House Price Prediction using Linear Regression

Task 1 – AI & Machine Learning Internship

Name: Aryan Marota
Internship Organization: Maincraft Technologies
Internship Duration: 6 Weeks
Task Title: House Price Prediction using Linear Regression
Tools Used: Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

---

1. Introduction

House price prediction is a classic regression problem in the field of machine learning and data science. Accurate prediction of housing prices is important for real estate companies, buyers, sellers, and policy makers, as it helps in making informed financial decisions.

The objective of this task is to build and evaluate a **Linear Regression model** that predicts median house prices using the **California Housing Dataset**. The task involves performing exploratory data analysis (EDA), training a regression model, evaluating its performance using standard metrics, and interpreting the results.

This project demonstrates the complete machine learning workflow, starting from data loading and exploration to model evaluation and visualization.

---

2. Dataset Description

The dataset used in this project is the **California Housing Dataset**, which is available directly from the Scikit-learn library. The dataset is based on data collected from the **1990 California census** and represents aggregated information about housing districts in California.

Dataset Characteristics

- Total number of records: **20,640**
- Total number of features: **8**
- Target variable: **Median House Value (MedHouseVal)**
- All variables are numerical
- No missing values present

Input Features

- **MedInc:** Median income of households in a district
- **HouseAge:** Median age of houses
- **AveRooms:** Average number of rooms per household
- **AveBedrms:** Average number of bedrooms per household

- **Population:** Population of the district
- **AveOccup:** Average occupancy per household
- **Latitude:** Geographic latitude
- **Longitude:** Geographic longitude

Target Variable

- **MedHouseVal:** Median house value (in hundreds of thousands)

Initial inspection using df.head(), df.info(), and df.describe() confirmed that the dataset is clean and suitable for regression modeling without additional preprocessing.

---

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the distribution of variables and their relationship with the target variable.

3.1 Data Distribution

The distribution of the target variable **MedHouseVal** shows a slightly right-skewed pattern. Most house prices lie between lower and mid-range values, while fewer observations are present at higher price ranges. A visible upper cap is observed, which is a known characteristic of this dataset.

This distribution suggests that predicting extremely high-priced houses may be more challenging for a linear model.

3.2 Correlation Analysis

A correlation heatmap was used to examine relationships among features.

Key observations from the correlation heatmap:

- **Median Income (MedInc)** has the strongest positive correlation with house prices.
- **Average Rooms** also shows a moderate positive relationship with the target.
- **Latitude and Longitude** indicate that geographical location plays a role in determining house prices.
- Features such as population and average occupancy show weak correlation with the target variable.

These observations suggest that income and location-related features are the most influential factors in predicting house prices.

---

4. Model Building

A **Linear Regression** model was chosen for this task due to its simplicity and interpretability. Linear regression helps in understanding how each feature contributes to the prediction of house prices.

4.1 Train-Test Split

- The dataset was split into:
    - **80% training data**
    - **20% testing data**
- A fixed random state was used to ensure reproducibility of results.

4.2 Model Training

The model was trained using the training dataset, where it learned a linear relationship between the input features and the target variable.

After training, predictions were generated on the unseen test dataset.

---

5. Model Evaluation

The performance of the model was evaluated using standard regression metrics:

5.1 Evaluation Metrics

- **Mean Absolute Error (MAE):** 0.533
- **Root Mean Squared Error (RMSE):** 0.746
- **R^2 Score:** 0.576

5.2 Interpretation of Results

- The **MAE** value indicates that, on average, the predicted house price differs from the actual price by approximately 0.53 units.
- The **RMSE** value shows that larger errors are moderately penalized, which is expected due to price variability.
- The **R^2 score of 0.576** indicates that the model explains around **57.6% of the variance** in house prices.

These results suggest that the Linear Regression model performs reasonably well and provides a solid baseline for house price prediction.

---

6. Visualization of Results

6.1 Actual vs Predicted Prices

The scatter plot comparing actual and predicted house prices shows that most points lie close to the ideal diagonal line, indicating good prediction performance for mid-range values.

However, deviations are observed for higher-priced houses, suggesting that the linear model struggles to capture complex patterns at extreme values.

6.2 Residual Analysis

The residual plot displays the difference between actual and predicted values against predicted prices.

Observations:

- Residuals are mostly centered around zero.
- Some patterns and increasing variance are visible at higher predicted values.
- A few outliers are present, indicating cases where the model significantly under- or over-predicts prices.

This suggests that while the model is reasonable, it does not fully capture non-linear relationships in the data.

---

7. Conclusion

In this task, a Linear Regression model was successfully implemented to predict house prices using the California Housing Dataset. The project covered all essential steps of a machine learning pipeline, including data exploration, model training, evaluation, and interpretation.

The results show that Linear Regression provides a good baseline performance, with median income emerging as the most influential feature. However, the presence of residual patterns and moderate R^2 score indicates that more advanced models such as Ridge Regression, Lasso Regression, Decision Trees, or ensemble methods could further improve performance.

Overall, this task provided valuable hands-on experience in applying machine learning techniques to a real-world regression problem.

---

8. Future Scope

- Use regularized regression techniques (Ridge, Lasso)
- Experiment with non-linear models
- Perform feature scaling and feature engineering
- Compare multiple regression models