Improving Predictions: Impact of Outliers

Achille Nguessie

November 05, 2024

Abstract

This report presents a comprehensive analysis and modeling of the California Housing dataset to predict median house values. The study involves exploratory data analysis, regression modeling, handling of outliers and multicollinearity, and the application of advanced machine learning algorithms. The final model achieves a significant improvement in predictive accuracy, demonstrating the efficacy of non-linear approaches over traditional regression models. For a detailed walkthrough of the project, including the code and data, please visit the GitHub page. You can also explore my personal work and other projects on my personal website.

Contents

1 Introduction **3**

2 Data Exploration and Preprocessing **3**

3 Regression Modeling **6**

3.1 Baseline Linear Regression Model 6

3.2 Residuals Analysis 6

4 Model Enhancement **7**

5 Non-linear Machine Learning Models **9**

6 Conclusion **11**

1 Introduction

The California Housing dataset is a widely used benchmark in machine learning for regression tasks. It contains data collected from the 1990 U.S. Census, encompassing various attributes that influence housing prices in California. This report aims to build predictive models to estimate the median house value based on the available features, improving model performance through data preprocessing and advanced algorithms.

2 Data Exploration and Preprocessing

The initial step involves loading the dataset and conducting a thorough exploratory data analysis (EDA) to understand the data's structure and underlying patterns.

A statistical summary provides insights into the central tendencies and dispersion of the features. The dataset includes the following features:

- **MedInc**: Median income in block group
- **HouseAge**: Median house age in block group
- **AveRooms**: Average number of rooms per household
- **AveBedrms**: Average number of bedrooms per household
- **Population**: Block group population
- **AveOccup**: Average number of household members
- **Latitude**: Block group latitude
- **Longitude**: Block group longitude

Figure 1 displays the histograms of the features, revealing that many variables are heavily right-skewed.

Statistical Summary of Features

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
count	20640.00	20640.00	20640.00	20640.00	20640.00	20640.00	20640.00	20640.00
mean	3.87	28.64	5.43	1.09	1425.48	3.07	35.63	-119.57
std	1.90	12.59	2.47	0.35	1132.46	10.39	2.14	2.00
min	0.50	1.00	0.85	0.33	3.00	0.69	32.54	-124.35
25%	2.56	18.00	4.44	1.01	787.00	2.43	33.93	-121.80
50%	3.54	29.00	5.23	1.05	1166.00	2.82	34.26	-118.49
75%	4.74	37.00	6.05	1.10	1725.00	3.28	37.71	-118.01
max	15.00	52.00	141.91	34.07	35682.00	1243.33	41.95	-114.31

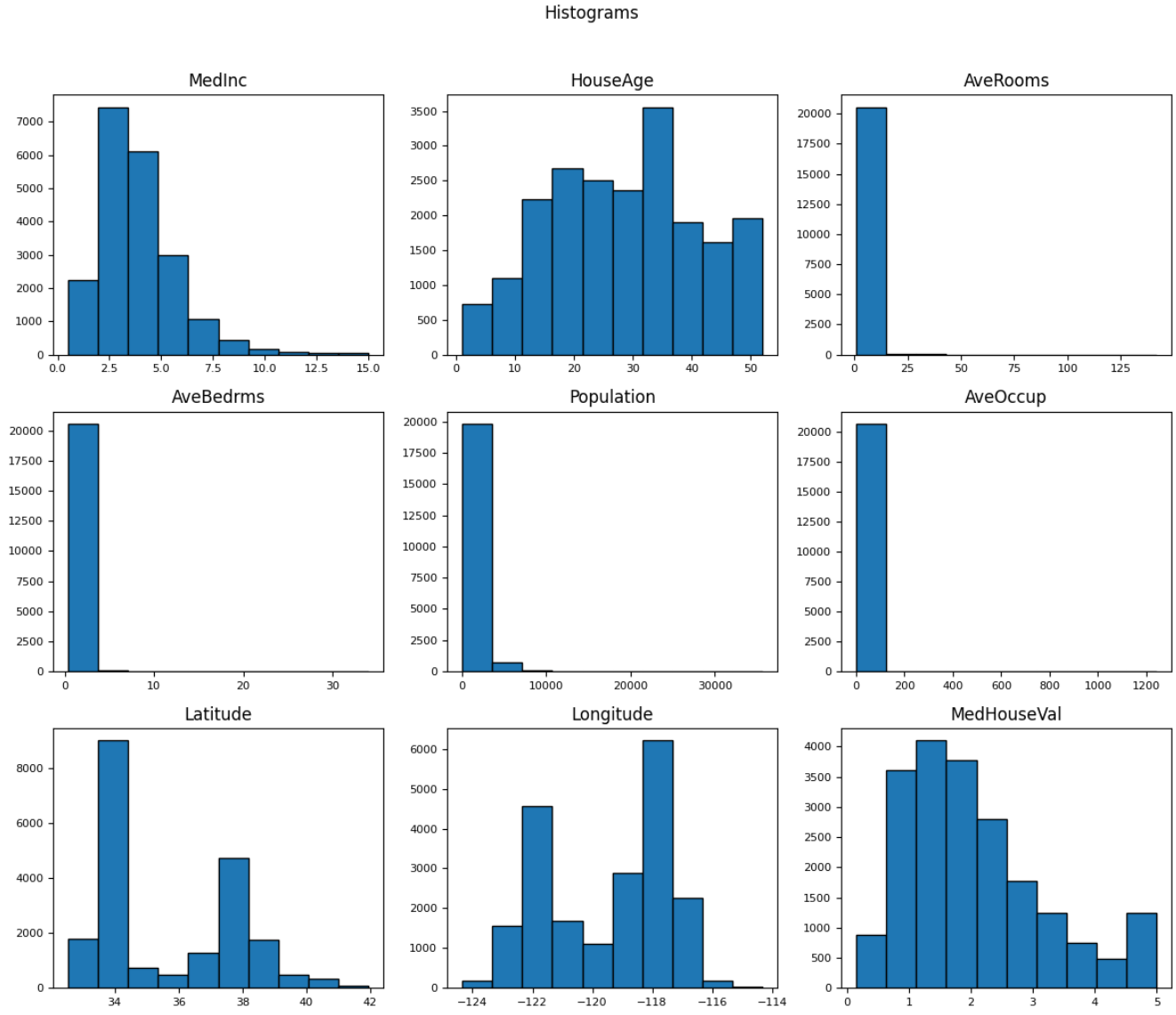

Figure 1: Histograms of Features

The correlation matrix in Figure 2 illustrates the linear relationships between pairs of variables.

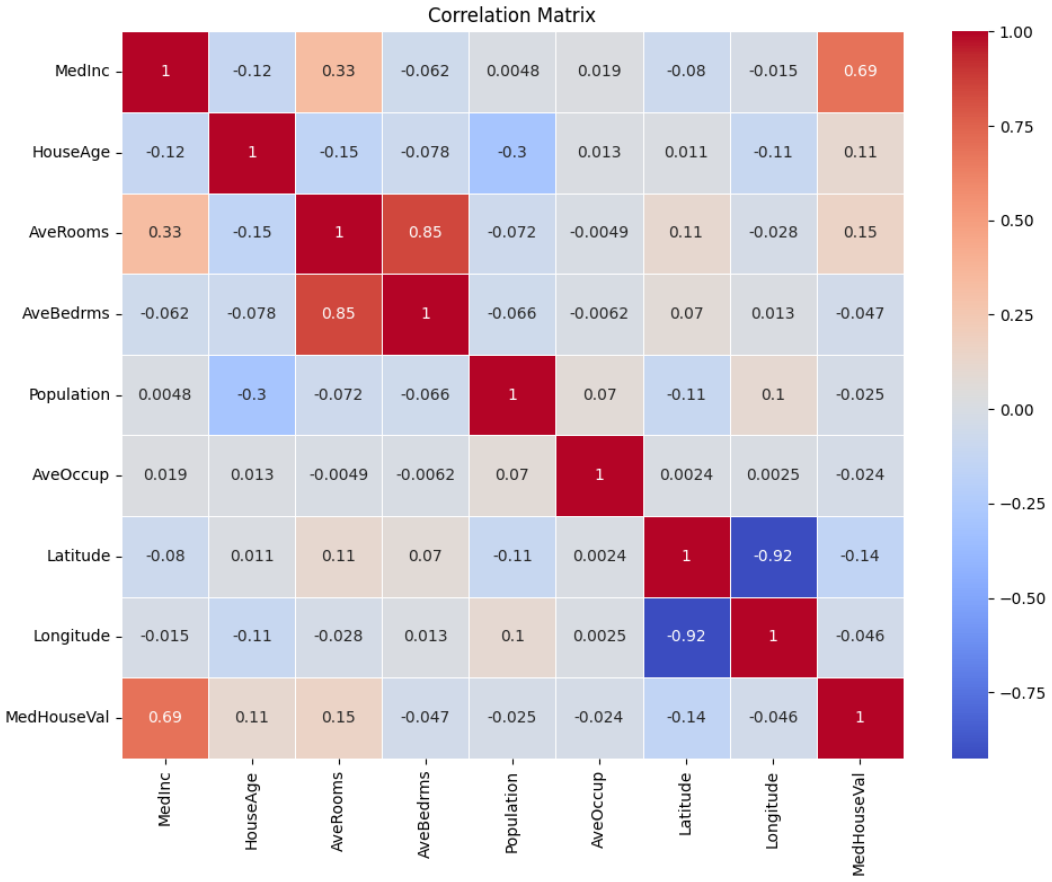

Figure 2: Correlation Matrix

Observations from EDA

- **High Correlation**: *AveRooms* and *AveBedrms* exhibit a strong positive correlation, suggesting multicollinearity.
- **Median Income**: Positively correlated with *AveRooms*, indicating wealthier areas have larger houses.
- **Weak Correlations**: Features like *Population* show weak correlations with *MedHouseVal*.
- **Skewness**: Many features are right-skewed, necessitating data transformation or outlier handling.
- The relationships captured by *Latitude* and *Longitude* provide evidence of spatial dependency, which may be valuable for geographic or location-based analyses.

3 Regression Modeling

A linear regression model is initially employed to establish a baseline performance. The dataset is split into training and testing sets in a 70:30 ratio. Standardization is applied to features to normalize the data.

3.1 Baseline Linear Regression Model

A linear regression model is trained using the standardized training data.

Table 1: Baseline Linear Regression Model Performance

Dataset	**MSE**	R^2
Training Set	0.523	0.606
Testing Set	0.531	0.596

3.2 Residuals Analysis

Figure 3 illustrates the residuals of the model, highlighting potential issues with heteroscedasticity and non-normality.

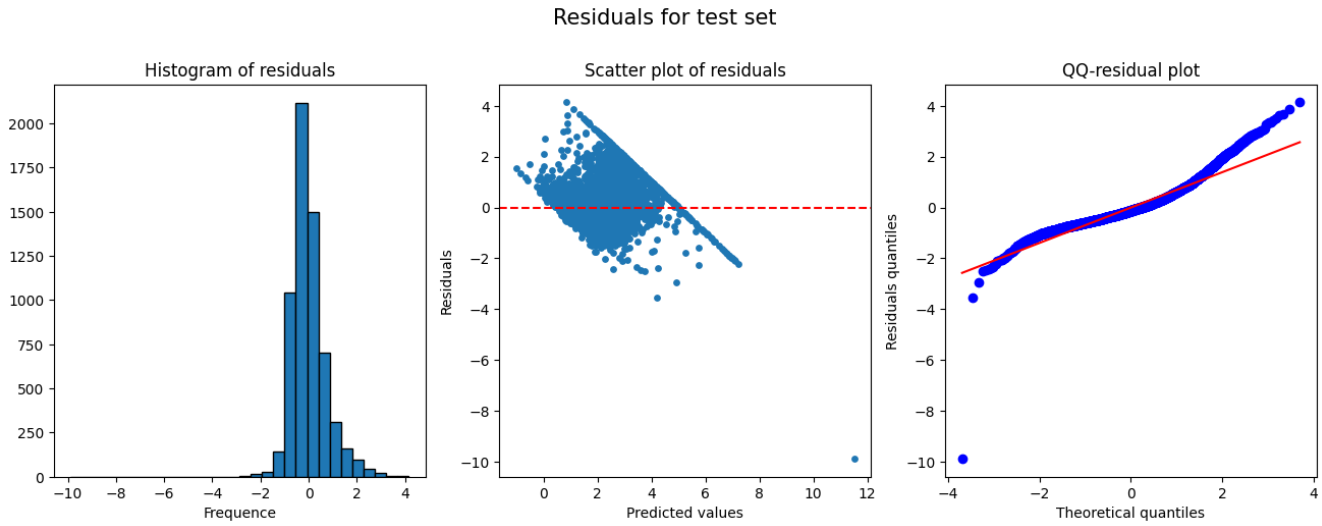

Figure 3: Residuals of the Baseline Linear Regression Model

- **Histogram of Residuals (left):**
  The distribution of residuals appears to follow a normal shape with a mean around zero, which is desirable. This suggests that the model does not have significant bias and that errors are generally centered around zero. However, there are a few extreme values, which could indicate the presence of some outliers or variance not accounted for by the model.

- **Scatter Plot of Residuals vs. Predicted Values (center):**
  This plot shows that the residuals are not perfectly random; there is a diagonal structure, suggesting heteroscedasticity (residuals increase or decrease with predicted values). This might indicate that the model has difficulty capturing certain trends or variances in the data. We can clearly see that there is an extreme outlier.

- **QQ Plot of Residuals (right):**
  This normality plot of residuals shows a deviation from the normal line, especially at the tails. This indicates that the residuals do not perfectly follow a normal distribution, which could impact the model's performance and interpretation, particularly if the analysis relies on the normality assumption. Also here the outlier is distinguishable.

4 Model Enhancement

To improve the model's performance, several data preprocessing steps are undertaken.

Removing Outliers

Outliers are identified using box plots for features with significant skewness (Figure 4).
Outliers are removed based on the whisker ranges obtained from the box plots (Figure 5).

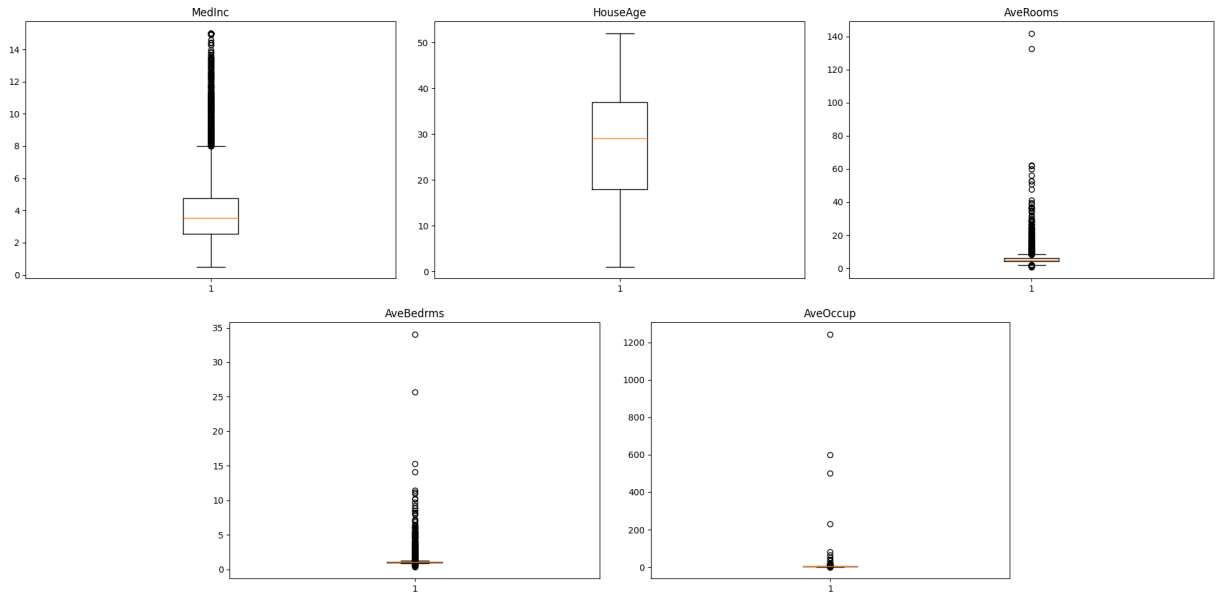

Figure 4: Box Plots before Outlier Detection

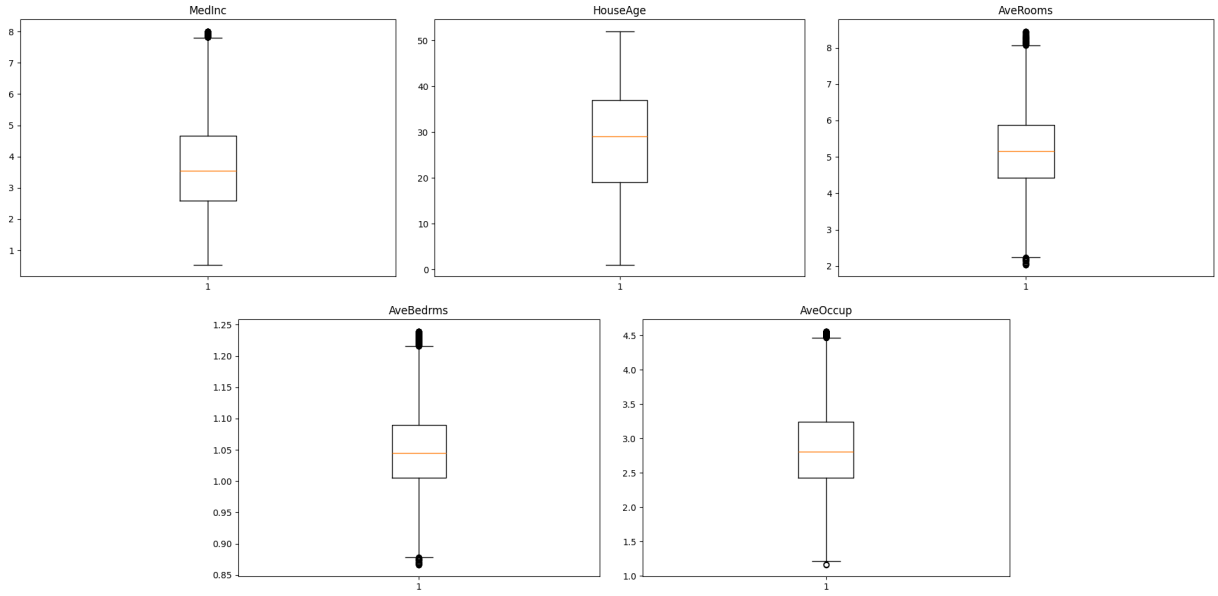

Figure 5: Box Plots before Outlier Detection

Retraining the Model

The linear regression model is retrained on the cleaned dataset.

Table 2: Model Performance After Outlier Removal

Dataset	**MSE**	R^2
Training Set	0.393	0.646
Testing Set	0.393	0.646

Addressing Multicollinearity

Due to high correlation between *AveRooms* and *AveBedrms*, *AveBedrms* is removed to reduce multicollinearity. Additionally, a new feature *RatioRoomsToBedrms* is created to capture the relationship between rooms and bedrooms.

Comparison of Regression Models

Table 3: Comparison of Regression Models

Model	**MSE**	R^2	**Adjusted** R^2
Baseline Model	0.531	0.596	0.595
Without Outliers	**0.393**	**0.646**	**0.646**
Without *AveBedrms*	0.406	0.635	0.635
With New Feature	0.399	0.641	0.641

5 Non-linear Machine Learning Models

To capture complex patterns in the data, non-linear models are explored using gradient boosting algorithms and random forests.

Hyperparameter Optimization

Optuna is used for hyperparameter tuning across several models: Random Forest, XGBoost, CatBoost, and LightGBM. The CatBoostRegressor emerged as the best model based on the validation MSE.

Table 4: Best Hyperparameters for CatBoostRegressor

Hyperparameter	**Value**
Model	CatBoost
n_estimators	914
learning_rate	0.110
depth	9

Model Performance

The CatBoost model is retrained on the combined training and validation sets for final evaluation.

Table 5: CatBoost Model Performance

Dataset	**MSE**	R^2
Training Set	0.070	0.959
Testing Set	0.170	0.853

Residuals Analysis

Figure 6 shows the residuals of the CatBoost model.

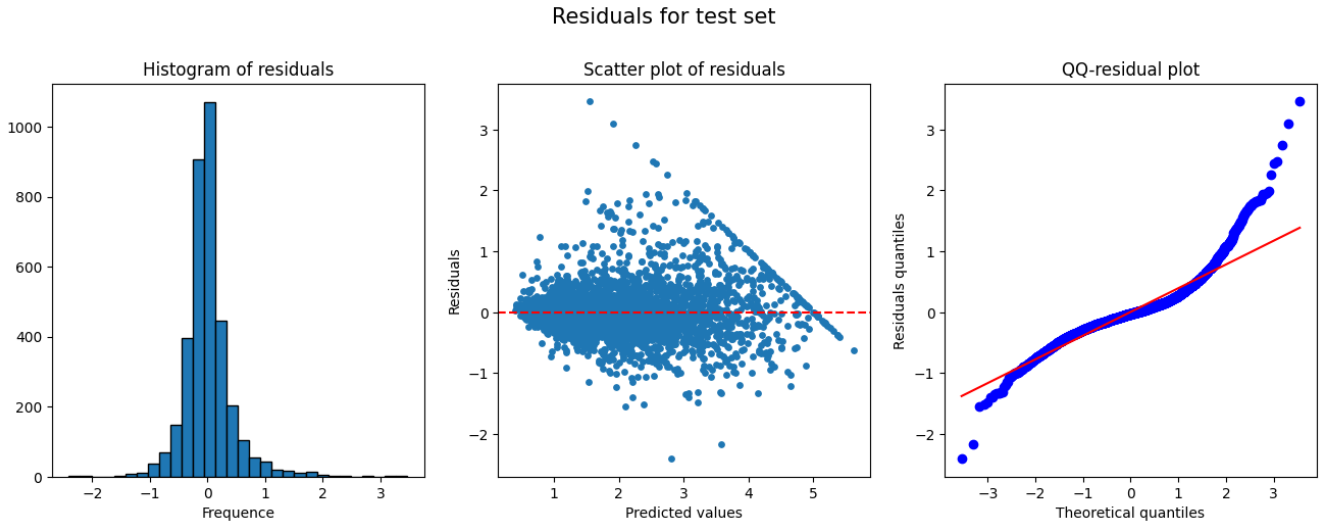

Figure 6: Residuals of the CatBoost Model

6 Conclusion

The study demonstrates that handling outliers and multicollinearity improves the performance of linear regression models. However, non-linear models like CatBoost significantly outperform traditional regression methods by capturing complex relationships in the data. The final CatBoost model achieves an R^2 of 0.853 on the test set, indicating a strong predictive capability.

References

[1] Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830.

[2] Akiba, T., et al. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework*. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

[3] Pace, R. Kelley, and Ronald Barry (1997). *Sparse Spatial Autoregressions*. Statistics and Probability Letters, 33, 291-297.