PREDICTING HOUSING PRICES : USING MACHINE LEARNING

Introduction

This report summarizes the process and results of building a machine learning model to predict housing prices using the California housing dataset. The key aspects covered include an overview of the dataset and its features, data preprocessing steps, model training and evaluation, and an interpretation of the model's performance and coefficients.

Dataset and Features

This is the dataset used in the second chapter of Aurélien Géron's recent book 'Hands-On Machine learning with Scikit-Learn and TensorFlow'. It serves as an excellent introduction to implementing machine learning algorithms because it requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size between being to toyish and too cumbersome.

The data contains information from the 1990 California census. So although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introductory dataset for teaching people about the basics of machine learning.

The California housing dataset contains information on various attributes of housing in California. The key features include:

longitude: Longitude coordinate of the house.
latitude: Latitude coordinate of the house.
housing_median_age: Median age of the house.
total_rooms: Total number of rooms in the house.
total_bedrooms: Total number of bedrooms in the house.
population: Population in the neighborhood.
households: Number of households in the neighborhood.
median_income: Median income of the households.
median_house_value: Median house value (target variable).
ocean_proximity: Proximity to the ocean (categorical variable, one-hot encoded). Values include [near bay, near ocean, inland, Island, <1h ocean].

Data Preprocessing

Loading and Inspecting the Data

- **Steps:**
    - Load the dataset using pandas
    - The dataset was loaded using the pandas library and then descriptive statistics was checked about it.
    - Display the first few rows of the dataset.

Handling Missing Values

- **Steps:**
    - Identify missing values.
    - Choose an appropriate method to handle missing values (e.g., mean/modal imputation).

Feature Scaling and Normalisation

- **Steps:**
    - Apply feature scaling (e.g., StandardScaler or MinMaxScaler) to ensure all features contribute equally to the model.

Splitting the Data

- **Steps:**
    - Split the data into training and testing sets (e.g., 80-20 split).

Model Development

4.1 Model Selection

- **Model 1 :** Choose linear regression model (e.g., Linear Regression from sklearn).
- **Model 2** : Choose Random Forest Regression model (e.g.,Random Forest Regressor from sklearn)

4.2 Training the Model

- **Steps:**
    - Train both the models on the training dataset.

- Display the model coefficients.

4.3 Evaluating the Model

- **Metrics:**
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - R^2 Score
- **Steps:**
  - Evaluate the model on the testing dataset.
  - Display the evaluation metrics.

Model Evaluation and Interpretation

Visualisation

- **Plots:**
  - Actual vs. Predicted prices.
  - Residual plot.

Interpretation

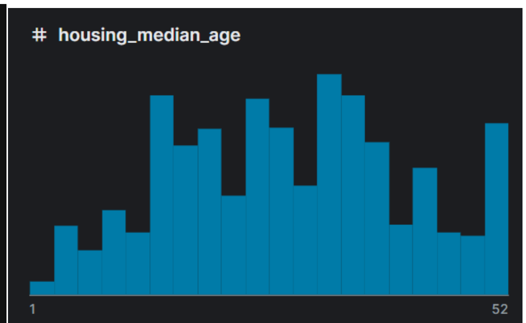
- **Coefficients:**
  - Interpret the model coefficients to understand the impact of different features on housing prices.
- **Findings:**
  - From the Above model and Explatory Data Analysis we find that the correlation between number of rooms and the price of our house is positively corelated to each other.

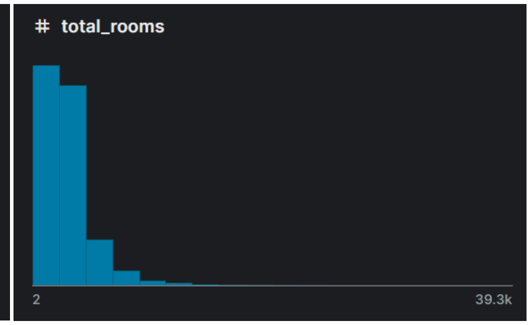
Conclusion

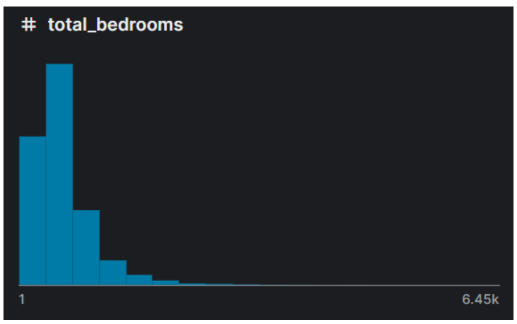
- **Summary:** Our random forest regressor model performs well with about 80% plus accuracy. This model was deployed as a locally run web application using streamlit library and with the help of this application we can check/test our own prices for Houses based on California Housing Price.

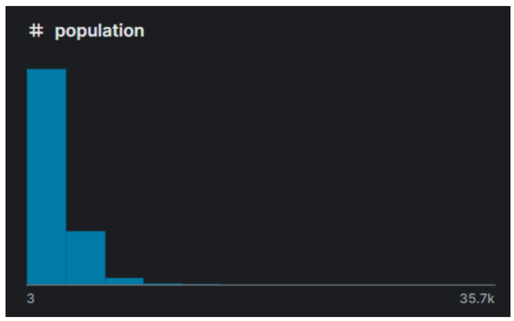
Photo Gallery

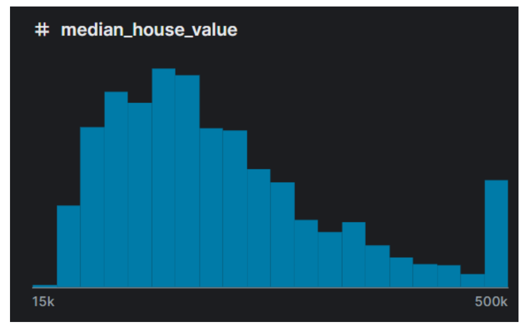
```
Information on California DF :

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
None
```

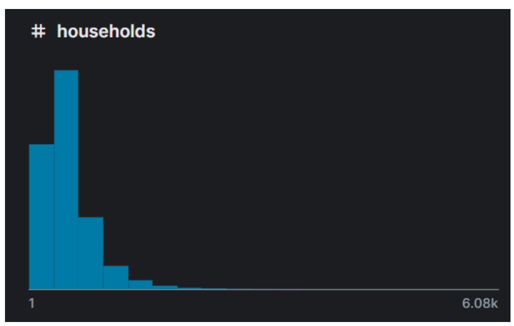


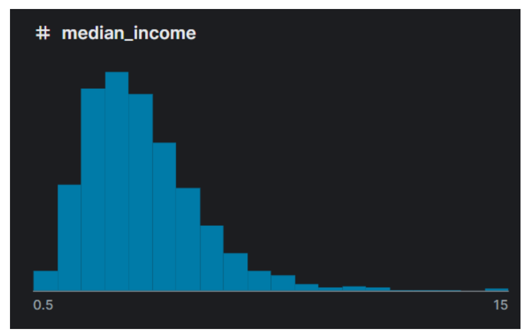








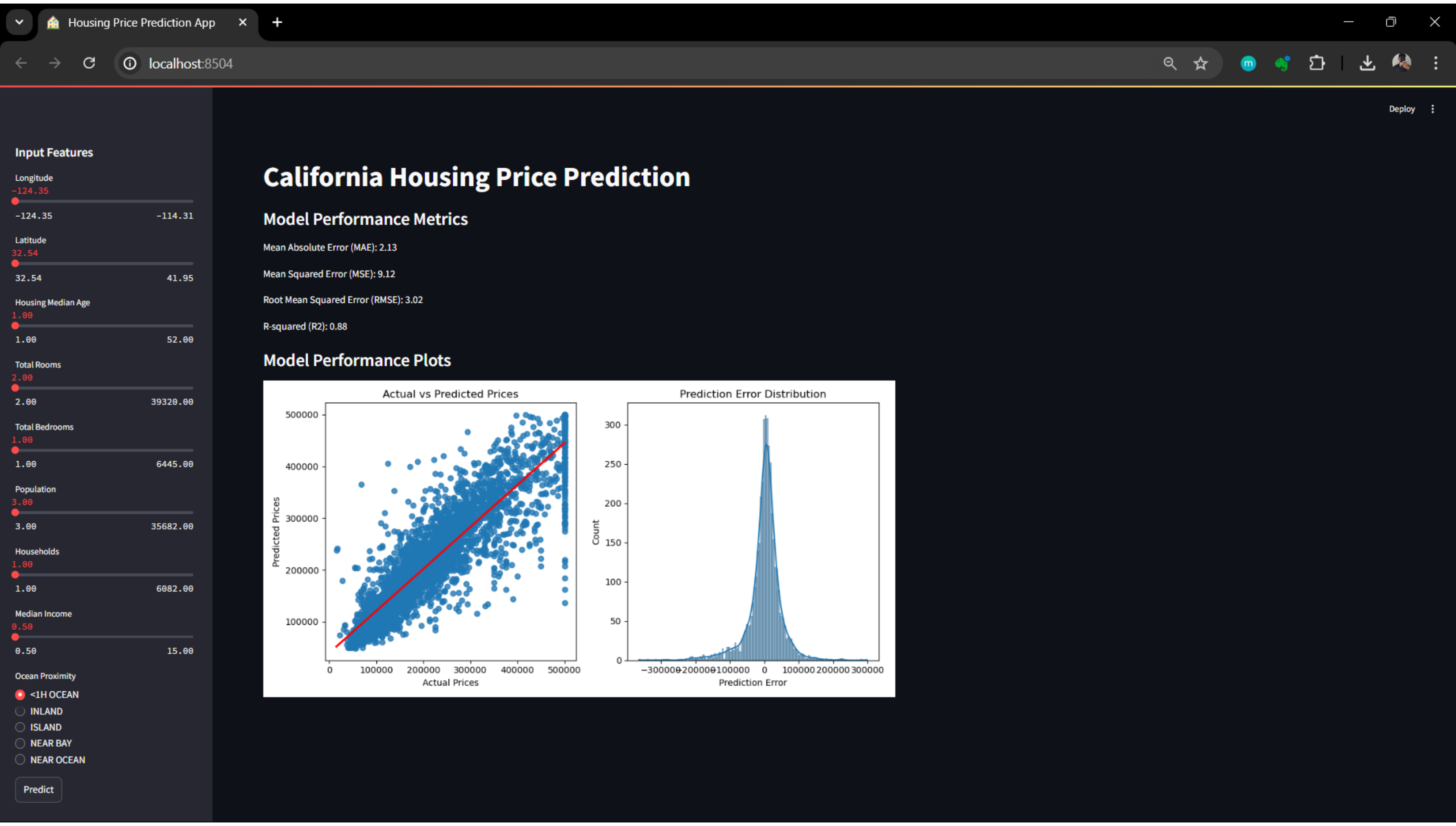




	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY
5	-122.25	37.85	52.0	919.0	213.0	413.0	193.0	4.0368	269700.0	NEAR BAY

```
Performance Metrics for Linear Regression Model
 ---------------------------------------------------
Linear Model R2 Score : 0.6257420882414747
Mean Absolute Error (MAE) : 50722.24170136072
Mean Squared Error (MSE) : 4904309277.46062
Root Mean Squared Error (RMSE) : 70030.77378881816
```

```
Performance Metrics for Random Forest Regressor Model
 ---------------------------------------------------
Linear Model R2 Score : 0.8177978442947841
Mean Absolute Error (MAE) : 31665.932480620155
Mean Squared Error (MSE) : 2387593406.909613
Root Mean Squared Error (RMSE) : 48863.006527531776
```


Housing Price Prediction App
localhost:8504
Deploy
Input Features
Longitude
-124.35
-124.35
-114.31
Latitude
32.54
32.54
41.95
Housing Median Age
1.00
1.00
52.00
Total Rooms
2.00
2.00
39320.00
Total Bedrooms
1.00
1.00
6445.00
Population
3.00
3.00
35682.00
Households
1.00
1.00
6082.00
Median Income
0.50
0.50
15.00
Ocean Proximity
<1H OCEAN
INLAND
ISLAND
NEAR BAY
NEAR OCEAN
Predict
California Housing Price Prediction
Model Performance Metrics
Mean Absolute Error (MAE): 2.13
Mean Squared Error (MSE): 9.12
Root Mean Squared Error (RMSE): 3.02
R-squared (R2): 0.88
Model Performance Plots
Actual vs Predicted Prices
Predicted Prices
Actual Prices
Prediction Error Distribution
Count
Prediction Error