House Price Prediction

Kavya Sree Kaitepalli
kavyasree.k20@iiits.in

Himaja Anchuri
himaja.a20@iiits.in

Abstract**—** This paper focusses on price prediction of houses using regression model. We perform regression on the data of house sales collected from the Washington city in the years of 2014-2015. We performed Sequential Feature Selection (SFS) to know the best set of attributes, and then performed linear regression with one attribute (which has the highest correlation with the response variable) and multiple linear regression with the best subset of variables provided by the SFS algorithm. We also implemented Principal Component Analysis (PCA) for dimensionality reduction. Additionally, we performed Decision Tree Regression in order to reduce root mean squared error. This paper uses Root mean squared error, coefficient of determination (R^2), adjusted coefficient of determination as the performance metrics.

Keywords—Regression, PCA, SFS

I. INTRODUCTION

One of the most necessary assets for human life is home. Now-a-days, the real-estate market is at its peak. The market demand for housing is increasing rapidly because our population is raising day by day. And the prices are also fluctuating frequently. People are looking to buy a new home in their budgets by analyzing the market strategies. The problem with estimating the prices of houses manually is that there is 25% chance of error. People who don't know the actual price of house may suffer loss of money. Hence, the aim of our project is to predict actual house price with more accuracy.

Through this project, we can identify the most important house attributes that influence the model ability to predict future house prices.

II. DATASET DESCRIPTION

The data consists of a total of 20 attributes and a price column for prediction.

The attributes are as follows:

1. id – unique id for the house
2. date – selling date of house
3. bedrooms - Number of Bedrooms in the house
4. bathrooms - Number of bathrooms in the house
5. sqft_living - square footage of the home
6. sqft_lot - square footage of the lot
7. floors - Total number of floors in house
8. waterfront - House which has a view to a waterfront
9. view – characteristics of the property
10. condition - How good the overall condition of the house is
11. grade - overall grade given to the housing unit (based on King County grading system)
12. sqft_above - square footage of house apart from basement
13. sqft_basement - square footage of the basement
14. yr_built – Year in which the house is built
15. yr_renovated - Year when house was renovated last
16. zipcode – zipcode of the area of house
17. lat - Latitude coordinate
18. long - Longitude coordinate
19. sqft_living15 – living area square footage of nearest 15 neighbouring houses
20. sqft_lot15 – lot area square footage of nearest 15 neighbouring houses
21. price – the variable to be predicted

III. METHODOLOGY

A. *Data Pre-Processing*

1.Dealing with missing data:
Check for missing values in the dataset. Variables with more than 50% missing data should be removed, if missing data is very less, we need to fill the missing values with mean values. But our dataset doesn't have any missing values.

2.Duplicated data:
In this step, we remove the duplicates from the dataset. In our data, there are houses with duplicated ids which have been removed.

3.Outliers:
The house sales data do not have many outliers. We found only one data point to be outlier since it has 33 bedrooms with only 1.75 bathrooms. So, we removed it.

4.Feature Engineering:
Extract the year in which the house was sold from the date attribute.
Replace attribute "yr_built" with attribute age by deducting the sold_year from the yr_built.

B. *Exploratory Data Analysis*

Exploratory data analysis is an essential step before building a regression model. This allows us to identify the data's hidden patterns.

1.Heat map:

Creating a Correlation Matrix for all the features. If there is very high correlation between two features, keeping both of them is not a good idea most of the time not to cause overfitting.

Plotting a heat map for the house data to know the correlation between attributes of the dataset.

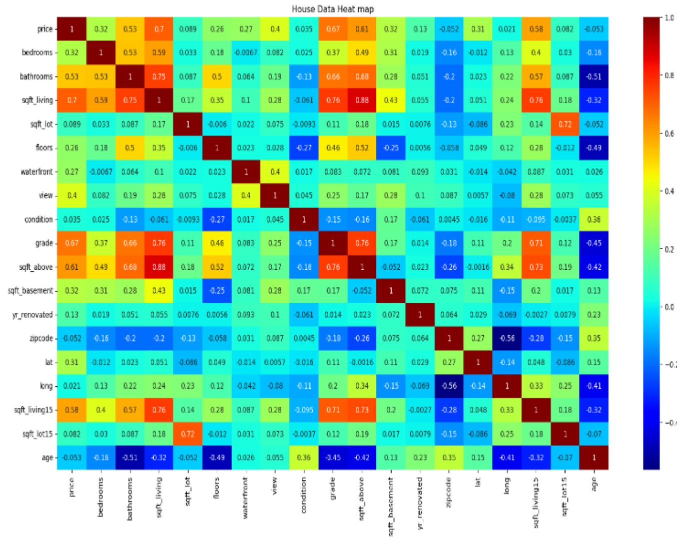


Fig. 1. Heat Map

From the above heat map, we see that there is considerable correlation between the attributes like sqft_lot and sqft_lot15, sqft_above and sqft_living, sqft_living and sqft_living15 etc.

2.Finding the attributes that highly affect the Price (variable to be predicted):

Correlation of all attributes with that of Price variable are shown below (in descending order):

```
price               1.000000
grade               0.867812
sqft_living         0.861394
sqft_living15       0.823776
bathrooms           0.782877
sqft_above          0.782465
bedrooms            0.611080
floors              0.541588
view                0.451881
sqft_basement       0.326447
long                0.218968
waterfront          0.181932
lat                 0.141801
sqft_lot15         -0.084143
sqft_lot           -0.096295
yr_renovated       -0.121734
condition          -0.357891
zipcode            -0.378659
age                -0.544039
```

Sqft_living and grade are highly affecting the price of house.

3.Scatter plot of price with some of the continuous attributes with attribute "condition" as hue:

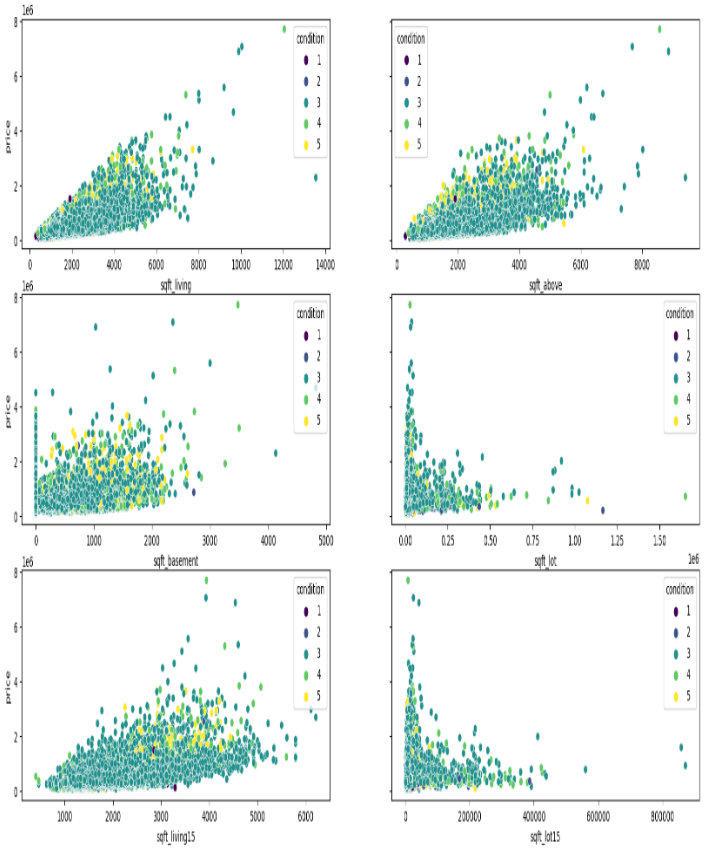


Fig 2

From the figure, we know that sqft_living and sqft_above has a positive correlation with the price attribute.

4.Box plot between Price and Bathrooms:

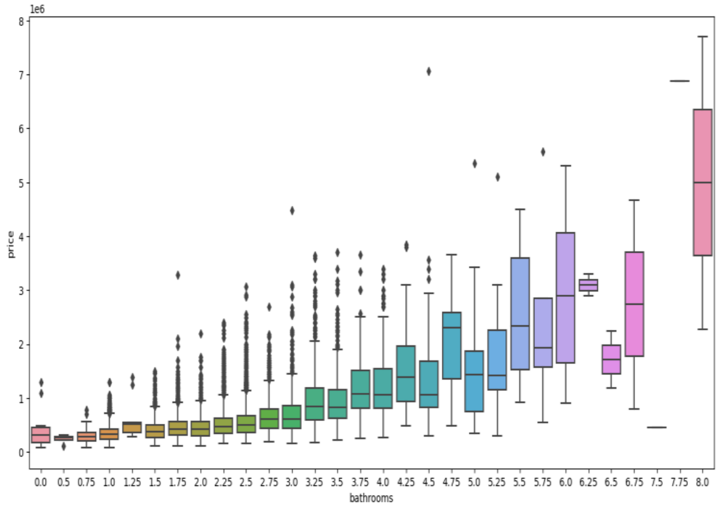


Fig 3

We observe a positive correlation for bathrooms with price. The price increases as the number of bathrooms increase.

5.Counting number of houses with respect to grades:

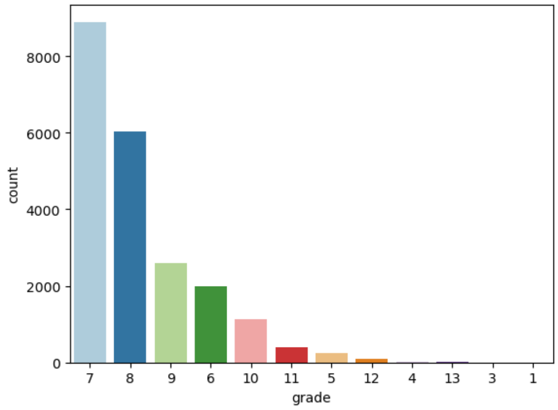


Fig4

Number of houses with grade of 7 and 8 are more in number, that is, we can say that most of the houses have a good grade and they are good enough.

6.Counting number of houses with respect to Floors:

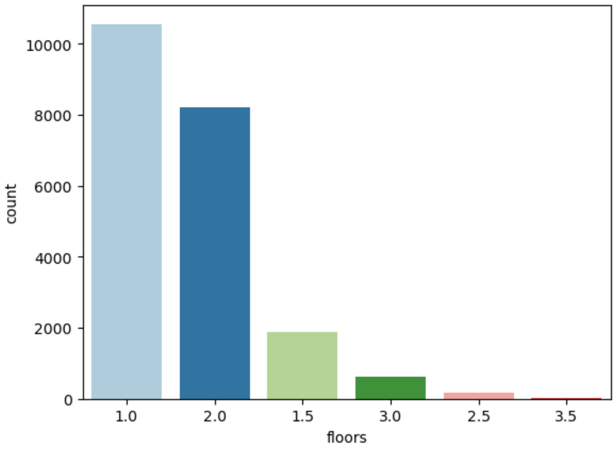


Fig5

There are more houses with a single floor and as the number of floors increase, count of houses is decreasing.

7.Counting number of houses with respect to Floors:

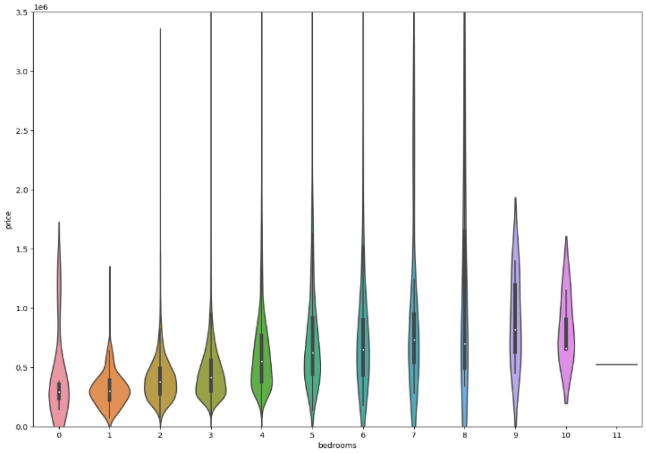


Fig 6

From the above violin plot, the height of violin plot says that the houses having bed rooms from 4-9 have higher prices compared to the houses with 1 bed room or more than 8 bed rooms, whereas the width of the violin plot says that the houses with 1-5 bedrooms are more compared with houses with more than 5 bed rooms.

8.Histograms for attributes to see if the data has outliers or if they need to be normalised:

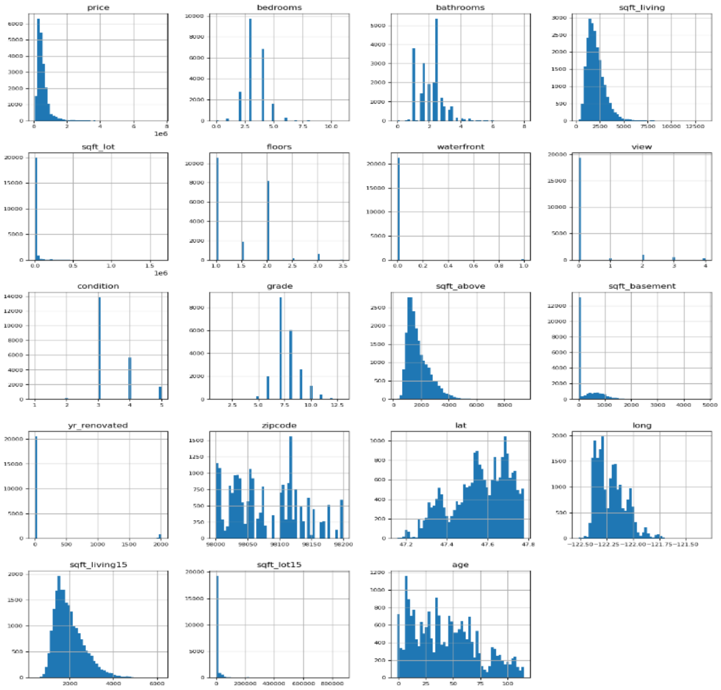


Fig 7

9. Durbin Watson Test:

The Durbin-Watson test is a statistical test used to detect the presence of autocorrelation at lag 1 in the residuals (prediction errors) from a regression analysis.
The test statistic always ranges from 0 to 4 where:

- d = 2 indicates no autocorrelation
- d < 2 indicates positive serial correlation
- d > 2 indicates negative serial correlation

We got a DW factor around 2.06 which means that there is no considerable autocorrelation between attributes.

C. ***Application of Regression Model***

The data set is split into training data (80%) and testing data (20%). The model is trained with the training data and the price is predicted for testing data. In this, we used k-fold cross validation technique for evaluating the performance of the regression model.
Firstly, we apply simple linear regression with independent variable as sqft_living (since the price is highly dependent on this attribute). After that, we perform multiple linear

regression considering all the attributes. Next, we apply PCA for dimensionality reduction and SFS for choosing the best set of attributes. Lastly, we also implement binary tree regression on the data to see if there is any considerable increase in the coefficient of determination.

1.Simple Linear Regression Model:

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. One variable, denoted x, is regarded as the predictor, explanatory, or independent variable. The other variable ,denoted y is regarded as the response outcome or dependent variable. Here this model is applied by taking price as response variable and sqft_living as independent variable.

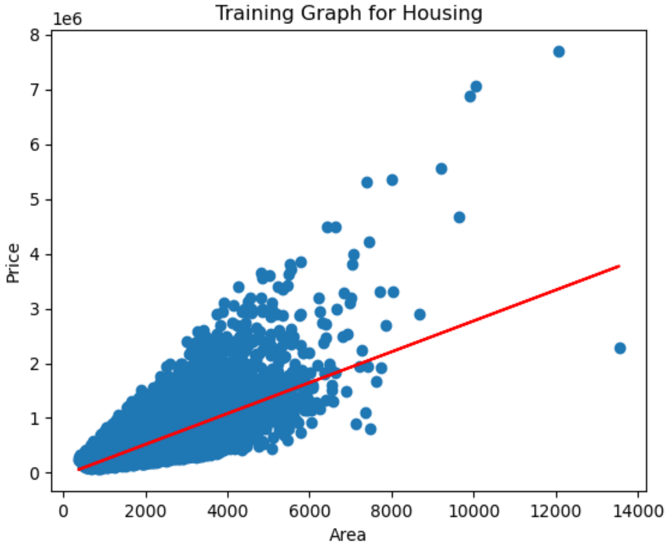
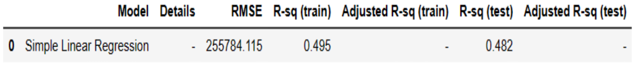


Fig8



	Model	Details	RMSE	R-sq (train)	Adjusted R-sq (train)	R-sq (test)	Adjusted R-sq (test)
0	Simple Linear Regression	-	255784.115	0.495	-	0.482	-

2.Multiple Linear Regression Model:

Here we have more than one regressor and one response variable.

Considering all the attributes (other than id):

We split the dataset into training and testing data and train the regression model with training data and then predict the price values of the testing data.

a) Metrics:

```
Mean absolute error: 126742.97436940242
Mean squared error: 44693189766.2733
Root mean squared error: 211407.63885506432
Coefficient of determination: 0.7018781318493
283
```

b) Sequential Feature Selection (SFS):

It is a technique used for feature selection in which features are sequentially added to an empty candidate set until the addition of further features does not decrease the criterion (metric).

Considering six features and using coefficient of determination as the metric:

The resulting subset of six features:

- `sqft_living`
- `waterfront`
- `view`
- `grade`
- `lat`
- `age`

Coefficient of determination when multiple linear regression is performed on the above subset is: 0.6831552566322299

After coefficient of determination reaching 0.6831552566322299, there is no considerable change in that value of R^2. The maximum R^2 attained (considering all the features) is only around 0.7.

c) Scree plot

The scree plot is used to determine the number of principal components to keep in principal component analysis (PCA).

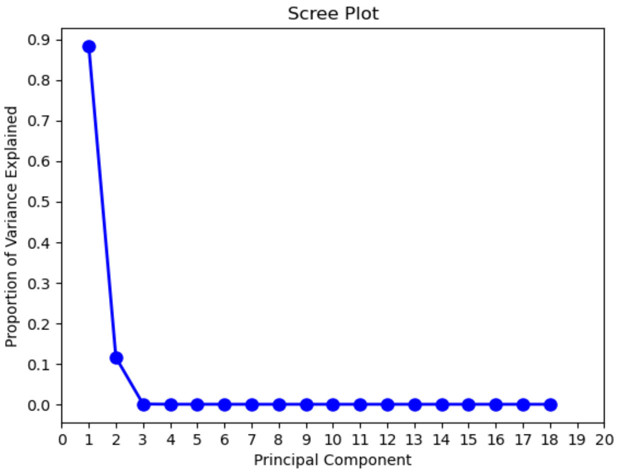


Fig 9

From the above scree plot, we see that most of the variance is explained by the first two principal components.

The variance explained by the first two components is around 99.55%.

Applying multiple linear regression considering only these two components gives out the coefficient of determination to be 0.5865451684901726.

Applying multiple linear regression considering all the attributes along with the principal components gives out coefficient of determination to be 0.7007624652546911.

Applying decision tree regression model considering all the attributes gives out coefficient of determination to be 0.74486958646959. The score is obtained by considering the mean for 10-fold cross validation.

3.Decision Tree Regression Model:

Decision tree regression is a machine learning algorithm that predicts continuous values by observing features of an object and training a model in the structure of a tree to predict data in the future to produce meaningful continuous output.

In decision tree regression, the tree is built by recursively splitting the data into smaller subsets based on the value of a feature that maximizes the reduction in variance of the target variable4. The final result is a tree with decision nodes and leaf nodes. The decision nodes ask questions about the features of the data, while leaf nodes represent predictions or decisions.

IV. Results

Plot between actual and predicted prices
Model: Simple Linear Regression
Attribute considered: sqft_living

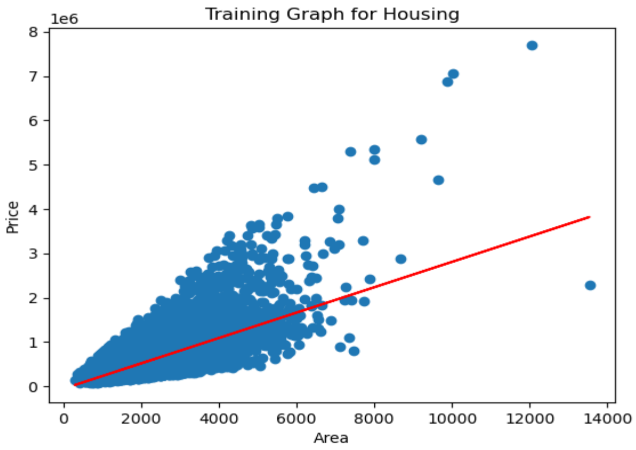


Fig 10

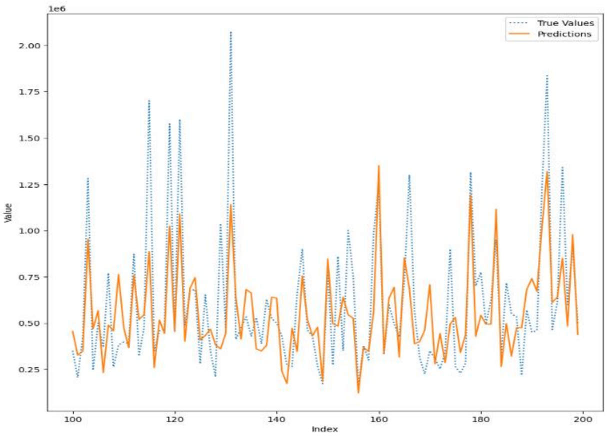
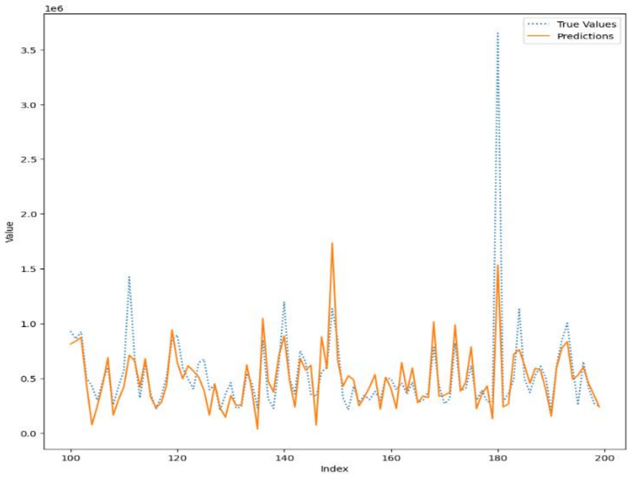


Fig 11
Plot between actual and predicted prices
Model: Multiple Linear Regression



Plot between actual and predicted prices
Model: Decision Tree Regression

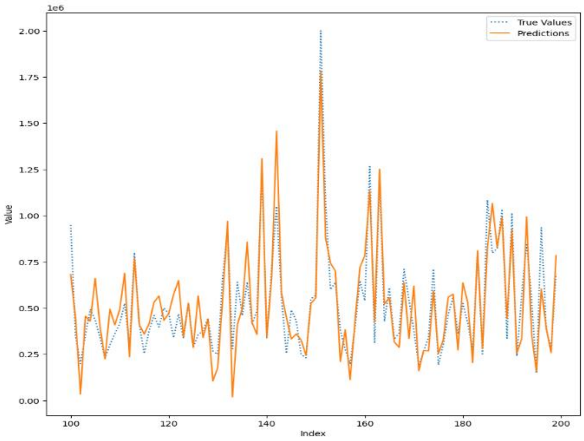


Table 1: Comparisons of Coefficient of determination for different models:

Sl.no	Model	Coefficient of determination
1	Simple linear regression	0.482
2	Multiple linear regression considering all attributes	0.7018781318493283
3	Multiple linear regression with PCs	0.5865451684901726.
4	Multiple linear regression considering both attributes and PCs	0.7007624652546911
5	Decision tree regression	0.74486958646959

V. Conclusion

From the simulation results shown above, it can be concluded that the proposed multiple linear regression model can effectively analyze and predict the housing price to some extent. Admittedly, the prediction accuracy is still limited at specific points, and the universality of the model still needs to be improved in further research.