Prediction of Real Estate Prices

Manasi Odassery
B94

Overview

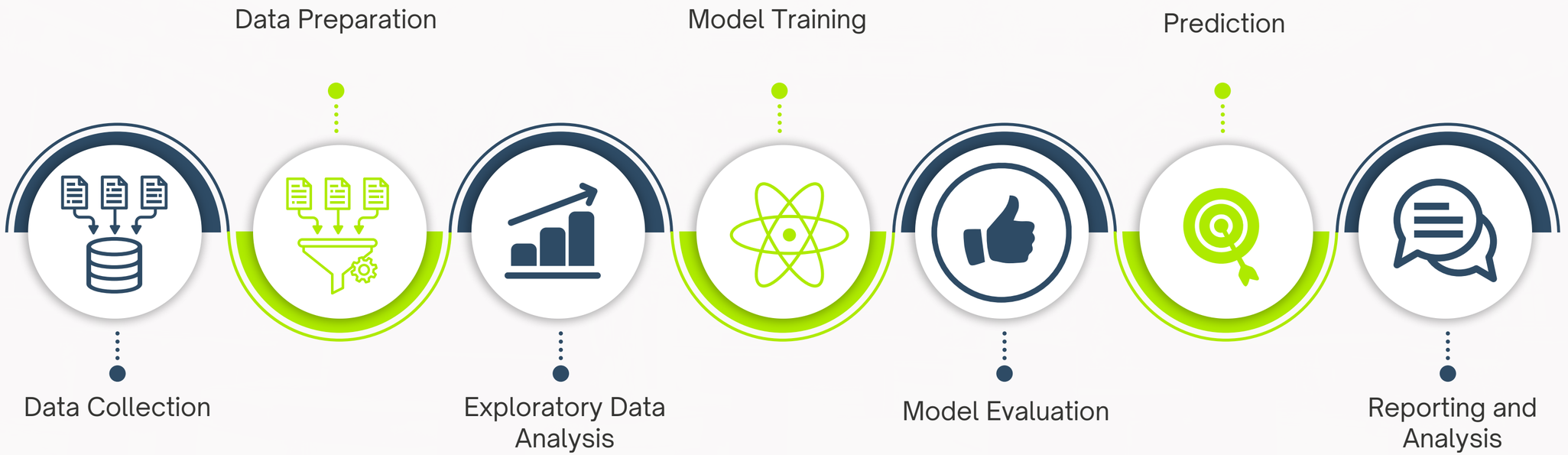
Objective

The primary goal of this project was to apply the best model using Machine Learning to predict real estate prices based on a set of features such as surface area, number of rooms, and geographical location.

This project aims to demonstrate how machine learning techniques can be utilized to derive insights that assist in making informed real estate investment decisions.

Methodology

The project followed a structured workflow



Data Collection

The dataset has the following columns:

- surface: The area of the property in square meters.
- rooms_new: The number of rooms in the property.
- zipcode_new: The zipcode of the property location.
- price_new: The price of the property.
- latitude: The latitude coordinate of the property.
- longitude: The longitude coordinate of the property.

Data Preparation

Data Formatting

The data appears to be in a single-column format, used delimiter (;) to segregate the data into proper columns.

Handling Missing Values

Handled missing values in rooms by filling them with the median value (since it's a count of rooms, median is more appropriate than mean).

Descriptive Statistics

Used descriptive statistics for data insights

Analysis:

Surface: The property's size varies significantly from 4.26 to 690 square metres, with an average of 108 square metres.

Rooms: The typical property has 3.66 rooms, the lowest being 0 and the most 9. The rooms column has 28 missing values, which were addressed by filling the missing values with the median.

Zipcode: Zipcodes range from 1011 to 1109.

Price: Prices vary from as low as 10,000 to as high as 6,400,000.

Latitude and Longitude: These show that the properties are mainly located at 52.07 latitude and 2.91 longitude.

Exploratory Data Analysis

Data Consistency Check

Ensured that all data adhered to expected formats and that there were no inconsistencies (Price)

EDA plots

1. Pairplot
2. Histogram plot for distribution of surface area
3. Histogram plot for distribution of rooms
4. Histogram plot for distribution of property prices
5. Boxplot for number of rooms by zipcode

Analysis:

The pairplot shows pairwise relationships between features.

The distribution of property prices is right-skewed, meaning there are more properties with prices on the lower end and fewer with very high prices.

There is a right skew in the distribution of surface area, with fewer properties having larger surface areas. This is typical for residential property data, where larger homes are less common.

There is a higher number of houses with 3 rooms.

The uniformity in the box plot could indicate a stable market targeting similar demographics across different zip codes.

Model Training

Train the data

We will train a KNN Regressor and a Multiple Regression model using these features.

After training, we'll evaluate and compare the models based on R^2 score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Analysis:

Here's how the two models performed:

For Linear Regression:

- **R^2 Score: 0.753 (75.3% variance explained)**
- **Mean Squared Error (MSE): 117,372,115,420.46**
- **Root Mean Squared Error (RMSE): 342,596.14**

For KNN:

- **R^2 Score: 0.825 (82.5% variance explained)**
- **Mean Squared Error (MSE): 83,353,815,170.56**
- **Root Mean Squared Error (RMSE): 288,710.61**

Model Evaluation

Analyze the data

In comparing the performance of Linear Regression and K-Nearest Neighbors (KNN) models for predicting real estate prices, the KNN model demonstrates superior accuracy, explaining 82.5% of the variance ($R^2 = 0.825$) with smaller prediction errors (MSE = 83,353,815,170.56 and RMSE = 288,710.61) than the Linear Regression model, which explains 75.3% of the variance ($R^2 = 0.753$) with higher errors (MSE = 117,372,115,420.46 and RMSE = 342,596.14).

Analysis:

The KNN Regressor model outperforms the Multiple Regression model across all metrics, explaining a higher percentage of the variance in prices and achieving lower errors.

Given these results, we will choose the KNN Regressor as the best model.

Prediction

Prediction of the model using new data

Overview:

The scaled new data is fed into the KNN model, which outputs predicted prices. The predictions are:

- 1,367,500 for a 200 sqm, 5-room property
- 666,900 for a 150 sqm, 4-room property
- 636,900 for a 100 sqm, 3-room property
- 370,700 for a 75 sqm, 2-room property
- 181,300 for a 50 sqm, 1-room property

Analysis:

Predicted prices indicate a trend: greater prices are expected for larger properties with more rooms. This fits nicely with the normal dynamics of the real estate market, where the number of rooms and size of the home are key factors in determining the price.

As the attributes get smaller, the prices likewise seem to be decreasing, indicating that the model is reacting to the input features rationally.

Conclusion

The analysis of the K-Nearest Neighbors (KNN) model applied to estate price prediction shows that it effectively utilizes key property features—such as surface area, number of rooms, and location—to estimate property prices.