Predicting House Prices using Machine

Learning

**Phase 4: Development Part 2**

**Malarvizihi.R**

**AU711021205013**

**Predicting House Prices using Machine Learning**

**Phase 4: Development Part 2**

**Table of Contents**

* **Introduction**
* **Data Collection and Preprocessing**
* **Data Sources**
* **Data Cleaning**
* **Feature Engineering**
* **Data Splitting**
* **Exploratory Data Analysis (EDA)**
* **Data Visualization**
* **Statistical Analysis**
* **Model Selection**
* **Linear Regression**
* **Decision Trees**
* **Random Forest**
* **XGBoost**
* **Evaluation Metrics**
* **Model Training and Evaluation**
* **Training the Models**
* **Model Evaluation**
* **Hyperparameter Tuning**
* **Grid Search**
* **Random Search**
* **Feature Importance**
* **Deployment**
* **Saving the Model**
* **Creating an API**
* **Conclusion**

# INTRODUCTION:

* Briefly introduce the problem of house price prediction and the importance.
* Mass valuations using automated systems are commonly used in the collection of recurring annual taxes, but can also be used for sporadic real estate taxes (property transfer taxes, capital gains tax, and inheritance and gifts taxes), banking (loans, mortgage risk), real estate portfolio estimation, real estate marketers, among others .
* For a mass appraisal model to be uniform and accurate in estimating real estate prices, it will be necessary for the data to be correct, complete and up-to-date.
* **Data Collection and Preprocessing**
* Explain how to collect and clean the dataset.
* Describe feature engineering and data splitting.
* In order to proceed further into this project, first of all we should collect the data, after collecting a dataset, then we should pre-process the data and then we do the exploratory data analysis on the given data set.
* This pre-processing of data consists of Handling missing values which means some of values in the dataset may be missing.
* We should handle those missing values. Handling duplicates which mean some of the values in the dataset may be their multiple times we should handle those duplicate values that means we should delete those duplicates
* **Exploratory Data Analysis (EDA)**
* Visualize the data using Python libraries like Matplotlib and Seaborn.
* Perform statistical analysis to gain insights into the dataset.
* By using matplotlib package in python we will do EDA on a given dataset. Normally we will do EDA on a given data set to establish a relation between given feature of the data set.
* Training: - This phase consists of 4 parts they are:- 1.Choosing a machine learning algorithm 2.Training 3.Evaluation 4.Hyperparameter tuning.
* Input code:

sns.pairplot(USA\_Housing)

* **Model Selection**
* Discuss various machine learning models suitable for regression tasks.
* Explain the choice of models and the significance of evaluation metrics.
* Select the model that performs best on your dataset and aligns with your project's goals. Consider the trade-off between model complexity and performance.
* *#Getting all Coulmn names*

USA\_Housing.columns

* **Model Training and Evaluation**
* Walk through the steps of training the selected models.
* Evaluate model performance using metrics like RMSE, MAE, etc .
* *#Getting all Coulmn names*

USA\_Housing.columns

* *# Columns as Features*

X = USA\_Housing[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',

'Avg. Area Number of Bedrooms', 'Area Population']]

* *# Price is my Target Variable, what we trying to predict*

y = USA\_Housing['Price']

* **Hyperparameter Tuning**
* Describe techniques like grid search and random search for hyperparameter optimization.
* Fine-tune model hyperparameters using techniques like Grid Search or Random Search.
* from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.4, random\_state=101)

* *#importing the Linear Regression Algorithm*

from sklearn.linear\_model import LinearRegression

* *#Training the Data Model*

lm.fit(X\_train, y\_train)

* **Feature Importance**
* Discuss the importance of feature selection and how to determine which features are most influential.
* Understanding feature importance can inspire feature engineering efforts. If certain interactions or transformations of features seem important, you can create new features to capture those relationships.
* Feature importance analysis is a versatile tool in predicting house prices using machine learning. It aids in feature selection, model selection, and model refinement, ultimately helping you build more accurate and interpretable models for real estate price predictions.
* from sklearn import metrics
* print('MAE:', metrics.mean\_absolute\_error(y\_test, predictions))

print('MSE:', metrics.mean\_squared\_error(y\_test, predictions))

print('RMSE:', np.sqrt(metrics.mean\_squared\_error(y\_test, predictions)))

* **Deployment**
* Explain how to save the trained model and create an API for predictions.
* API Development: If your model is deployed as a service, you may need to create an API (Application Programming Interface) that allows other software components to interact with and query your model for prediction.
* User Training: If your model is intended for use by non-technical users, provide training and support to ensure users can utilize it effectively.
* predictions = lm.predict(X\_test)
* plt.scatter(y\_test, predictions, edgecolor='black')
* from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.4, random\_state=101)

* Program Output

Avg. Area Income ... Address

0 79545.458574 ... 208 Michael Ferry Apt. 674\nLaurabury, NE 3701...

1 79248.642455 ... 188 Johnson Views Suite 079\nLake Kathleen, CA...

2 61287.067179 ... 9127 Elizabeth Stravenue\nDanieltown, WI 06482...

3 63345.240046 ... USS Barnett\nFPO AP 44820

4 59982.197226 ... USNS Raymond\nFPO AE 09386

... ... ... ...

4995 60567.944140 ... USNS Williams\nFPO AP 30153-7653

4996 78491.275435 ... PSC 9258, Box 8489\nAPO AA 42991-3352

4997 63390.686886 ... 4215 Tracy Garden Suite 076\nJoshualand, VA 01...

4998 68001.331235 ... USS Wallace\nFPO AE 73316

4999 65510.581804 ... 37778 George Ridges Apt. 509\nEast Holly, NV 2...

[5000 rows x 7 columns]

0 False

1 False

2 False

3 False

4 False

...

4995 False

4996 False

4997 False

4998 False

4999 False

Name: Avg. Area Number of Bedrooms, Length: 5000, dtype: bool

* **Conclusion**
* Summarize the key findings and the effectiveness of machine learning in predicting house prices.
* The machine learning model using linear regression algorithm is very helpful in predicting the house prices for real estate customers. Here we have used a supervised learning approach in machine learning field which will yield us a best possible result.
* The linear regression algorithm is used for this project because it is very simple to implement and hence gives accurate prediction of house price. Here in this project, we used python programming language.
* We also used different python packages like NumPy, pandas, matplotlib etc. For importing the dataset, and also for doing data pre-processing we used pandas. For doing exploratory data analysis we used matplotlib package in python.