House Price Prediction using Linear Regression

Internship Task Report – Artificial Intelligence & Machine Learning

Student Name: Sumit Chhadiwal

Internship Program: AI & ML Internship (Maincrafts)

Task: Linear Regression Model – California Housing Dataset

Date: 14/02/2026

1. Introduction

Machine Learning plays a crucial role in solving real -world prediction and decision -making problems. One of the most fundamental supervised learning techniques is Linear Regression , which is used to model the relationship between a dependent variable and one or more independent variables.

This project focuses on building and evaluating a Linear Regression model to predict house prices using the California Housing dataset . The objective of this task is to understand and implement the complete machine learning workflow including:

• Data loading

• Exploratory Data Analysis (EDA)

• Data preprocessing

• Feature selection

• Model training

• Model evaluation

• Result interpretation and reporting

This project serves as a foundational step toward more advanced machine learning models and provides valuable practical experience for internship and academic purposes.

2. Dataset Description

The California Housing dataset is a publicly available dataset provided by the scikit-learn library. It contains housing -related information collected from the 1990 California census.

Dataset Characteristics:

• Total records: 20,640

• Total features: 8 input features + 1 target variable

• Dataset type: Numerical

Input Features:

Feature Name Description

MedInc Median income in block group

HouseAge Median house age

AveRooms Average number of rooms

AveBedrms Average number of bedrooms

Population Block group population

AveOccup Average house occupancy

Latitude Latitude coordinate

Longitude Longitude coordinate

Target Variable:

• MedHouseVal – Median house value (in $100,000 units)

This dataset is well -suited for regression analysis due to its size, numerical features, and real-world relevance.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the structure, quality, and relationships within the dataset.

3.1 Data Inspection

• The dataset was loaded using fetch_california_housing() from scikit -learn.

• The info() function was used to verify data types and memory usage.

• The describe() function was used to analyze statistical properties such as mean, standard deviation, minimum, and maximum values.

3.2 Missing Values Analysis

The dataset was checked for missing values using:

df.isnull().sum()

Result:

• No missing values were found in any feature.

• Therefore, no data imputation was required.

3.3 Correlation Analysis

A correlation heatmap was generated to understand relationships between features and the target variable.

Key observations:

• Median Income (MedInc) showed the strongest positive correlation with house prices.

• Location features (Latitude and Longitude) also influenced house prices.

• Some features showed weak correlation, indicating limited predictive power.

EDA helped confirm that the dataset is clean and suitable for model training.

4. Feature Selection and Data Splitting

All eight input features were selected for training the model, while MedHouseVal was used as the target variable.

The dataset was split into:

• Training set: 80%

• Testing set: 20%

Using:

train_test_split(X, y, test_size=0.2, random_state=42)

This ensures the model is evaluated on unseen data to measure generalization performance.

5. Model Building

A Linear Regression model was implemented using

sklearn.linear_model.LinearRegression .

Steps followed:

1. Initialize the model

2. Fit the model using training data

3. Generate predictions on the test dataset

Linear Regression attempts to learn the best -fitting linear equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Where:

• y = predicted house value

• x = input features

• β = learned coefficients

6. Model Evaluation

The trained model was evaluated using standard regression metrics:

6.1 Evaluation Metrics

Metric Description Value

MAE Mean Absolute Error 0.5332001304956556

RMSE Root Mean Squared Error 0.7455813830127763

R² Coefficient of Determination 0.575787706032451

Metric Explanation:

• MAE measures the average absolute difference between actual and predicted

values.

• RMSE penalizes larger errors more strongly and indicates prediction accuracy.

• R² Score shows how much variance in the target variable is explained by the

model.

6.2 Visualization

Two plots were generated:

1. Actual vs Predicted Scatter Plot – to visualize prediction accuracy

2. Residual Plot – to analyze error distribution

The residuals were mostly centered around zero, indicating acceptable model

performance.

7. Results and Discussion

The Linear Regression model successfully learned general trends in the housing data.

The predictions were reasonably accurate for many samples; however, some deviation

exists due to:

• Non-linear relationships in the data

- Feature interactions not captured by linear models

- Presence of outliers

- Limited model complexity

Despite these limitations, the model provides a strong baseline solution and demonstrates the effectiveness of basic regression techniques.

8. Future Improvements

The following improvements can significantly enhance performance:

- Feature scaling using StandardScaler or MinMaxScaler

- Removing or treating outliers

- Polynomial feature transformation

- Regularization (Ridge and Lasso regression)

- Advanced algorithms such as:

o Random Forest Regressor

o Gradient Boosting Regressor

o XGBoost

- Hyperparameter tuning using GridSearchCV or RandomizedSearchCV

- Cross-validation for more reliable evaluation

9. Conclusion

This project successfully demonstrates the complete machine learning workflow for a regression problem. It covers data exploration, preprocessing, model training, evaluation, and result interpretation using industry -standard tools such as Python and scikit-learn.

The project strengthens foundational understanding of supervised learning and provides practical experience essential for future data science and machine learning roles. It also serves as a strong portfolio project for internship evaluation.

10. Tools and Technologies Used

- Python

- Pandas

- NumPy

- Scikit-learn

- Matplotlib

- Seaborn

- Jupyter Notebook

End of Report