
California Housing Price Prediction – Linear Regression Report

1. Introduction

The goal of this project is to build and evaluate a **Linear Regression model** to predict median house values in California districts using the **California Housing dataset**.

The dataset contains demographic, geographic, and housing-related features from the 1990 U.S. Census.

---

2. Dataset Overview

- **Source:** California Housing dataset (Scikit-learn)
- **Records:** 20,640
- **Features:** 8 numerical predictors + target (MedHouseVal)
- **Key Features:**
- MedInc – Median income in block group
- HouseAge – Median house age
- AveRooms – Average rooms per household
- AveBedrms – Average bedrooms per household
- Population – Block group population
- AveOccup – Average household size
- Latitude, Longitude – Geographic coordinates
- **Target:** Median house value (in $100,000s)

---

3. Exploratory Data Analysis (EDA) Summary

3.1 Data Quality

- No missing values in dataset.
- Features have varying scales; MedInc and AveRooms show skewness.

3.2 Key Insights

- **Strongest correlation:** MedInc (0.69) with target.
- **Negative correlation:** Longitude (-0.05) and Latitude (-0.14) indicate location impact.
- **Outliers:** High-income districts with extreme house values.
- **Geospatial patterns:** Coastal areas tend to have higher prices.

3.3 Visual Highlights

- Scatter plot of MedInc vs. MedHouseVal shows a clear positive trend.
- Heatmap confirms MedInc as the most predictive feature.

---

4. Model Development

4.1 Preprocessing

- Standardized features using StandardScaler.
- No categorical encoding needed (all features numeric).
- Split: 80% training, 20% testing.

4.2 Model

- **Algorithm:** Ordinary Least Squares Linear Regression.
- **Implementation:** sklearn.linear_model.LinearRegression.

---

5. Model Evaluation

Metric	Train Set	Test Set
R^2 Score	0.606	0.602
MAE ($100k)	0.53	0.54
RMSE ($100k)	0.74	0.75

Interpretation:

- The model explains ~60% of variance in house prices.
- Average prediction error is about $54,000.
- Performance is consistent across train and test sets, indicating low overfitting.

---

6. Limitations

- **Linear assumption:** Cannot capture complex non-linear relationships.
- **Geospatial complexity:** Latitude/Longitude effects are not fully modeled.
- **Feature interactions:** Not explicitly included.

---

7. Improvement Ideas

1. **Feature Engineering**

- Create interaction terms (e.g., MedInc × Latitude).

- Add polynomial features for non-linear patterns.

2. **Geospatial Modeling**

- Incorporate distance to coast or urban centers.
- Use clustering to capture neighborhood effects.

3. **Model Upgrade**

- Try **Ridge/Lasso Regression** for regularization.
- Explore **Tree-based models** (Random Forest, Gradient Boosting).

4. **Data Enrichment**

- Add economic indicators, crime rates, or school quality scores.

---

8. Conclusion

The Linear Regression model provides a solid baseline for predicting California housing prices, achieving an R^2 of ~0.60. While it captures key trends, especially the strong influence of median income, further improvements in feature engineering and model complexity could significantly enhance predictive accuracy.
