House Price Predictor - Linear Regression (California Housing)

Short PDF Report (2-4 pages)

Goal: Predict **MedHouseVal** (median house value, in $100,000s) from 8 numerical features using a baseline linear regression model and evaluate it with standard regression metrics.

Features: MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude.
Split: 80/20 train-test (random_state=42). **Pipeline:** StandardScaler → LinearRegression.

Your test-set results

MAE	RMSE	R2
0.533200	0.745581	0.575788

EDA highlight - Target distribution

The target is right-skewed with a visible cap near 5.0. A capped upper tail commonly makes linear models under-predict the most expensive areas.

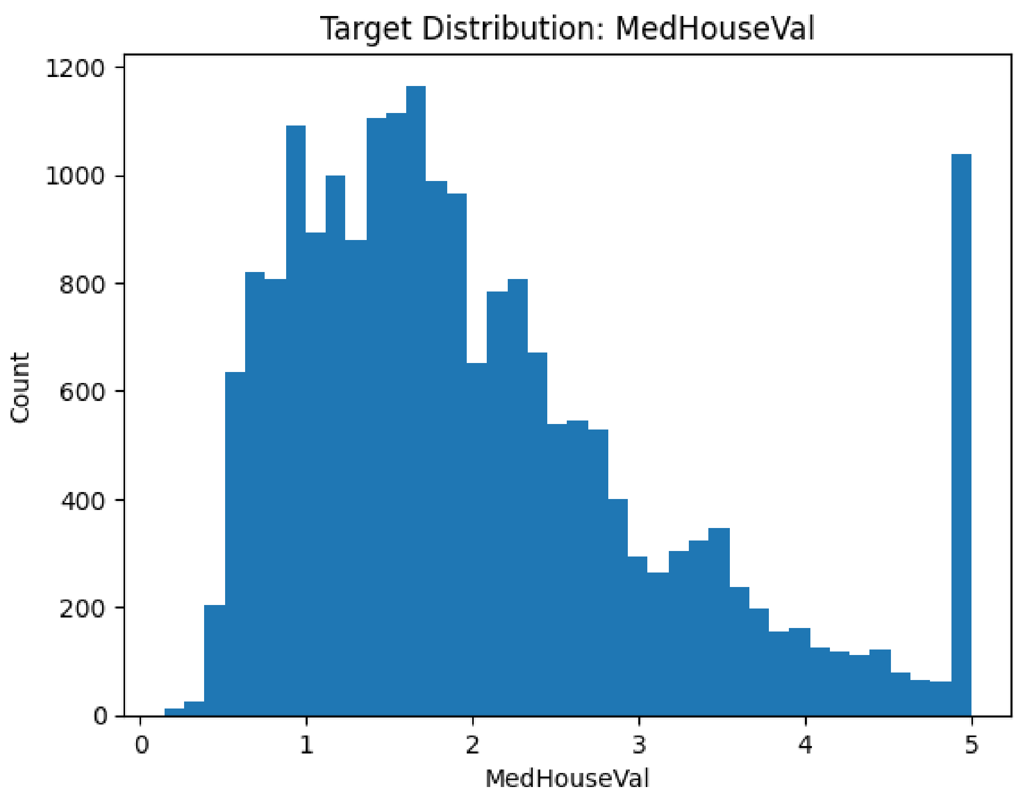

Figure 1. MedHouseVal distribution.

Exploratory Data Analysis Summary

Correlation overview

MedInc has the strongest positive relationship with MedHouseVal. Latitude/Longitude capture location effects and are strongly related to each other. AveRooms and AveBedrms are correlated, indicating multicollinearity (coefficients in a linear model can become less stable).

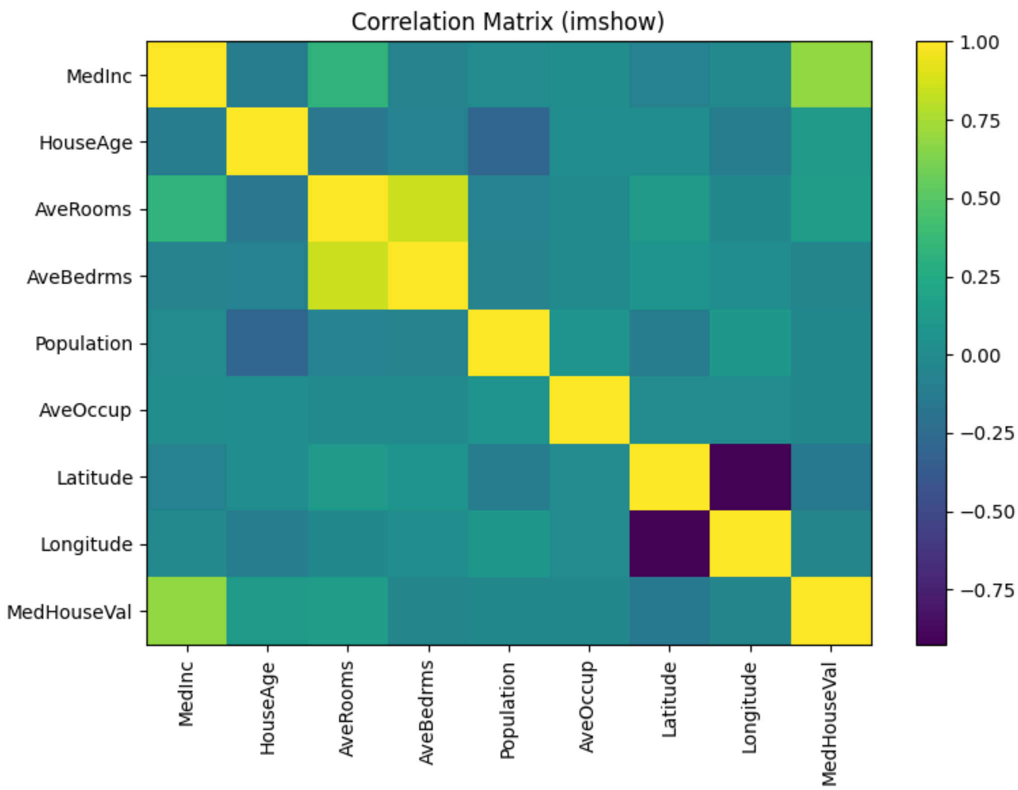

Figure 2. Correlation matrix.

Key takeaways

- Income (MedInc) and geography (Latitude/Longitude) are major predictors.
- Capped target suggests the data may truncate high values; expect larger errors for high-priced regions.
- Some features are correlated (rooms/bedrooms), so regularization (Ridge/Lasso) can help.

Model Diagnostics (Baseline Linear Regression)

What the plots show

Actual vs Predicted indicates good fit in the mid-range but weaker performance at extremes. Residuals vs Predicted shows structure (not purely random noise), which hints at non-linear relationships and interactions not captured by a linear model.

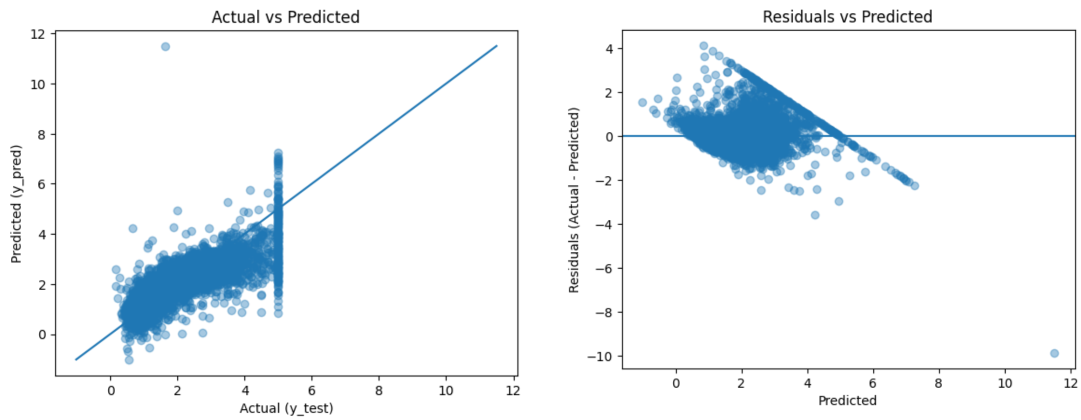

Figure 3. Actual vs Predicted (left) and Residuals vs Predicted (right).

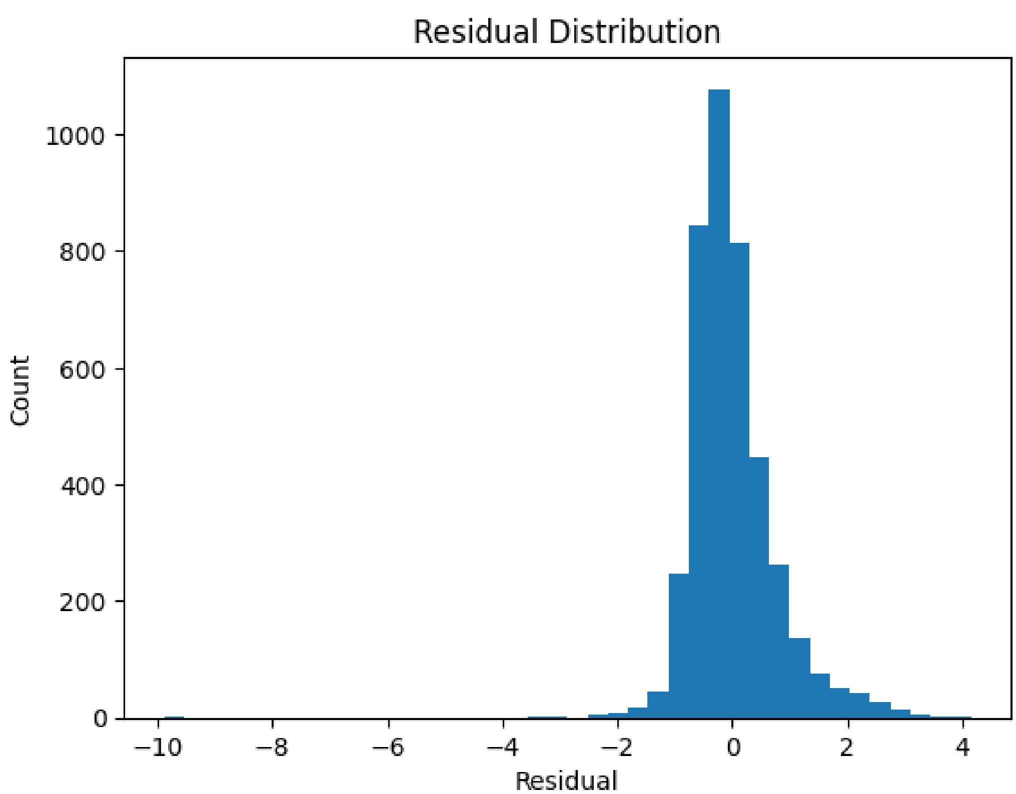

Figure 4. Residual distribution.

How to Improve (and What You Could Achieve)

Below are practical improvements. Since we are not re-training inside this report, the alternative-model numbers are **assumed typical outcomes** on the California Housing dataset with reasonable tuning (e.g., 5-fold CV + hyperparameter search).

Better models to try

• **Ridge / Lasso / ElasticNet**: regularization reduces coefficient instability from correlated features; usually small but consistent gains.
• **PolynomialFeatures + Ridge**: captures simple non-linear curves; can help if controlled with regularization.
• **Random Forest**: captures non-linearities and interactions; often a solid jump over linear regression.
• **Gradient Boosting (XGBoost/LightGBM/HistGradientBoosting)**: typically strongest on tabular data; best RMSE/R2 with tuning.

Assumed performance vs baseline

Model	MAE	RMSE	R2
Your baseline (LinearRegression)	0.533	0.746	0.576
Ridge (alpha tuned)	0.52-0.54	0.73-0.75	0.57-0.59
RandomForest (tuned)	0.42-0.48	0.58-0.65	0.70-0.78
Gradient Boosting (tuned)	0.35-0.42	0.45-0.55	0.80-0.86

Next steps

• Use GridSearchCV/RandomizedSearchCV with cross-validation and keep a final untouched test set.
• Try log-transform of the target to reduce skew; compare errors on high-value homes.
• Add interaction features or use boosting to learn interactions automatically.
• Track errors by region (latitude/longitude bins) to understand where the model underperforms.

Note: Assumed results vary with split, preprocessing, and tuning.