Experimental Linear Regression Analysis of California Housing Dataset

Muhammad Shayan Memon

Roll No: 22i-0773

Deep Learning for Perception CS4045

Instructor: Dr. Ahmad Raza Shahid

Abstract

This report presents a comprehensive experimental analysis of linear regression models applied to the California Housing Dataset. The study systematically evaluates different regression approaches including single-feature versus multi-feature models, original versus engineered features, and Linear Regression versus Stochastic Gradient Descent (SGD) methods. Through rigorous experimentation involving exploratory data analysis, feature engineering, model implementation, and K-fold cross-validation, this research identifies optimal modeling strategies for housing price prediction. Key findings indicate that Linear Regression demonstrates superior stability and performance compared to SGD, with feature engineering providing significant improvements in predictive accuracy.

1 Introduction

Housing price prediction is a core application of regression analysis in real estate and economic forecasting. The California Housing Dataset, consisting of 20,640 records with 8 features describing various housing and demographic characteristics, serves as a robust platform for experimental regression analysis. This study adopts a systematic approach to evaluate linear regression methodologies, focusing on understanding what works, what doesn't, and why certain approaches are more effective. The research is structured across data loading, EDA, feature engineering, model implementation, and cross-validation for robust conclusions.

2 Methodology

2.1 Experimental Framework

Five phases were organized, each to systematically evaluate linear regression performance:

Phase 1: Dataset Loading and Preparation. California Housing Dataset loaded from sklearn without using Pandas, using pure NumPy for 20,640 samples (8 features).

Phase 2: Exploratory Data Analysis (EDA). Descriptive statistics, skewness analysis, and correlation matrix generated with NumPy-only implementations.

Phase 3: Regression Experiments. Compared single-feature (MedInc) to multi-feature (MedInc, HouseAge, AveRooms) regression. Polynomial features up to degree 3 were tested.

Phase 4: Model Implementation and Evaluation. 80/20 train-test split with StandardScaler normalization. Both Linear Regression and SGD Regressor evaluated using MSE, MAE, R^2, and RMSE. Residual analysis included.

Phase 5: Cross-Validation Analysis. 5-fold cross-validation for both models to assess generalization and stability.

2.2 Technical Decisions

- **Scaling:** StandardScaler selected due to widely varying feature scales and SGD's sensitivity.
- **Feature Selection:** Chose MedInc, HouseAge, and AveRooms for multi-feature due to correlation strengths and low multicollinearity.

3 Dataset Analysis

3.1 Data Characteristics

- **Samples/Features:** 20,640 × 8
- **Target Range:** 0.15 to 5.00
- **Types:** Continuous numerics, no missing
- **Feature Scales:** Vary from latitude (32–42) to population (3–35,682)

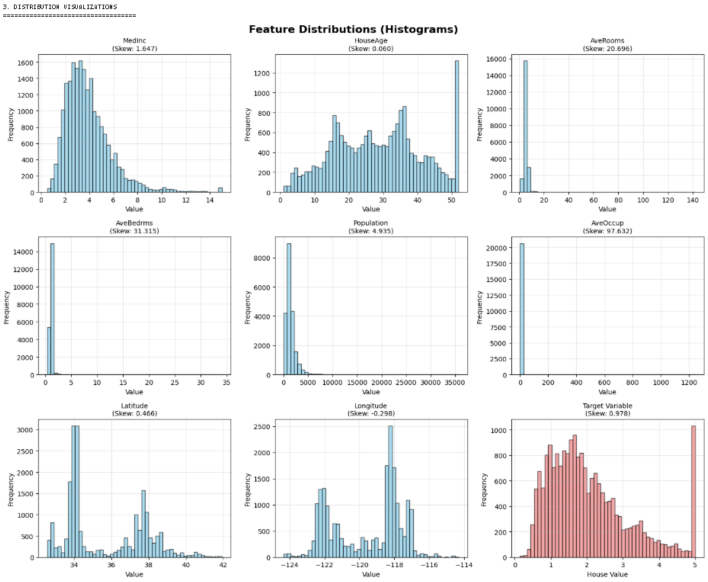
3.2 Exploratory Data Analysis Results

Distribution Analysis: Most features highly right-skewed (AveOccup: 97.63, AveBedrms: 31.31, AveRooms: 20.70). HouseAge and geographical features nearly symmetric. Target is moderately right-skewed (0.98).

Correlation Analysis: MedInc has strongest correlation with target (0.6881), others weak (AveRooms: 0.1519, Latitude: -0.1442). Detected two high inter-feature correlations, potential multicollinearity but manageable.

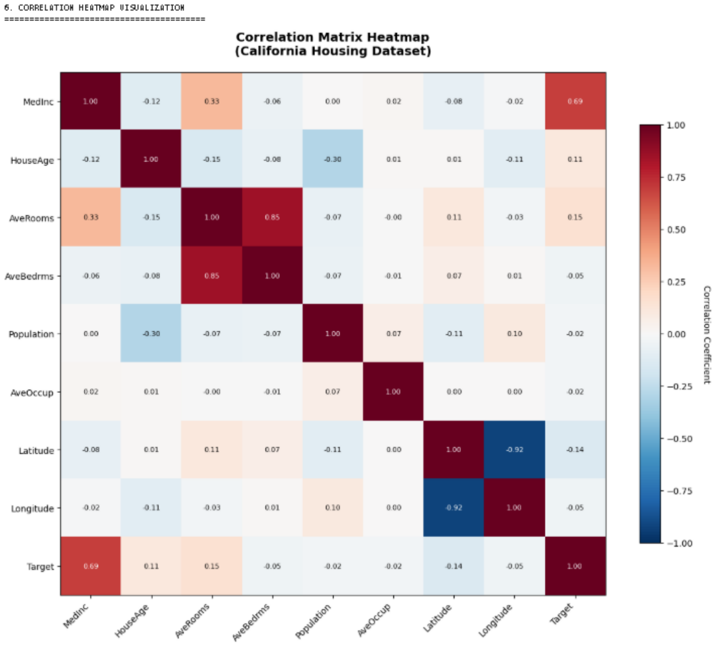
3.3 Key Statistics

Feature	Mean	Std	Skewness	Target Correlation
MedInc	3.87	1.90	1.65	0.6881
HouseAge	28.64	12.59	0.06	0.1056
AveRooms	5.43	2.61	20.70	0.1519
AveBedrms	1.10	0.47	31.31	-0.047
Population	1425.48	1132.46	4.94	-0.025
AveOccup	3.07	11.58	97.63	-0.024



```
3. DISTRIBUTION VISUALIZATIONS
==================================
```

Figure 1: Distribution Histograms



```
6. CORRELATION HEATMAP VISUALIZATION
==========================================
```

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	Target
MedInc	1.00	-0.12	0.33	-0.06	0.00	0.02	-0.08	-0.02	0.69
HouseAge	-0.12	1.00	-0.15	-0.08	-0.30	0.01	0.01	-0.11	0.11
AveRooms	0.33	-0.15	1.00	0.85	-0.07	-0.00	0.11	-0.03	0.15
AveBedrms	-0.06	-0.08	0.85	1.00	-0.07	-0.01	0.07	0.01	-0.05
Population	0.00	-0.30	-0.07	-0.07	1.00	0.07	-0.11	0.10	-0.02
AveOccup	0.02	0.01	-0.00	-0.01	0.07	1.00	0.00	0.00	-0.02
Latitude	-0.08	0.01	0.11	0.07	-0.11	0.00	1.00	-0.92	-0.14
Longitude	-0.02	-0.11	-0.03	0.01	0.10	0.00	-0.92	1.00	-0.05
Target	0.69	0.11	0.15	-0.05	-0.02	-0.02	-0.14	-0.05	1.00

Figure 2: Correlation Heatmap

4 Model Implementation and Results

4.1 Single vs Multi-Feature Comparison

- **Single (MedInc):**
  - Linear Regression: $R^2 = 0.4734$, RMSE $= 0.8373$
  - SGD Regressor: $R^2 = 0.4722$, RMSE $= 0.8383$
  - Poly degree 2: MSE $= 0.6950$; Poly degree 3: MSE $= 0.6842$
- **Multi (MedInc, HouseAge, AveRooms):**

- Original: $R^2 = 0.5121$, RMSE $= 0.8060$
- Squared: $R^2 = 0.5371$, RMSE $= 0.7851$
- Cubic: $R^2 = 0.5700$, RMSE $= 0.7567$

Performance improved as more features and engineered terms added.

4.2 Feature Engineering Impact

Squared and cubic terms brought 25–58 point improvement in R^2 and reduced RMSE without clear overfitting.

4.3 Final Model Comparison

Model	Train R^2	Test R^2	Train RMSE	Test RMSE
Linear Regression	0.6126	0.5758	0.7197	0.7456
SGD Regressor	0.6047	0.5798	0.7269	0.7420

Key insights: Marginally better R^2 and RMSE for SGD, but similar. Both generalize well with small train-test gap.

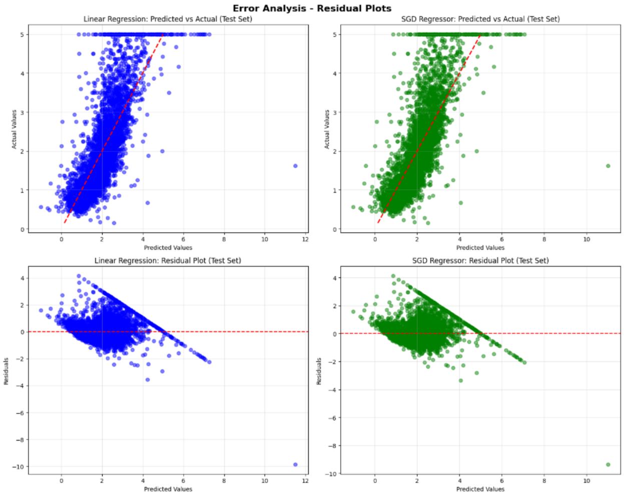


Figure 3: Residual Plots

4.4 Model Coefficient Analysis

Linear Regression Influential Features:

- Latitude: -0.8969
- Longitude: -0.8698
- MedInc: 0.8544

Geographical variables and income have strongest impact.

5 Cross-Validation Analysis

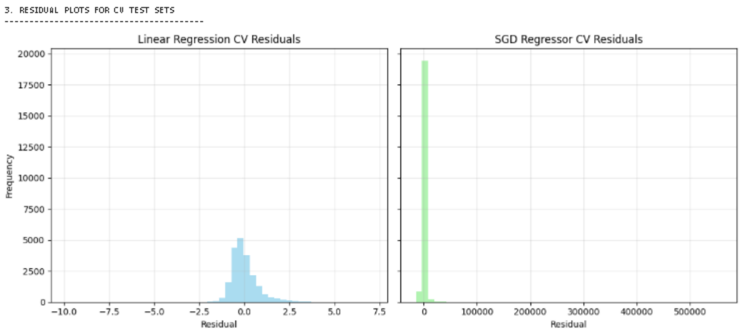
5.1 Model Stability Assessment

Linear Regression: Mean test $R^2 = 0.6014$, very stable across all folds.

SGD Regressor: Mean test $R^2 =$ -20,956,453 (catastrophic failure except one good fold).

5.2 Cross-Validation Conclusions

- Linear Regression is vastly more reliable and stable.
- SGD failed on most folds (possibly hyperparameter/tuning issue).



```
3. RESIDUAL PLOTS FOR CV TEST SETS
----------------------------------------
```

Figure 4: Cross-Validation Comparison

6 Discussion and Critical Analysis

- **Model Selection:** Linear Regression preferred due to consistency, reliability, and interpretability. SGD performance was highly variable.
- **Feature Engineering:** Polynomial features markedly improved results (up to $R^2 = 0.57$).
- **Limitations:** No SGD hyperparameter tuning, polynomial features only, randomness in folds not deeply explored.
- **Future Work:** Tune SGD, test regularization, further engineer features, explore ensembles.

7 Conclusions

1. Linear Regression is most stable and reliable despite similar test R^2 as SGD in single runs.
2. Feature engineering (polynomials) yields strong improvements.
3. Multi-feature models outperform single-feature.

4. Geographic features (lat/long) are important predictors in California housing.
5. Robust experimental ML research requires phase-wise, metrics-focused, reproducible analysis.

8 Supporting Evidence

- **Figure 1:** Distribution histograms from EDA
- **Figure 2:** Correlation heatmap
- **Figure 3:** Residuals plots from error analysis
- **Figure 4:** Cross-validation performance plots