ML REPORT:

Programming Project1- Linear Models

Submitted to: Professor Poonam Kumari

8th March 2024

Keerthana Allam

Nihith Nath

Introduction.

This report presents the findings from the exploratory data analysis (EDA) and subsequent modeling efforts on the California Housing dataset, which was collected from the 1990 California census for predicting median house values across the state.

1. Exploratory Data Analysis Findings.

Dataset Overview

The California Housing dataset contains information collected from the 1990 California census. It comprises features that could be used to predict median house values in the census block groups across California.

Data Features and Descriptions

The dataset consists of 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 house occupancy
- **`Latitude`**: block group latitude
- **`Longitude`**: block group longitude
- **`MedHouseVal`**: median house value for households within a block group

Data Quality Checks

- **Missing Values**: Preliminary checks for null or missing values across the dataset indicate that there are no such values, meaning the data is relatively clean and can be used for analysis without needing imputation.
- **Data Types**: The features are primarily numerical, which is suitable for various statistical analyses and machine learning models.

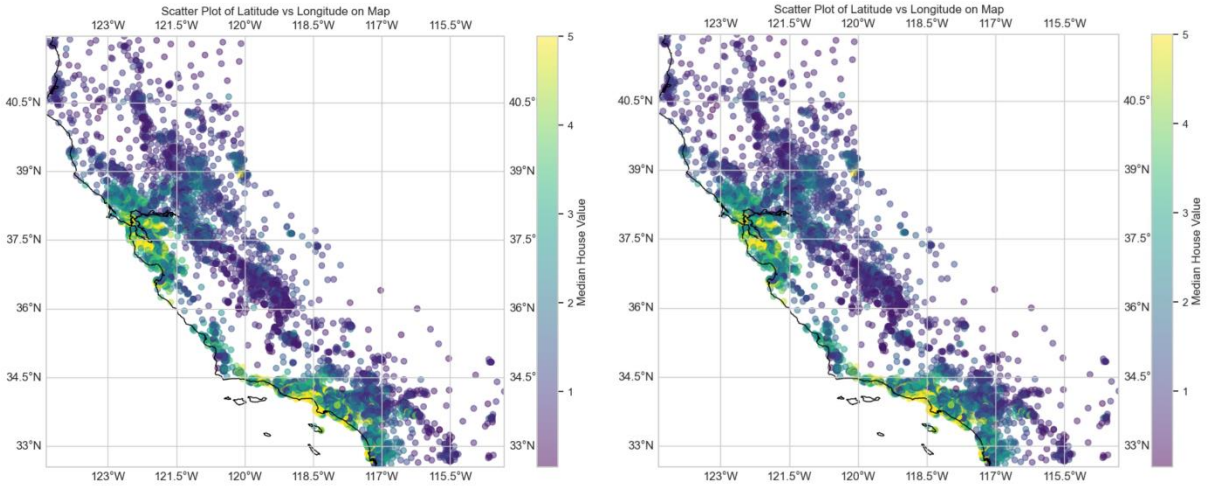
Univariate Analysis

- **`MedInc`**: This feature has a relatively wide range of values but seems to be right-skewed, indicating a smaller number of block groups with very high median income.
- **`HouseAge`**: The age of houses seems to be fairly uniformly distributed, with a slight increase for older houses, indicating a good number of older housing blocks.

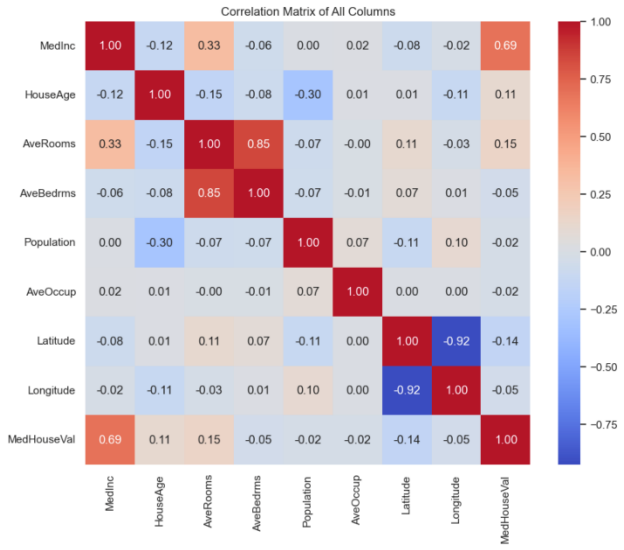
- **`AveRooms and AveBedrms`**: Both average rooms and bedrooms per household have a right-skewed distribution, showing that the majority of the households have fewer rooms and bedrooms, with some outliers having significantly more.
- **`Population`**: Population distribution is heavily right-skewed with several outliers, which suggests that most block groups have a relatively low population, but a few have very high numbers.
- **`AveOccup`**: Average occupancy per household also shows a right-skewed distribution, indicating that most households have fewer occupants.
- **`MedHouseVal`**: The median house value is the target variable for prediction and seems to have a wide range but is capped at 5, which might indicate data capping during the data collection process.

Multivariate Analysis

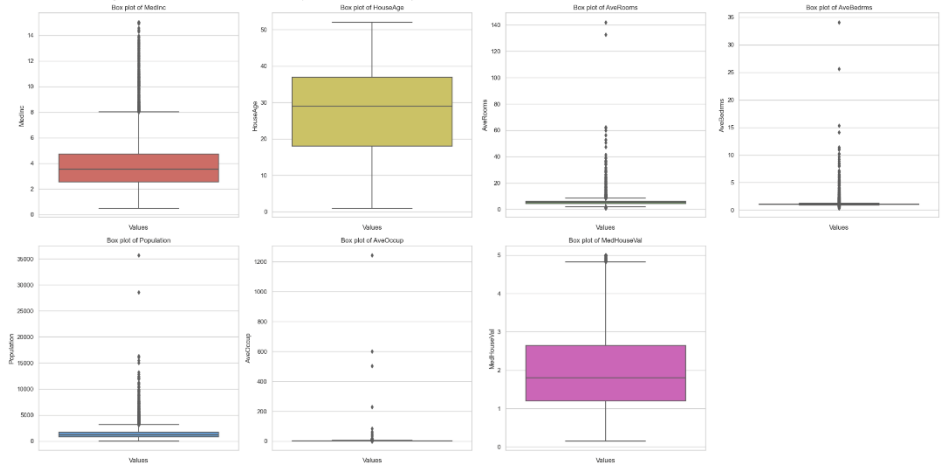
- **Scatter Plots**: The scatter plots for latitude and longitude show the geographical distribution of median income and house values. There seems to be a concentration of higher values in coastal areas.



- **Correlation Matrix**: The matrix indicates some expected relationships, such as a positive correlation between the average number of rooms and bedrooms and a negative correlation between latitude and median house value, possibly indicating that house values decrease as one moves northward and a strong positive relation between Median Income Value with Median House Value.



Outliers

- **Boxplots**: Boxplots for each feature reveal the presence of outliers, particularly in the **`AveRooms`**, **`AveBedrms`**, and **`Population`** features. This suggests that while there are common trends, there are exceptional cases that may need to be considered in any predictive modeling.
- 

Summary of Findings/Observations

- The dataset appears clean with no missing values, making it ready for further analysis without the need for data cleaning steps.
- The target variable **`MedHouseVal`** shows a wide distribution of values, indicating variability in house values across California.
- There is a geographical pattern in the data where certain locations have higher house values and income levels.
- The most significant positive correlation is between **`MedInc`** and **`MedHouseVal`**, suggesting that median income is a good predictor of median house value.

- ## 2. Simple Linear Regression Summary

Objective: To predict median house values based on median income in California.

Method: Used simple linear regression with median income as the predictor.

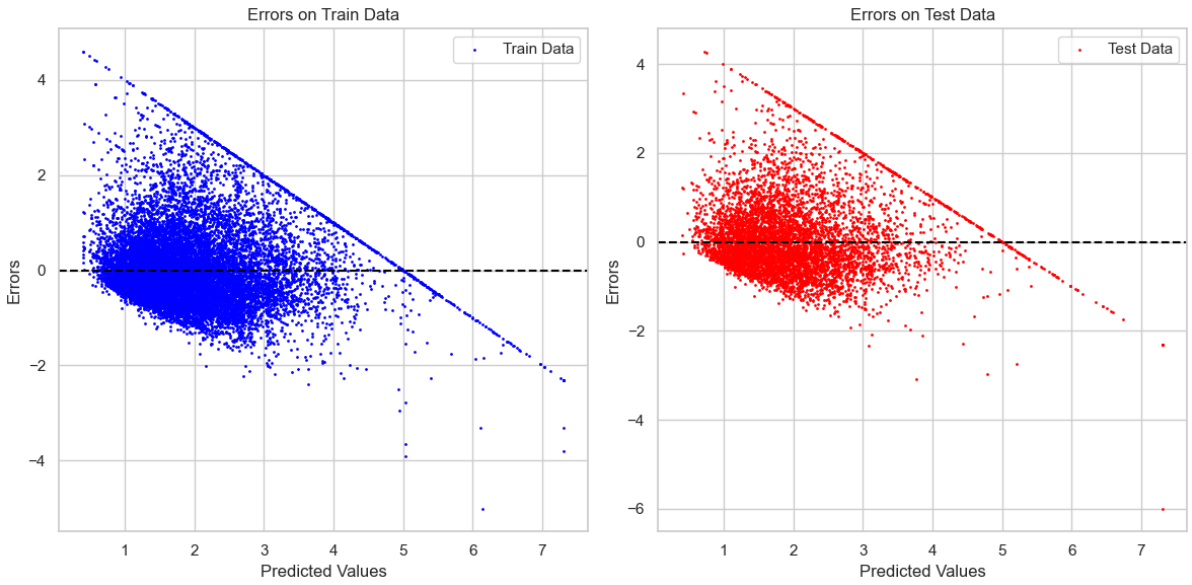
Performance:

- **Train MAE**: 0.617
- **Test MAE**: 0.618
- **Train MSE**: 0.716
- **Test MSE**: 0.720
- **Train R^2**: 0.460
- **Test R^2**: 0.463

Results

- The linear regression line in the plot displays the general trend of the predictions by the model relative to the actual data points.

- The presence of heteroscedasticity and the relatively wide spread of residuals suggest that a simple linear model might not be the best fit for this dataset.



- The R^2 score is satisfactory, but there is significant room for improvement, possibly by incorporating more features or trying a different kind of model.

3. Multiple Linear Regression Model

The model was expanded from a simple linear regression that used only the median income as a predictor to a multiple linear regression model. Model Performance:

The performance of the multiple linear regression model was compared to the simple linear regression model based on median income alone.

- **Multiple vs. Simple Linear Regression**: The multiple linear regression model has a lower MSE and a higher R^2 compared to the simple linear regression model, indicating improved prediction accuracy and a better fit to the data.
- **Error Distributions**: The error plots for both models show a pattern in the residuals, which may suggest that further feature engineering or a more complex model could be necessary to capture all the underlying relationships.

Mean Squared Error (Train): 0.5508
Mean Squared Error (Test): 0.5634
R-squared: 0.5749
The weights and bias of the model are as follows:
Weights:
Feature 1: 0.8167
Feature 2: 0.1775
Feature 3: -0.1303
Feature 4: 0.1438
Feature 5: 0.0162

Feature 6: -0.0490
Feature 7: -0.4857
Feature 8: -0.4503
Bias: 2.0709
The increased R^2 value in the multiple linear regression model indicates a better fit, suggesting that the additional features provide more information about the variability in the median house values.

Final Model Equation:
Given the weights and bias, the final equation for the multiple linear regression model is:
Y=0.8167·MedInc+0.1775·HouseAge−0.1303·AveRooms+−0.4503·Longitude+2.0709y

4. Locally Weighted Linear Regression Model

LWLR was implemented using a Gaussian kernel function. We tuned the hyperparameter tau, which dictates the bandwidth of the kernel and hence the degree of locality. A range of tau values (0.1, 0.5, 1, 10) was explored to determine the best fit for the model.

The dataset was sampled at 10%, with 'MedHouseVal' as the target variable and the remaining features, excluding 'Population', 'AveOccup', and 'AveBedrms', as predictors. The data was split with 30% reserved for testing to validate the model's predictions.
Results
The results indicated varying levels of accuracy across different tau values:

- Tau=0.1 provided the lowest MSE and the highest R^2, signifying the closest fit to the data.

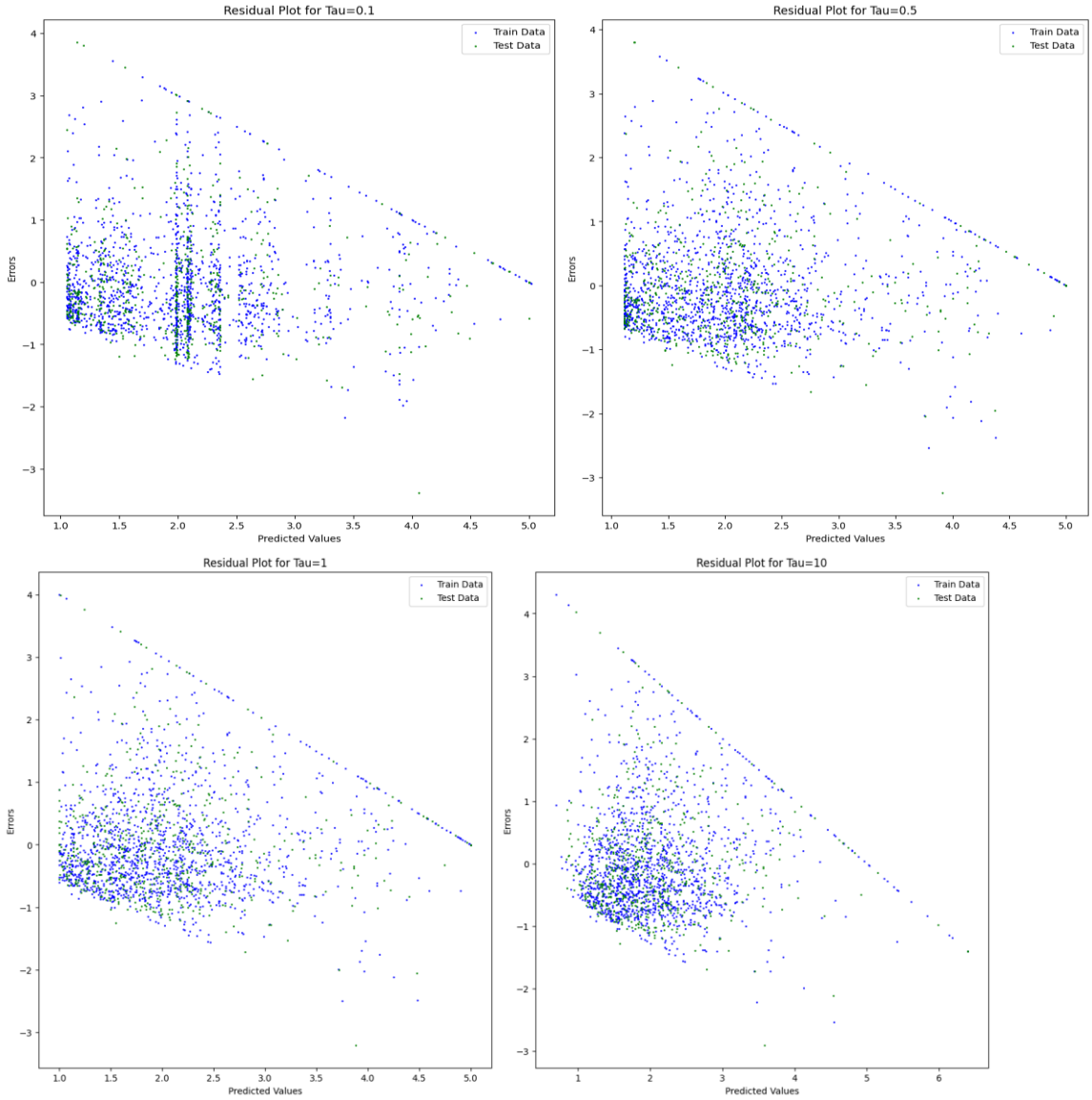
- As tau increased, both the training and testing MSE gradually increased, and R^2 decreased. This indicates a decline in model performance, with the model's predictions diverging from actual values.

Tau=0.1:
Train MSE=0.6765291478967527, Test MSE=0.7041670735408138, Train R^2=0.4974141957858451, Test R^2=0.45931076742080346

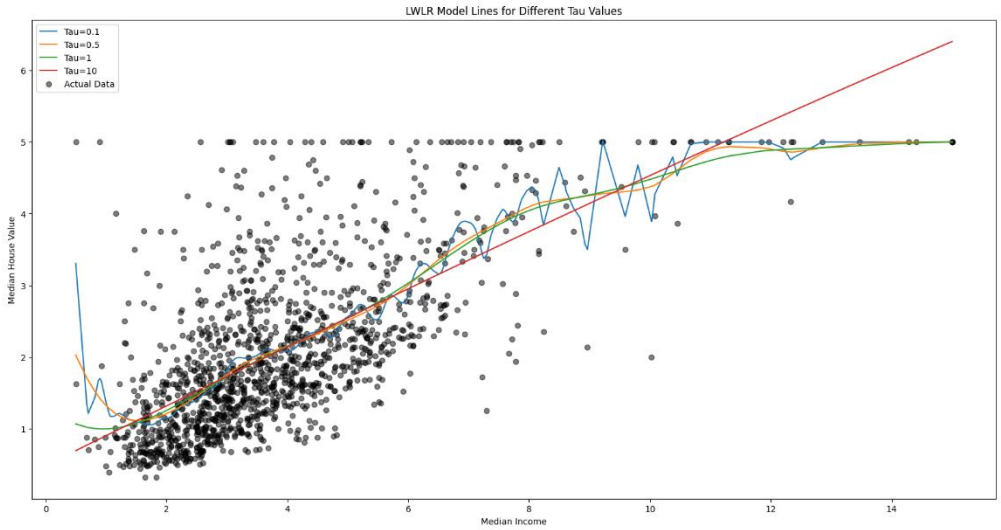
Tau=0.5:
Train MSE=0.6996748372827815, Test MSE=0.6994083642460799, Train R^2=0.4802195265682172, Test R^2=0.4629647054893633

Tau=1:
Train MSE=0.7092789058762872, Test MSE=0.7007709369654351, Train R^2=0.4730847733165634, Test R^2=0.46191846458199315

Tau=10:
Train MSE=0.7308246853082726, Test MSE=0.7127206283853093, Train R^2=0.45707865899479316, Test R^2=0.45274298659367807



We understand that plotting should be multi-dimensional, but the below plot was to just check how the polynomial looks like for a single feature assuming that other features are not not impacting the house value.



5. Model Comparison and Selection

Comparison

- **Training Error**: LWLR (tau=0.1) had the lowest training error, suggesting an excellent fit to the training data. However, MLR was more balanced, offering a robust model that did not overfit.
- **Testing Error**: LWLR with higher tau values and MLR showed more generalizable results on the test data compared to SLR and LWLR with a low tau.
- **Error Distribution**: Residual plots revealed patterns across models, especially for LWLR with lower tau values, indicating potential model complexity issues.
- **R-squared (R^2)**: All models displayed moderate R^2 scores, with Multi Linear Regression slightly leading, suggesting room for including more complex models like polynomial regression or machine learning algorithms.

Metric	SLR	MLR	LWLR (Tau=0.1)	LWLR (Tau=0.5)	LWLR (Tau=1)	LWLR (Tau=10)
Train MSE	0.716	0.5508	0.6765	0.6997	0.7093	0.7308
Test MSE	0.720	0.5634	0.7042	0.6994	0.7008	0.7127
Train R^2	0.460	N/A	0.4974	0.4802	0.4731	0.4571
Test R^2	0.463	0.5749	0.4593	0.4629	0.4619	0.4527

Final Recommendations

- Use MLR for a balanced approach incorporating multiple predictors.
- Consider LWLR with carefully tuned tau values for localized predictions.
- Employ cross-validation for model robustness.
- Explore advanced regression techniques for further improvements.

Conclusion

The exploration and analysis of the California Housing dataset through various linear models have provided valuable insights into predicting median house values. The multiple linear regression model, in particular, offers a promising approach for incorporating multiple predictors effectively.