HOUSING SALE PRICE PREDICTION

By
Obwoge
Getrude

OVERVIEW

For this project, I have used various regression methods to predict the price of houses in the northwestern county,USA. The models start with a very basic simple linear regression advancing towards a complex polynomial regression. The main aim here is to attain a model that will best predict the price of a house with a metric error close to zero.

PROJECT WORKFLOW

BUSINESS UNDERSTANDING

Houses are slowly becoming the biggest investment one can venture into in this era. The newest form of housing investment is the one where investors are now remodeling houses by improving particular features of the house with the goal of increasing its market value. To get a maximum return on investment, these investors ought to know which features of a house increase its value.

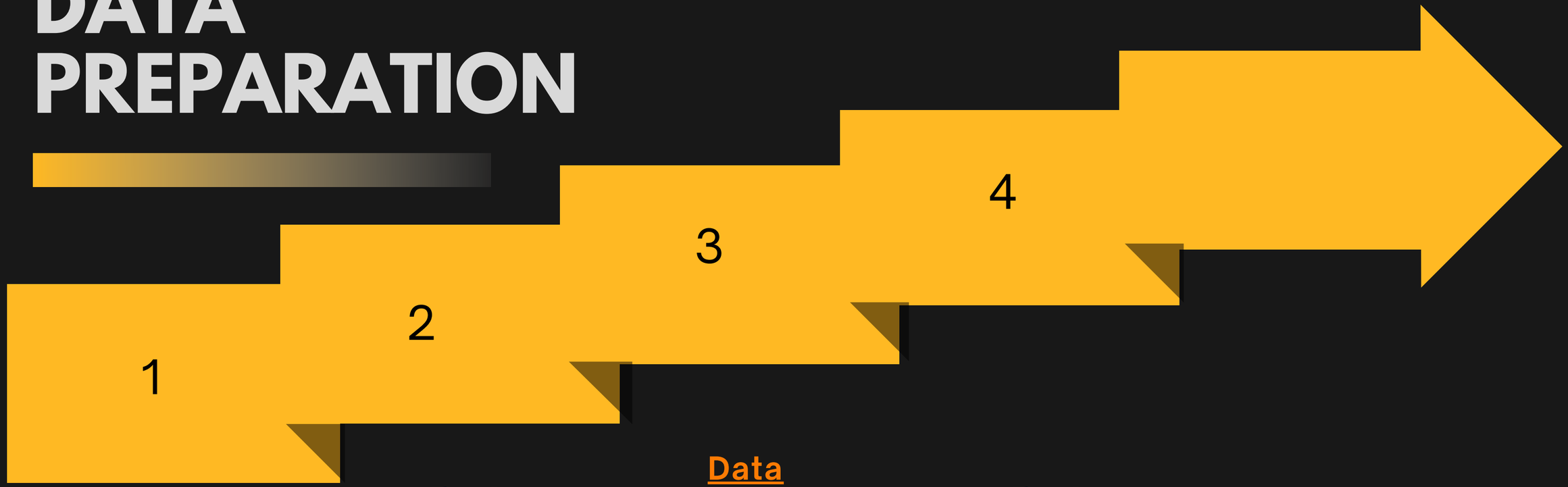
PROBLEM STATEMENT

The main problem that investors willing to get into the housing market face is understanding the key factors that affect house prices. We will perform regression analysis and find the relationship between the predictor variables and the target variable (price). The analysis will help determine which variables impact the price significantly and hence provide knowledge for investors on what features to invest on in the housing market.

DATA UNDERSTANDING

The dataset used in this analysis is the KC housing data set which contains data for 21597 properties in King County and 21 variables.It was obtained from Kaggle, which is a Google LLC subsidiary,an online community for data scientist.It allows users to find and publish realible data sets.

DATA PREPARATION



Loading Data

Importing libraries,reading and loading of dataset

Data Cleaning

Removing null and duplicate values and checking for outliers,anomalies

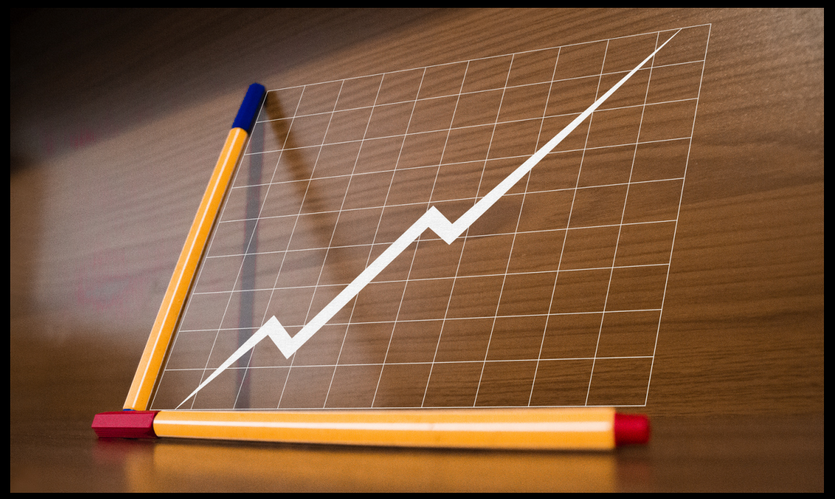
Data Preprocessing

One hot encoding categorical features,standardization and logtransformation of features

Data Splitting

Splitting data into train for raining the model and test for testing the model.

MODELLING



Model	Adjusted R	MAE
Simple	49.4	0.4741
Multiple	73.1	0.3258
Polynomial	84.7	0.2560

EVALUATION

- The model that performed best was the polynomial regression model after standard scaling.
- Its predicted values had the lowest mean absolute error of 0.2560 from the actual data points.
- There was not a great difference between Train Mean Squared Error(0.2560) and Test Mean Squared Error(0.2324) therefore we have not overfitted our model and it is okay. This means that our model will perform in a similar way on different data.
- The model also explains 84.7% variations in house prices.

CONCLUSIONS

- Several factors had a significant influence on price that is:

- Square footage of (the living space,the lot,the basement and Square footage of house apart from basement,
- The year the house was built and when it was renovated
- The grade of the house in terms of in terms of the construction and design of the house
- The location
- The number of bedrooms and bathrooms a house has.

-Waterfronts and views barely impact price,on the contrary houses with neither had very high prices.

RECOMMENDATIONS

RECOMMENDATION 1

Remodel houses to achieve grade 7 in terms of the construction and design of the house in the King County area, meeting grading requirements for the region.

RECOMMENDATION 2

Remodelled houses should attain the average condition of maintance.

RECOMMENDATION 3

During remodelling ensure houses have a maximum of 3 floors,bathrooms and bedrooms.

NEXT STEPS

- More features should be included in the analysis for example number of houses sold or even profit realized.
- Updated data should be used that is the data should be less than 5years old.
- Data spanning atleast five years instead of 2 will make better predictions.

Do you have any questions?


Thank you