houses with highest Coef2 ;
Scenario 2: Define Coef_final as mean value of Coef1 and Coef2 and use all budget for selecting houses with highest Coef_final
Finally after reviewing the results, I decided to go for **Scenario 1 ,** because I believe a house with one very good advantage is easier to sell with higher profit than a house with two middle weighted advantage.
I also did Kmeans clustering using all features together and found out that my selected scenario is only from cluster 0 and cluster 1 among the 4 clusters which also showed that the selection scenarios makes scene.

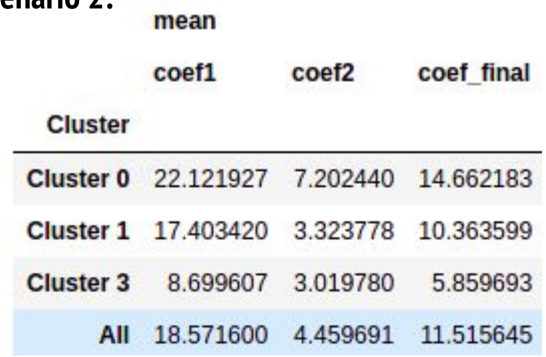


Scenario 1:



	mean		
	coef1	coef2	coef_final
Cluster			
Cluster 0	31.563721	9.680155	20.621938
Cluster 1	23.359307	3.911313	13.635310
All	25.911792	5.706064	15.808928

Scenario 2:



	mean		
	coef1	coef2	coef_final
Cluster			
Cluster 0	22.121927	7.202440	14.662183
Cluster 1	17.403420	3.323778	10.363599
Cluster 3	8.699607	3.019780	5.859693
All	18.571600	4.459691	11.515645