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 Boston 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

Each record in the database describes a Boston suburb or town. The data was drawn from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. The attributes are defined as follows (taken from the UCI Machine Learning Repository1)

1. **ZN: proportion of residential land zoned for lots over 25,000 sq.ft.**
2. **INDUS: proportion of non-retail business acres per town**
3. **CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)**
4. **NOX: nitric oxides concentration (parts per 10 million)**
   1https://archive.ics.uci.edu/ml/datasets/Housing
   123
   20.2. Load the Dataset 124
5. **RM: average number of rooms per dwelling**
6. **AGE: proportion of owner-occupied units built prior to 1940**
7. **DIS: weighted distances to five Boston employment centers**
8. **RAD: index of accessibility to radial highways**
9. **TAX: full-value property-tax rate per $10,000**
10. **PTRATIO: pupil-teacher ratio by town 12. B: 1000(Bk−0.63)2 where Bk is the proportion of blacks by town 13. LSTAT: % lower status of the population**
11. **MEDV: Median value of owner-occupied homes in $1000s**
    We can see that the input attributes have a mixture of units.

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.

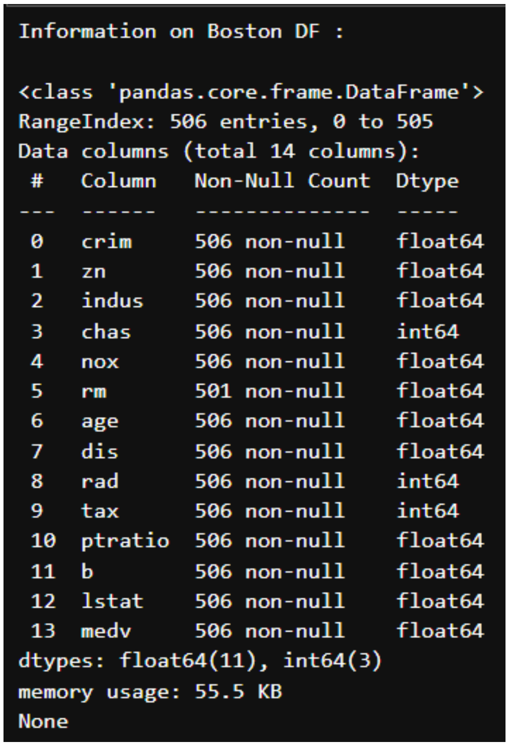
Interpretation

- **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 Boston DF :

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   crim     506 non-null    float64
 1   zn       506 non-null    float64
 2   indus    506 non-null    float64
 3   chas     506 non-null    int64
 4   nox      506 non-null    float64
 5   rm       501 non-null    float64
 6   age      506 non-null    float64
 7   dis      506 non-null    float64
 8   rad      506 non-null    int64
 9   tax      506 non-null    int64
 10  ptratio  506 non-null    float64
 11  b        506 non-null    float64
 12  lstat    506 non-null    float64
 13  medv     506 non-null    float64
dtypes: float64(11), int64(3)
memory usage: 55.5 KB
None
```

Boston Housing Dataset first 10 Entries:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
5	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7