About Dataset

The dataset pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data. It contains one row per census block group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes data. It can bd downloaded here.

Columns:

1. longitude: A measure of how far west a house is

2. latitude: A measure of how far north a house is

3. housing_median_age: Median age of a house within a block

4. total_rooms: Total number of rooms within a block

5. total_bedrooms: Total number of bedrooms within a block

6. population: Total number of people residing within a block

7. households: Total number of households, a group of people residing within a home unit, for a block

8. median_income: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

9. median_house_value: Median house value for households within a block (measured in US Dollars)

10. ocean_proximity: Location of the house with reference to ocean/sea

Objective:

The goal of this analysis is to process and explore the dataset, utilize it to develop regression models that predict the median housing value for districts, train a machine learning algorithm, and report on the model's performance.

Exploratory Data Analysis:

The dataset comprises 20,640 observations and 10 columns, with most attributes being numerical, except for **ocean_proximity**, which is categorical. The **total_bedrooms** attribute has 20,433 non-null values, indicating 207 missing entries. The histogram reveals that **total_bedrooms** is skewed and contains outliers. Filling these missing values with the median is advisable to avoid the influence of outliers on the mean. There are no duplicate entries in the dataset.

The ocean_proximity attribute contains five categorical values: **<1H OCEAN**, **INLAND**, **NEAR OCEAN**, **NEAR BAY**, and ISLAND. These values were converted into numerical format for regression analysis.

The target variable, **median_house_value**, shows a mild correlation with most features except for **median_income**, which exhibits a strong linear relationship, as seen in the scatter plot. Features such as **households**, **population**, **total_bedrooms**, and **total_rooms** display strong correlations with each other. Although multicollinearity affects the coefficients and p-values, it does not impact prediction accuracy or the goodness-of-fit. Since the primary focus is on prediction rather than understanding the role of each feature, multicollinearity reduction is unnecessary. Additionally, **longitude** and **latitude** are highly correlated.

Data visualization through box plots and histograms highlights the presence of outliers and the distribution of data. The **median_house_value** shows an unusual peak near its maximum value of around $500k, suggesting potential outliers. The features **households**, **population**, **total_bedrooms**, **total_rooms**, and **median_income** have skewed distributions and a broad range on the x-axis, indicating numerous outliers. The distributions for **population**, **total_bedrooms**, and **total_rooms** are skewed towards smaller values and are interconnected.

Conclusions from EDA:

Identified and processed outliers in several features.
Overall data distribution is approximately normal.
Strong correlation between **median_house_value** and **median_income**.
High correlation among some features; multicollinearity is disregarded as the primary focus is on prediction.
The **<1H OCEAN** category in **ocean_proximity** is the most frequent.

Data Preparation:
Analysing and preparing the dataset involved key steps: data cleaning, data reduction, and data transformation. Strategies for each step are outlined below.

Data Cleaning:

Handling missing values: Missing values can be removed or filled in with mean, median, or mode values. Here I decided to fill the missing values with the median value as that feature was skewed and had few outliers.

Removing duplicates: duplicated rows should be identified and removed not to impact the accuracy of data. This dataset has no duplicates.

Handling outliers: few features has outliers. Outliers in Median_house_value, median_income and households were spotted and removed.

Data Transformation:

Normalisation: Min-Max scaling was used to scale numerical features with huge ranges to a similar range to prevent certain features from dominating the analysis.

Encoding categorical variables: Converted categorical variables of ocean_proximity feature into numerical representations using label encoding where each category was assigned a Unique Integer value.

Machine learning model:

In this analysis, I utilised multiple machine learning models to predict housing values. The models include Multiple Linear Regression and Decision Tree Regression, as well as ensemble methods like Bagging Regressor, AdaBoost Regressor, and Random Forest Regression. Each model was assessed based on its performance metrics to determine the most effective approach.

1. Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical technique that models the relationship between a dependent variable and multiple independent variables by fitting a linear equation. It assumes a linear relationship between the predictors and the target variable.

Performance Evaluation:

- **Training Score:** 0.592
- **Testing Score:** 0.599
- **MSE:** $3555499321
- **RMSE:** $59628
- **MAE:** $44844
- **Adjusted R^2:** 0.598
- **MAPE:** 28.56%

MLR provides a baseline for comparison, capturing linear relationships but struggling with complex patterns and non-linearity, resulting in moderate accuracy.

2. Decision Tree Regression

Decision Tree Regression creates a tree-like model that splits the data into subsets based on the values of input features. It can capture non-linear relationships and interactions between variables effectively.

Performance Evaluation:

- **Training Score:** 0.761
- **Testing Score:** 0.644
- **MSE:** $3158258573
- **RMSE:** $56198
- **MAE:** $38505
- **Adjusted R^2:** 0.643
- **MAPE:** 22.47%

Decision Tree Regression performed better than MLR by capturing non-linear patterns but was prone to overfitting, as indicated by the high training score compared to the testing score.

3. Bagging Regressor

Bagging Regressor, or Bootstrap Aggregating, is an ensemble technique that improves model stability and accuracy by combining the predictions of multiple decision trees. It reduces variance and enhances generalisation.

Performance Evaluation:

- **Training Score:** 0.807
- **Testing Score:** 0.731
- **MSE:** $2383877724
- **RMSE:** $48825
- **MAE:** $34073
- **Adjusted R^2:** 0.731
- **MAPE:** 20.29%

Bagging Regressor improved over single decision trees by reducing variance, leading to better generalisation and more accurate predictions.

4. AdaBoost Regressor

AdaBoost Regressor (Adaptive Boosting) sequentially applies weak learners (typically decision trees) to the data, focusing on the mistakes of previous models. Each subsequent model corrects errors made by the previous ones, enhancing performance.

Performance Evaluation:

- **Training Score:** 0.854
- **Testing Score:** 0.747
- **MSE:** $2245030115
- **RMSE:** $47382
- **MAE:** $36895
- **Adjusted R^2:** 0.746
- **MAPE:** 25.35%

AdaBoost Regressor achieved higher accuracy by iteratively correcting errors, making it more effective than bagging in handling complex patterns.

5. Random Forest Regression

Random Forest Regression constructs an ensemble of decision trees and aggregates their predictions to improve accuracy and robustness. It leverages the strength of multiple trees to mitigate overfitting and capture intricate data patterns.

Performance Evaluation:

- **Training Score:** 0.965
- **Testing Score:** 0.772
- **MSE:** $2021954958
- **RMSE:** $44966
- **MAE:** $30222
- **Adjusted R^2:** 0.772
- **MAPE:** 17.62%

Random Forest Regression outperformed other models by effectively balancing bias and variance, providing the best generalisation and prediction accuracy among the tested models.

Model Comparison

A comparison of model performances is summarised below:

Model	Training R^2	Testing R^2	MSE	RMSE	MAE	Adjusted R^2	MAPE
Multiple Linear Regression	0.592	0.599	$3,555499321	$59,628	$44,844	0.598	28.56%
Decision Tree Regression	0.761	0.644	$3,158,258,573	$56,198	$38,505	0.643	22.47%
Bagging Regressor	0.807	0.731	$2,383,877,724	$48,825	$34,073	0.731	20.29%
AdaBoost Regressor	0.854	0.747	$2,245,030,115	$47,382	$36,895	0.746	25.35%
Random Forest Regression	0.965	0.772	$2,021,954,958	$44,966	$30,222	0.772	17.62%

Conclusions:

- **Random Forest Regression** demonstrated superior performance, excelling in prediction accuracy and generalisability, making it the most effective model.
- **AdaBoost Regressor** and **Bagging Regressor** also provided strong results, with AdaBoost being particularly effective in improving performance through error correction.
- **Decision Tree Regression** and **Multiple Linear Regression** offered simpler, interpretable models but lagged in accuracy compared to ensemble methods.

Overall, ensemble methods like Random Forest and AdaBoost provided significant improvements over individual models, effectively capturing complex patterns and delivering higher accuracy.

Reference:
California Housing : https://github.com/ageron/handson-ml/tree/master/datasets/housing