

House Price Prediction Using Linear Regression

Internship Task: Artificial Intelligence & Machine Learning – Task 1
Student Name: Akhlesh Rajput
University: University of Hyderabad
Program: MCA
Internship Provider: Maincrafts Technology

INTRODUCTION

House price prediction is an important problem in the real estate industry. Accurate prediction of house prices helps buyers, sellers, and investors make better decisions. In this project, a Linear Regression model is developed to predict median house prices in California using historical housing data.

The objective of this task is to understand the complete machine learning workflow, including data loading, data exploration, preprocessing, model training, evaluation, and result interpretation. Python and the Scikit-learn library are used to implement the model.

DATASET DESCRIPTION

The California Housing dataset is a publicly available dataset provided by Scikit-learn. It contains information collected from the 1990 California census. Each row represents housing information for a district.

Target Variable:
Median House Value (MedHouseVal)

Input Features Include:

- Median Income
- House Age
- Average Rooms
- Average Bedrooms
- Population
- Average Occupancy
- Latitude

- Longitude

The dataset does not contain missing values, which makes it suitable for direct model training.

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis was performed to understand the structure and characteristics of the dataset.

- Dataset size and column information were examined.
- Statistical summary such as mean, minimum, and maximum values was analyzed.
- Missing value check showed no null values.
- A correlation heatmap was used to observe relationships between features and house price.

It was observed that median income has a strong positive correlation with house price, indicating it is an important feature.

DATA PREPROCESSING

The dataset was divided into:

- Features (X): All input columns
- Target (y): Median house value

The dataset was split into training and testing sets using an 80:20 ratio. Training data is used to train the model, while testing data is used to evaluate performance.

MODEL BUILDING

Linear Regression algorithm was chosen for this task because it is simple, fast, and widely used for regression problems.

The model learns the relationship between input features and the target variable by minimizing prediction errors.

The Linear Regression model was trained using the training dataset.

MODEL EVALUATION

Three evaluation metrics were used:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R² Score

Lower MAE and RMSE values indicate better prediction accuracy.
R² score shows how well the model explains the variance in house prices.

The model achieved good performance on the test dataset, showing that Linear Regression is suitable as a baseline model.

RESULTS AND VISUALIZATION

Two important plots were created:

- Actual vs Predicted Prices Plot
- Residual Plot

The Actual vs Predicted plot shows that most points are close to the diagonal line, indicating good predictions.

The Residual plot shows that errors are randomly distributed around zero, which suggests the model fits reasonably well.

CONCLUSION

In this project, a Linear Regression model was successfully implemented to predict house prices using the California Housing dataset. The model demonstrated satisfactory performance and provided useful insights into important features affecting house prices.

This project helped in understanding the complete machine learning pipeline.

FUTURE IMPROVEMENTS

The model performance can be improved by:

- Applying feature scaling
- Using regularization techniques such as Ridge or Lasso Regression
- Trying advanced models like Random Forest or Gradient Boosting
- Performing hyperparameter tuning

TOOLS & TECHNOLOGIES USED

- Python
- Pandas
- NumPy
- Scikit-learn
- Matplotlib
- Seaborn
- Jupyter Notebook