

Regression Benchmark Report — California Housing

1. Introduction

This report summarizes the performance of seven machine learning regression models trained to predict median house value (MedHouseVal) using eight standardized socioeconomic and geographic features from the California Housing dataset. The objective is to build a reproducible pipeline for model comparison and understand how different algorithms perform on non-linear tabular data.

2. Goal

Predict median house values in California districts using a set of housing, income, demographic, and geographic predictors.

The target variable (MedHouseVal) is measured in units of \$100,000.

3. Models Compared

- Linear Regression
- Ridge Regression
- Lasso Regression
- Support Vector Regression (RBF Kernel)
- Random Forest Regressor
- Gradient Boosting Regressor
- Multi-Layer Perceptron (MLP) Regressor

These models span linear, regularized, kernel-based, ensemble, and neural network approaches.

4. Evaluation Metrics

4.1. MAE — Mean Absolute Error

Represents the average magnitude of errors in predictions.

- Same units as the target value (\$100k).
- Example: MAE = 0.33 corresponds to roughly \$33,000 average prediction error.

Lower values indicate better performance.

4.2. RMSE — Root Mean Squared Error

Penalizes larger errors more heavily.

- Often larger than MAE.
- Indicates how large typical prediction errors are.

Lower values indicate better performance.

4.3. R² — Coefficient of Determination

Measures how much variance in the target variable is explained by the model.

- $R^2 = 1.0$ indicates perfect prediction.
- $R^2 = 0.0$ indicates no predictive power.
- On this dataset, R^2 values above 0.75 are generally considered strong.

Higher values indicate better performance.

5. Results (Test Set)

Model	MAE	RMSE	R ²
Random Forest	0.326	0.503	0.807
MLP Regressor	0.343	0.513	0.799
Gradient Boosting	0.372	0.542	0.776
SVR (RBF)	0.377	0.569	0.753
Lasso	0.533	0.745	0.577
Ridge	0.533	0.746	0.576
Linear Regression	0.533	0.746	0.576

6. Ranking by R² (Best to Worst)

1. Random Forest — R² = 0.807
2. MLP Regressor — R² = 0.799
3. Gradient Boosting — R² = 0.776
4. SVR (RBF Kernel) — R² = 0.753
5. Lasso Regression — R² = 0.577
6. Ridge Regression — R² = 0.576
7. Linear Regression — R² = 0.576

7. Interpretation of Results

The California Housing dataset contains several **non-linear relationships** and **feature interactions**, including:

- The effect of median income varies strongly by location.
- Room-to-bedroom ratios matter in complex, non-linear ways.
- Latitude and longitude create geographic clusters that influence prices.
- Population and housing density interact with income and age features.

Because of this:

7.1. Ensemble Tree Models

Random Forest and Gradient Boosting perform strongly because they naturally capture non-linearities and interactions without requiring feature engineering.

Random Forest achieved the best performance overall with R² = 0.807.

7.2. Neural Network (MLP Regressor)

Also learns non-linear behavior and performed nearly as well as Random Forest.

More sensitive to hyperparameters and training stability.

7.3. SVR (RBF Kernel)

Non-linear kernel helps, but computational cost increases with dataset size.

Moderate performance compared to ensembles.

7.4. Linear Models

Linear, Ridge, and Lasso models assume linear relationships.

Their performance plateaued around R² ≈ 0.576 because they cannot capture complex

interactions.

8. Key Takeaways

- **Random Forest is the top-performing model**, achieving the highest R^2 and lowest errors.
- Compared to linear models, Random Forest reduced average error from approximately \$53,000 to \$32,600.
- Ensemble tree models are generally strong default choices for real estate and other tabular datasets with non-linear patterns.
- The experiment demonstrates practical skills in data preprocessing, modeling, evaluation, and reproducibility.

9. Files and Artifacts

- Metrics file:
`reports/results/regression_results.json`
- Feature importance plots and coefficient plots:
`reports/figures/feature_importance_*.png`
- Script for regenerating this report:
`script/scriptmake_report.py`

10. Conclusion

This benchmark highlights the importance of model selection in regression tasks involving non-linear tabular data.

Tree-based ensemble methods and neural networks outperform linear models significantly, demonstrating the value of flexible learning algorithms when feature interactions are present.

The results and artifacts provide a strong foundation for demonstrating machine learning engineering and model evaluation skills in internship applications.