Great Learning
POWER AHEAD

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Capstone Project

HOUSE PRICE PREDICTION

A tool to predict the cost of the house based on different features that helps a buyer find the perfect match and a seller sell in the right price. This tool works like a real estate agent without the brokerage being incurred and offering the best deals to both the parties.

Flow of the Presentation

1. Introduction to the Business Problem – Data & Around the Data
2. Understanding the Business Problem – Objective, Scope & Constraints
3. Exploratory Data Analysis Summary
4. Data Transformation and Assumptions
5. Modelling – Building and Tuning
6. Business Recommendations

INTRODUCTION TO THE BUSINESS PROBLEM

- THE VALUE OF A HOUSE JUST CANNOT BE DEFINED IN CURRENCY .
- REAL ESTATE INDUSTRY IS ONE OF THE MOST UNPREDICTABLE INDUSTRY WHICH NEEDS TO BE KEPT IN CHECK.
- THE PRICE OF THE HOUSE IS NOT JUST THE LOCATION AND THE SIZE OF PROPERTY, LOT OF OTHER FEATURES ALSO CONTRIBUTE IN DEFINING THE ACTUAL COST.

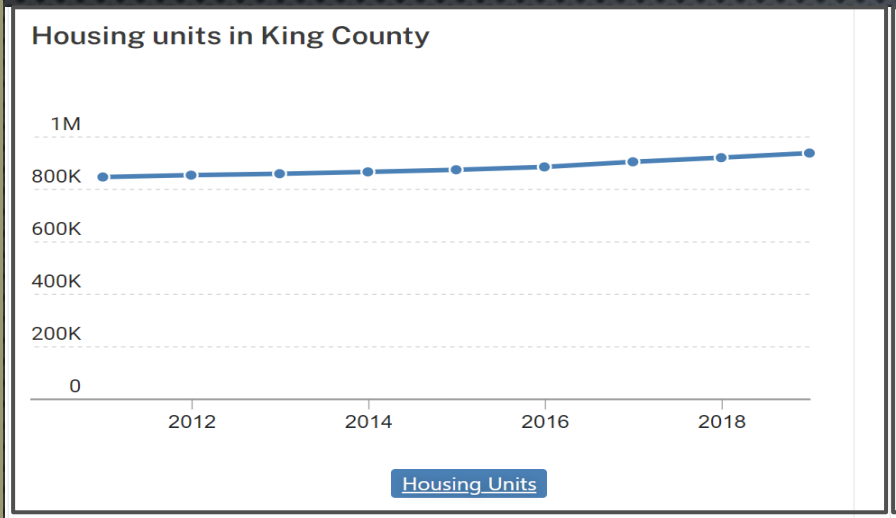
BUYING A HOUSE BY NOT OVER-PAYING AND SELLING THE HOUSE AT THE RIGHT PRICE NEEDS NO SKILL ANYMORE, WHY PAY A BROKERAGE WHEN WE ARE HERE ?

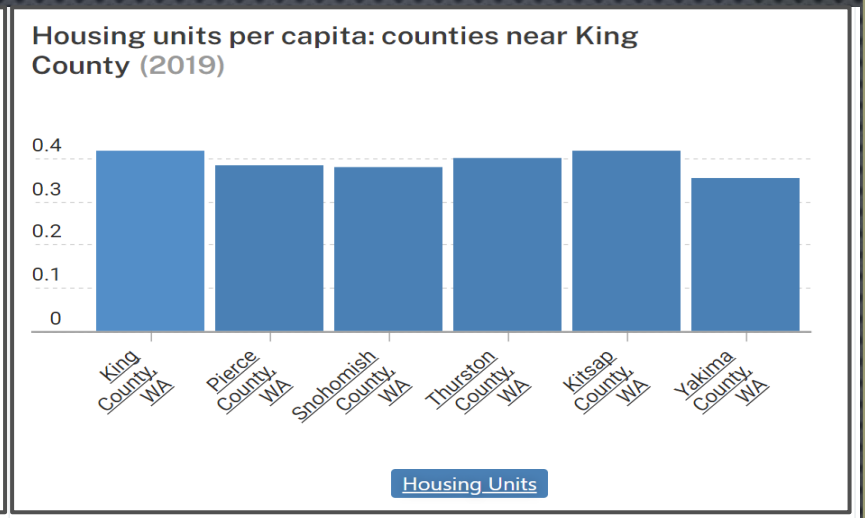
About the data and around the data

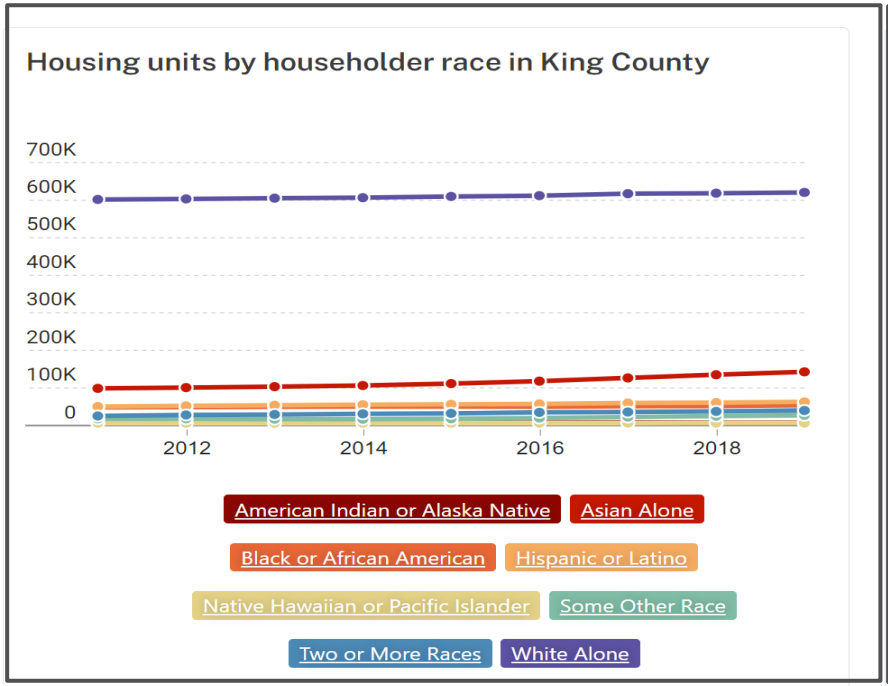
- The data provided to us are of over 21600 houses that have been sold in 2014-2015 in King County, Washington. We have been provided with over 20 features and not just the property sizes

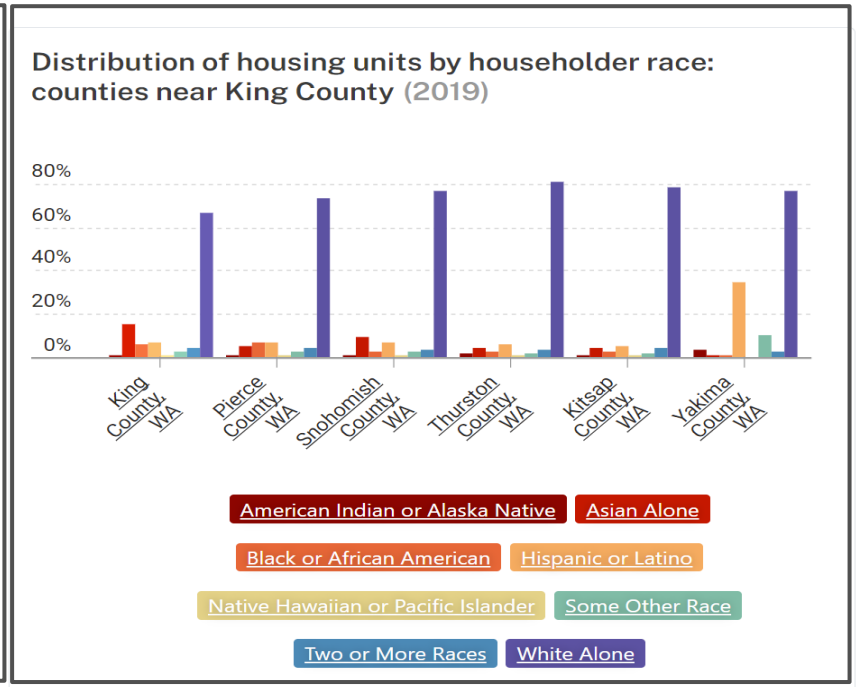
- Almost 29% of Washington's population i.e Around 2.29 Million Residents resides in King County (as per Wikipedia and kingcounty.gov)
- So it indeed is a big market for real estate to thrive and flourish.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 22 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Price                     21613 non-null  float64
 1   Bedrooms                  21505 non-null  float64
 2   Bathrooms                 21505 non-null  float64
 3   Total_Living_Area         21596 non-null  float64
 4   Land_Plot_Area            21571 non-null  float64
 5   Total_Floors              21571 non-null  object 
 6   Sea_view                  21612 non-null  object 
 7   Sight                     21556 non-null  float64
 8   House_Condition           21556 non-null  object 
 9   Quality_Rating            21612 non-null  float64
 10  Living_Area_excl_Basement 21612 non-null  float64
 11  Basement_Area             21612 non-null  float64
 12  Yr_built                  21612 non-null  object 
 13  Yr_renovated              21613 non-null  float64
 14  Zipcode                   21613 non-null  float64
 15  Latitude                  21613 non-null  float64
 16  Longitude                 21613 non-null  object 
 17  Total_Living_Area_2015    21447 non-null  float64
 18  Land_Plot_Area_2015       21584 non-null  float64
 19  Furnished                 21584 non-null  float64
 20  Total_Property_Area       21584 non-null  object 
 21  Yr_sold                   21613 non-null  int64  
dtypes: float64(15), int64(1), object(6)
memory usage: 3.6+ MB
```









- MORE INFORMATION W.R.T HOUSING IN KING COUNTY

Housing units, July 1, 2019, (V2019)	**970,301**	**3,195,004**
Owner-occupied housing unit rate, 2015-2019	56.9%	63.0%
Median value of owner-occupied housing units, 2015-2019	$549,200	$339,000
Median selected monthly owner costs -with a mortgage, 2015-2019	$2,477	$1,886

BUSINESS PROBLEM UNDERSTANDING

OBJECTIVE

- To collect the data and additional parameters that may contribute some value in deciding the price of the house.
- Build a system that analyses all the input features and predicts the price of the house.

BUSINESS PROBLEM UNDERSTANDING

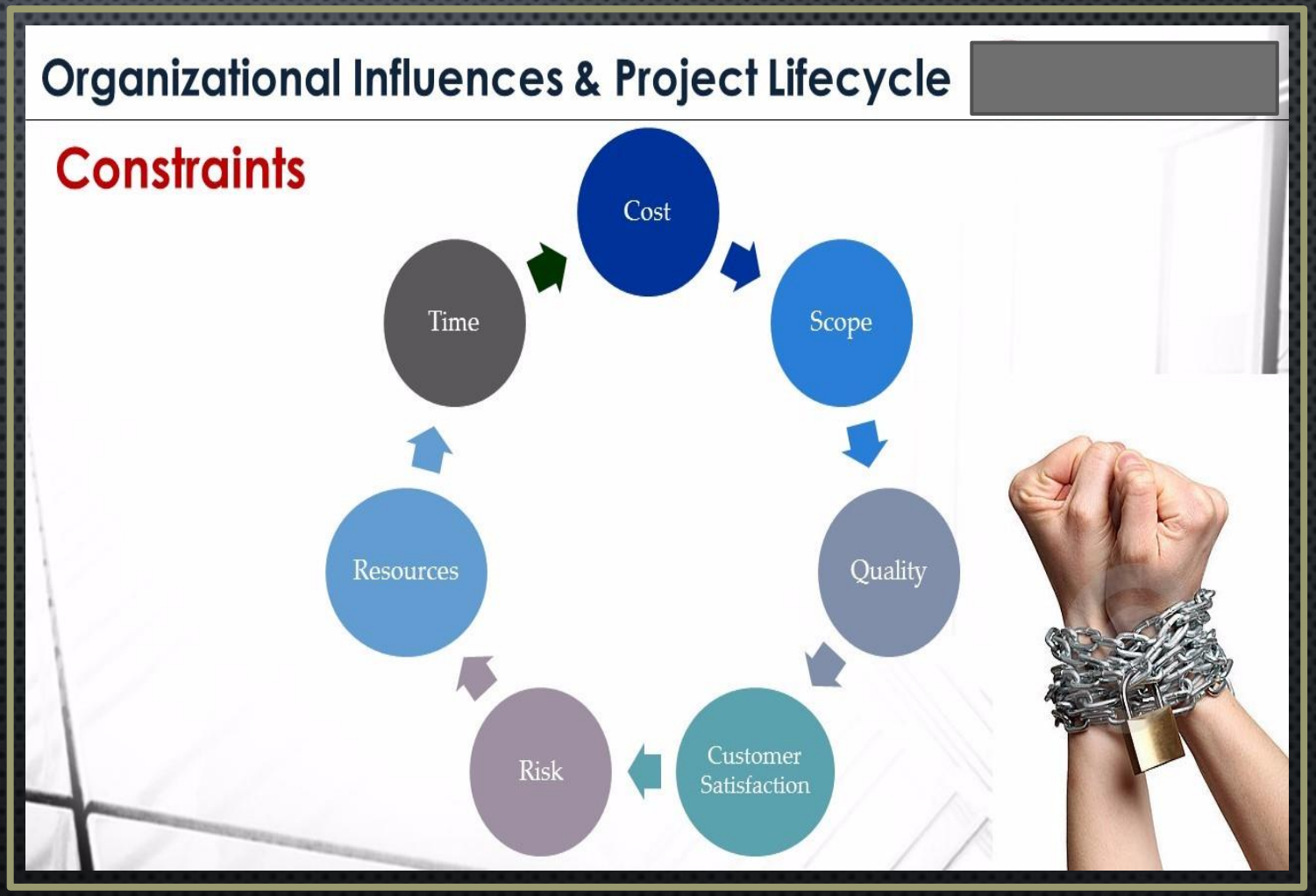
SCOPE

- W.R.T SELLER
- W.R.T BUYER
- ELIMINATE AGENT BROKERAGE
- BUILD A CENTRAL SYSTEM CONTAINING ALL THE DATA AND DEVELOP AN INTERACTIVE WEBSITE AND MOBILE APPLICATION.

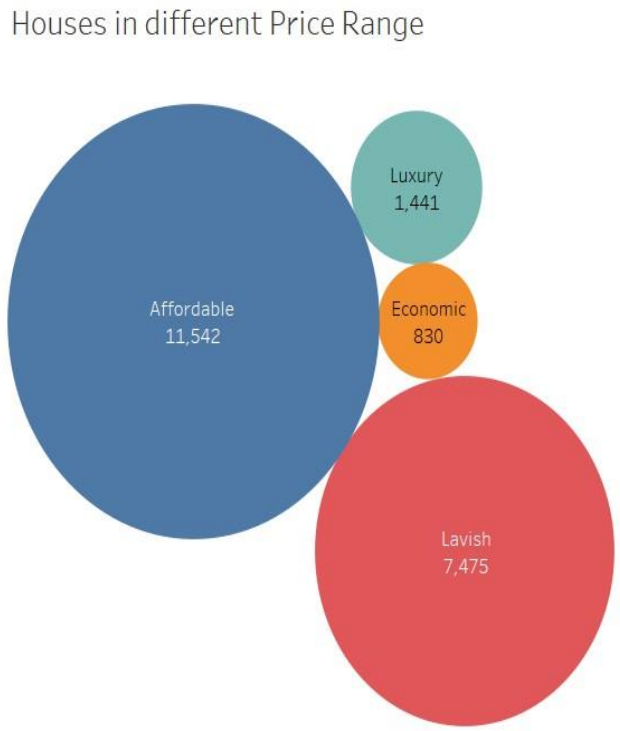
BUSINESS PROBLEM UNDERSTANDING

CONSTRAINTS

- OUTLIERS
- MISSING DATA
- COST OF DATA COLLECTION
- TIME TO MAKE THE SYSTEM LIVE
- UPDATING RESOURCES AND DATA WHEN SOLD



EXPLORATORY DATA ANALYSIS SUMMARY



Houses in different Price Range Stats

Price (group)	Count of Houses	Avg. Price	Min. Price	Max. Price
Economic	830	168,361	75,000	200,000
Affordable	11,542	352,628	200,126	500,000
Lavish	7,475	678,531	500,007	1,000,000
Luxury	1,441	1,535,982	1,010,000	7,700,000

- We have made 4 clusters to group houses by prices and we can see the distribution.
- From the EDA we did , following were our derivations : -
  - House Conditions, Quality Ratings, Furniture, #Bedrooms, Sea View, Sight, Renovated, Total Living Area and Property Area have a direct influence in the price variation.
  - Few houses – Renovated & Sea View
  - Most of the variables have outliers but they don't need to be treated.
  - Plot area and Total Property area have High correlation.
  - Some variables don't add direct value to the analysis

DATA ASSUMPTIONS AND TRANSFORMATIONS

ASSUMPTIONS

- CID DOESN'T ADD VALUE , SO WE DELETE IT.
- OUTLIERS ARE GENUINE , SO WE DON'T TREAT THEM.
- VARIABLES WITH HIGH CORRELATION HINDER THE MODEL, SO THEY WERE REMOVED.
- DATA COLLECTED IS CORRECT.

TRANSFORMATIONS

- HOUSE SOLD DATE IS CONVERTED TO YEAR SOLD.
- DATA WITH $ WERE CONVERTED TO NULLS AND WE IMPUTED VALUES USING KNNIMPUTER.
- BASEMENT AREA AND RENOVATED YEAR CONVERTED TO MEANINGFUL VARIABLES.
- MISSING VALUE TREATMENT.

MODEL BUILDING

THE PROCESS OF MODELING MEANS TRAINING A MACHINE LEARNING ALGORITHM TO PREDICT THE LABELS FROM THE FEATURES, TUNING IT FOR THE BUSINESS NEED, AND VALIDATING IT ON HOLDOUT DATA

- SPLIT TRAIN AND TEST SETS IN PROPORTION 80 : 20 RESPECTIVELY AND FURTHER SPLIT TRAINING SET INTO TRAIN AND VALIDATION SET IN PROPORTION OF 80 : 20 AGAIN.
- DIFFERENT MODELS USED IN THE REGRESSION PROBLEM ARE: -
  - LINEAR REGRESSION
  - RIDGE REGRESSION
  - LASSO REGRESSION
  - KNN REGRESSOR
  - DECISION TREE REGRESSOR
  - ENSEMBLE TECHNIQUES-
    - RANDOM FOREST REGRESSOR
    - BAGGING REGRESSOR
    - ADABOOST REGRESSOR
    - GRADIENT BOOST REGRESSOR

Method	Train Set Score	Train RMSE	Train MAE	Validation Set Score	VL RMSE	VL MAE
Linear Regression	0.751	185,664.026	117,361.680	0.722	176,957.798	115,768.905
Ridge	0.749	186,456.303	118,055.010	0.720	177,421.823	116,283.410
Lasso	0.751	185,665.279	117,375.103	0.722	176,941.666	115,759.006
KNN	0.672	213,190.487	129,999.900	0.467	244,777.503	157,857.103
Decision Tree	1.000	2,194.017	71.669	0.674	191,411.034	101,401.362
Random Forest	0.982	49,693.283	26,077.543	0.865	123,146.018	69,935.498
Bagging	0.974	60,012.467	30,181.366	0.849	130,444.046	74,111.953
AdaBoost	0.275	316,966.423	286,823.650	0.087	320,524.309	288,111.433
GradientBoost	0.903	115,856.969	73,362.355	0.853	128,725.192	77,148.948

MODEL TUNING

- SINCE THE ENSEMBLE TECHNIQUES WERE DOING BETTER THAN THE LINEAR REGRESSION, WE TUNED THE MODELS BASED ON ENSEMBLE TECHNIQUES AND TRIED TO EXPERIMENT WITH DECISION TREE AND KNN AND CHECK PERFORMANCE ON VALIDATION AND TEST SET.
- AS EXPECTED, ENSEMBLE METHODS HAD EVEN BETTER PERFORMANCE WHEN MODEL PARAMETERS WERE TUNED.

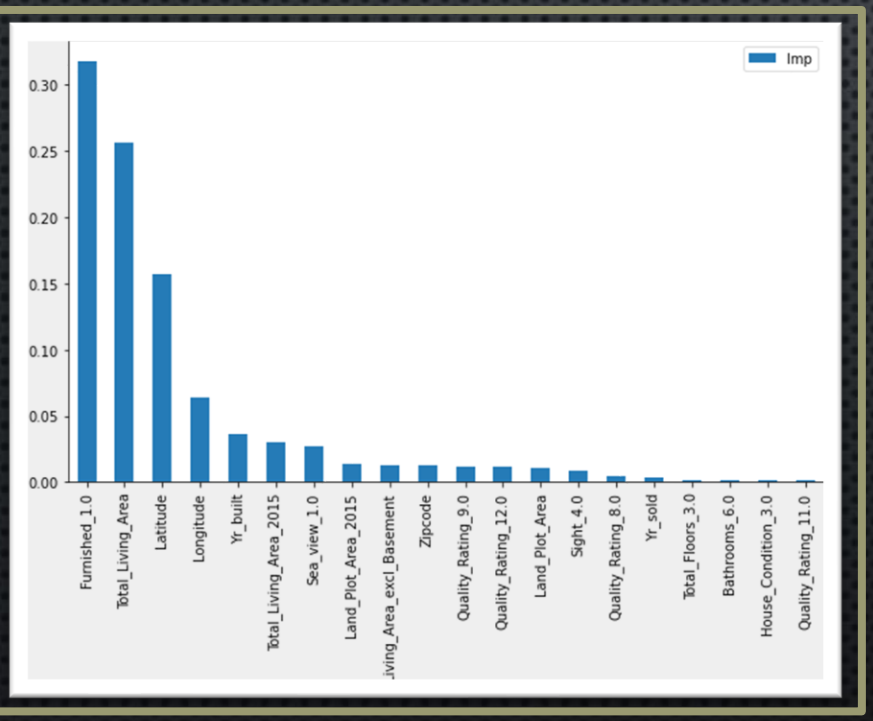
Method	Train Set Score	Train RMSE	Train MAE	Validation Set Score	VL RMSE	VL MAE	Test Set Score	Test RMSE	Test MAE
KNN	1.000	2,194.017	71.669	0.498	237,702.556	152,253.945	0.514	262,188.676	154,702.075
Decision Tree	0.892	122,034.639	72,534.623	0.785	155,525.183	89,081.875	0.791	171,813.629	92,277.095
Random Forest	0.936	93,935.140	46,211.195	0.857	126,769.836	71,050.483	0.877	132,031.955	72,040.905
Bagging	0.983	49,093.732	25,811.597	0.864	123,585.231	69,504.228	0.886	127,061.881	70,130.719
AdaBoost	0.683	209,661.074	154,433.178	0.595	213,525.795	155,223.654	0.647	223,506.030	157,095.559
GradientBoost	0.968	66,625.267	37,403.887	0.899	106,494.064	61,809.652	0.901	118,570.452	64,707.826

20 IMPORTANT FEATURES IN DIFFERENT MODELS

```
['Yr_built','Latitude','Total_Living_Area','Bathrooms_7.75','Is_renovated_1','Sea_view_1.0','Furnished_1.0',
 'House_Condition_3.0','House_Condition_5.0','Quality_Rating_3.0','Quality_Rating_4.0','Quality_Rating_5.0',
 'Quality_Rating_6.0','Quality_Rating_7.0','Quality_Rating_8.0','Quality_Rating_9.0','Quality_Rating_10.0',
 'Quality_Rating_11.0','Quality_Rating_12.0']]
```

Linear Regression

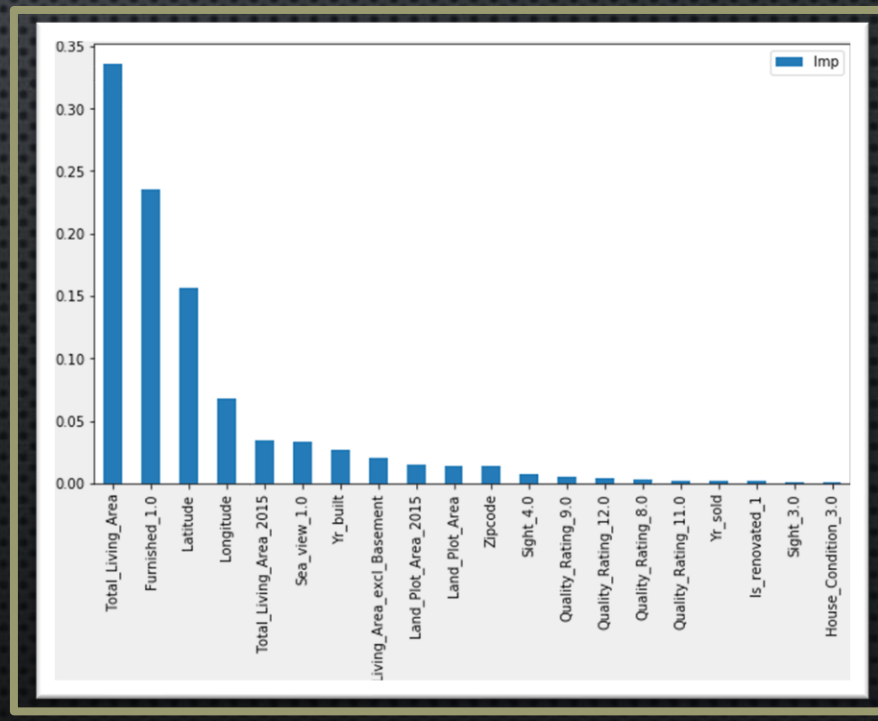
using Coefficients



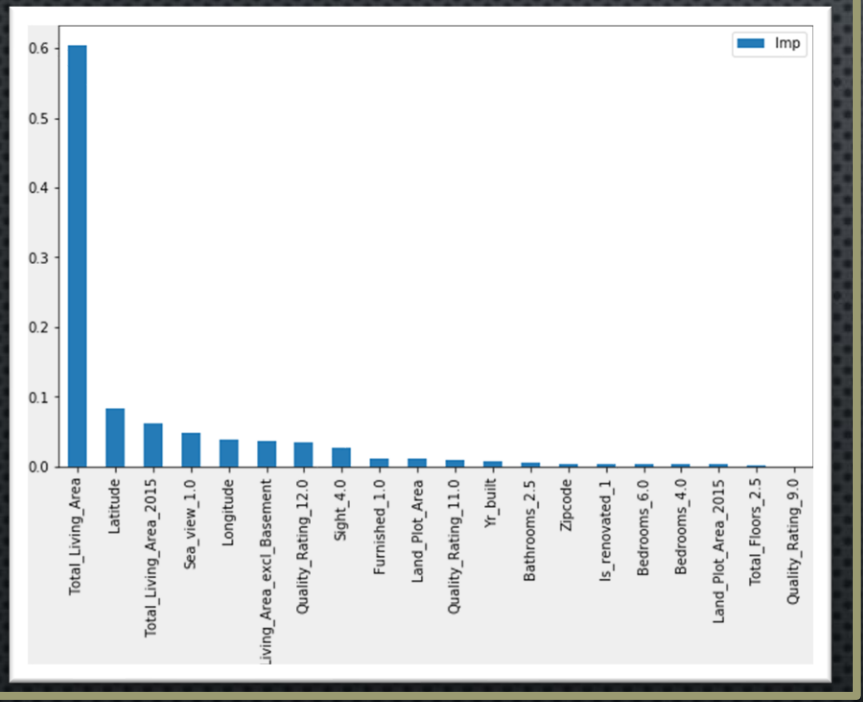
Using feature_importance

Decision Tree

Random Forest



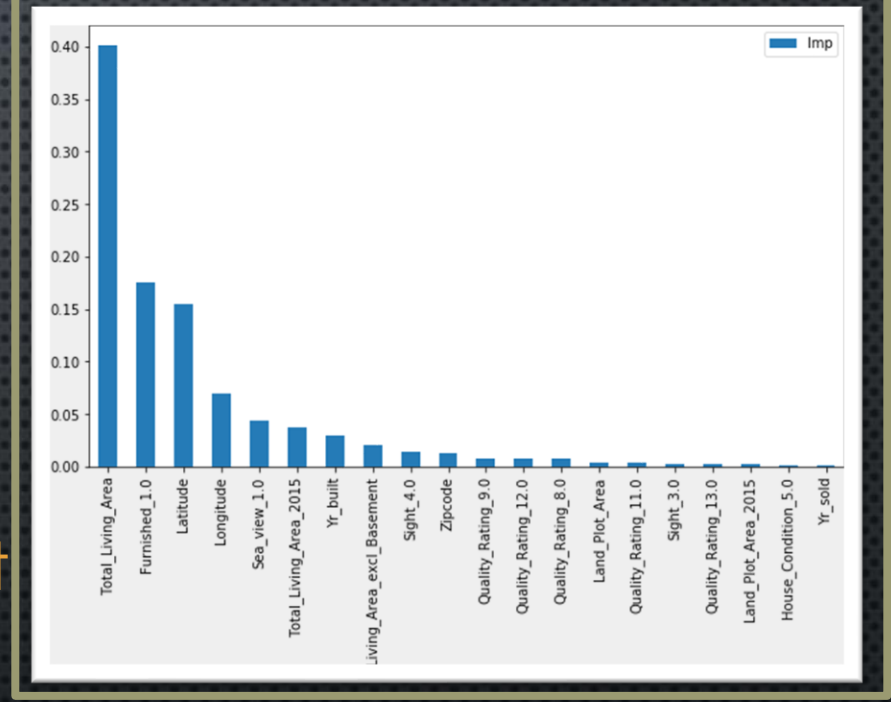
20 IMPORTANT FEATURES IN DIFFERENT MODELS



Using feature_importance

Ada Boost

Gradient Boost

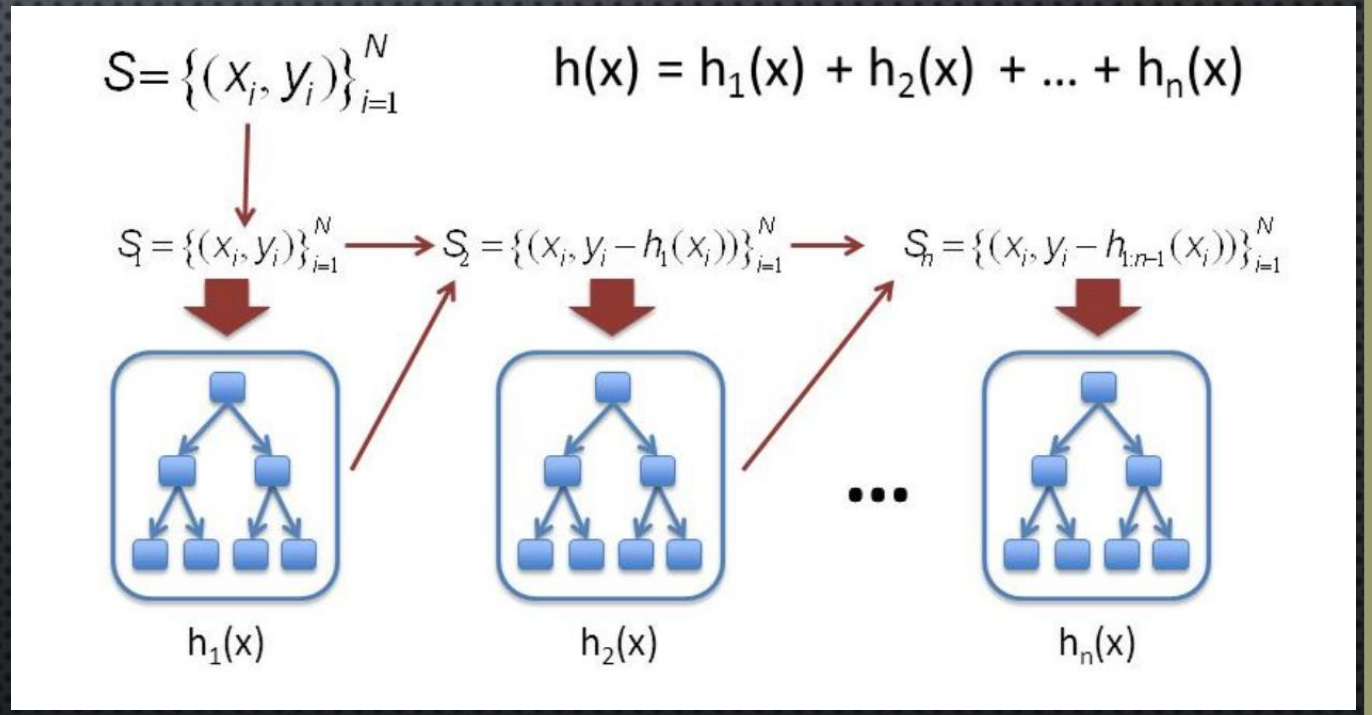



```
[['Basement_Present_1','Bathrooms_2.25','Bathrooms_2.5','Bedrooms_4.0','Bedrooms_6.0','Furnished_1.0',
'House_Condition_3.0','House_Condition_5.0','Is_renovated_1','Land_Plot_Area','Land_Plot_Area_2015','Latitude',
'Living_Area_excl_Basement','Longitude','Quality_Rating_11.0','Quality_Rating_12.0','Quality_Rating_13.0',
'Quality_Rating_7.0','Quality_Rating_8.0','Quality_Rating_9.0','Sea_view_1.0','Sight_2.0','Sight_3.0','Sight_4.0',
'Total_Floors_2.0','Total_Living_Area','Total_Living_Area_2015','Yr_built','Yr_sold','Zipcode','Price']]
```

Common Features

MODEL APPROACH USED

- The linear regression models do well but their performance is not up to the mark, they have reasonably higher RMSE and low R2 scores.
- Ensemble techniques do pretty well and hence we tune them and Based on the model performances, we can conclude that Gradient Boosting is the most optimal model with the least RMSE and best R2 score of around 90% on the test set.



INSIGHTS AND INFERENCES

- Quality Ratings, House Conditions, Furniture, Renovations, Sea View and the Location of the house were the major contributors apart from the #Bedrooms and house, plot and property areas.
- About 10-15% population live in the extreme (poor or rich) where as around 85% live in houses between 200,000 $ and 1,000,000 $.

- Only ~1% houses had a sea view and ~4% houses were renovated. Around 70% houses had a House Condition of 3.0 and had Quality Rating of 7 or 8.

Clusters made on House Pricing Category(Luxury, Lavish, Affordable and Economic)-

Cluster	Yr_built	Bedrooms	Bathrooms	Floors	Seaview	Furnished	HouseCondition	Rating	LivingArea_2015	LandPlotArea_2015	LivingArea_excl_Basement	BasementArea	TotalLivingArea	LandPlotArea	Price
C1	1975	4	3.5	2	1	1	3	10	3243	20483	3481	786	4266	27586	$19,75,656
C2	1972	4	2.5	1.5	1	1	3	9	2443	12308	2284	412	2696	14373	$7,98,341
C3	1982	3	2.25	1.5	0	0	3	8	2324	164221	2271	260	2530	249590	$5,62,848
C4	1970	3	2	1	0	1	3	7	1739	9464	1502	221	1723	9978	$3,70,771

RECOMMENDATIONS AND BUSINESS SCOPE

- Collect the data of houses sold and bought for the last 5 years and feed the data to the system.
- Develop interactive website and mobile application for the same.
- Plan online marketing campaigns as "brokerage killers" and "Best Houses at Right Price".
- Target the right audience.
- We can tie-up with the state for making this service a public service and charging a royalty and maintenance fee.
- Timely updates and model optimization.
- Build models for renting apartments and leasing properties.
- Integrate with financial agencies that provide loans and ease of money transfer wallets.

Thank you

Submitted by-

Aaditya Desai