PREDICTING HOUSE PRICES USING MACHINE LEARNING

E.Yogapriya (420721104057)

ABSTRACT :

The real estate industry plays a pivotal role in the economic well-being of individuals and nations. Buying or selling a home is often one of the most significant financial transactions in a person's life, and having an accurate understanding of property values is essential. Traditional methods of determining house prices rely on manual appraisal, market trends, and expert judgment. However, these methods can be subjective, time-consuming, and prone to human bias.

INTRODUCTION :

In response to these challenges, this project proposes the use of machine learning techniques to predict house prices. Machine learning algorithms have demonstrated remarkable capabilities in various domains, including finance, healthcare, and natural language processing. By harnessing the power of data and advanced predictive models, we aim to develop a system that can provide more accurate and objective estimates of house prices.

The housing market is a cornerstone of the global economy, and accurately predicting house prices is of paramount importance to buyers, sellers, and real estate professionals alike. In recent years, the advent of machine learning techniques has revolutionized the field of real estate, enabling us to develop more accurate and efficient predictive models. This project aims to leverage the power of machine learning to create a robust and reliable system for predicting house prices.

DATA COLLECTION:

We will gather a comprehensive dataset that includes information about various features of properties, such as location, size, number of bedrooms, bathrooms, and other relevant attributes. This dataset will serve as the foundation for our predictive model.

DATASET LINK: https://www.kaggle.com/datasets/vedavyasv/usa-housing

Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
79545.45857	5.682861322	7.009188143	4.09	23086.8005	1059033.558	208 Michael Ferry Apt. 674
79248.64245	6.002899808	6.730821019	3.09	40173.07217	1505890.915	188 Johnson Views Suite 079
61287.06718	5.86588984	8.51272743	5.13	36882.1594	1058987.988	9127 Elizabeth Stravenue
63345.24005	7.188236095	5.586728665	3.26	34310.24283	1260616.807	USS Barnett
59982.19723	5.040554523	7.839387785	4.23	26354.10947	630943.4893	USNS Raymond
80175.75416	4.988407758	6.104512439	4.04	26748.42842	1068138.074	06039 Jennifer Islands Apt. 443
64698.46343	6.025335907	8.147759585	3.41	60828.24909	1502055.817	4759 Daniel Shoals Suite 442
78394.33928	6.989779748	6.620477995	2.42	36516.35897	1573936.564	972 Joyce Viaduct
59927.66081	5.36212557	6.393120981	2.3	29387.396	798869.5328	USS Gilbert
81885.92718	4.42367179	8.167688003	6.1	40149.96575	1545154.813	Unit 9446 Box 0958
80527.47208	8.093512681	5.0427468	4.1	47224.35984	1707045.722	6368 John Motorway Suite 700
50593.6955	4.496512793	7.467627404	4.49	34343.99189	663732.3969	911 Castillo Park Apt. 717
39033.80924	7.671755373	7.250029317	3.1	39220.36147	1042814.098	209 Natasha Stream Suite 961
73163.66344	6.919534825	5.993187901	2.27	32326.12314	1291331.518	829 Welch Track Apt. 992
69391.38018	5.344776177	8.406417715	4.37	35521.29403	1402818.21	PSC 5330, Box 4420
73091.86675	5.443156467	8.517512711	4.01	23929.52405	1306674.66	2278 Shannon View
79706.96306	5.067889591	8.219771123	3.12	39717.81358	1556786.6	064 Hayley Unions
61929.07702	4.788550242	5.097009554	4.3	24595.9015	528485.2467	5498 Rachel Locks
63508.1943	5.94716514	7.187773835	5.12	35719.65305	1019425.937	Unit 7424 Box 2786
62085.2764	5.739410844	7.091808104	5.49	44922.1067	1030591.429	19696 Benjamin Cape
86294.99909	6.62745694	8.011897853	4.07	47560.77534	2146925.34	030 Larry Park Suite 665
60835.08998	5.551221592	6.517175038	2.1	45574.74166	929247.5995	USNS Brown
64490.65027	4.21032287	5.478087731	4.31	40358.96011	718887.2315	95198 Ortiz Key
60697.35154	6.170484091	7.150536572	6.34	28140.96709	743999.8192	9003 Jay Plains Suite 838
59748.85549	5.339339881	7.748681606	4.23	27809.98654	895737.1334	24282 Paul Valley
56974.47654	8.287562194	7.312879971	4.33	40694.86951	1453974.506	61938 Brady Falls

DATA PREPROCESSING:

Data preprocessing is a crucial step in any machine learning project. We will clean, transform, and normalize the data to ensure its quality and compatibility with machine learning algorithms. This process includes handling missing values, outliers, and categorical variables.

EXPOLATORY DATA ANALYSIS :

Exploratory Data Analysis (EDA) is an essential first step in any data-driven project, such as predicting house prices using machine learning. EDA involves exploring and understanding the dataset's characteristics to reveal insights that inform subsequent data preprocessing and model development.

During EDA, we load and inspect the dataset to grasp its structure. Visualizations like histograms, box plots, and scatter plots help visualize data distributions and relationships. EDA also involves identifying missing values and their patterns. Feature analysis assesses how features relate to the target variable (house prices), while outlier detection helps spot anomalies. Statistical tests may evaluate relationships between categorical variables and house prices.

REGRESSION:

In this project, we will employ regression analysis, such as linear regression or ensemble techniques like random forests and gradient boosting, to model the relationship between various property features and house prices. Regression will enable us to predict house prices based on historical data, providing valuable insights for the real estate market.

REGRESSION TECHNIQUE:

Linear Regression:

- Basic regression method assuming a linear relationship between features and house prices.
- Simple and interpretable.
- Suitable when there's a linear or near-linear relationship between predictors and the target variable.

Polynomial Regression:

- Extends linear regression by introducing polynomial terms to model non-linear relationships.
- Appropriate when the relationship between features and house prices is curvilinear.

Decision Tree Regression:

- Utilizes decision trees to capture non-linear relationships.
- Effective for complex, non-linear datasets.
- Prone to overfitting but can be mitigated with ensemble methods.

Random Forest Regression:

- Ensemble technique that combines multiple decision trees for improved accuracy and reduced overfitting.
- Robust and versatile, suitable for various data types.

Gradient Boosting Regression:

- Ensemble method that builds multiple decision trees sequentially to correct errors of the previous trees.
- Achieves high predictive accuracy.
- Popular algorithms include Gradient Boosting, XGBoost, and LightGBM.

Support Vector Regression (SVR):

- Uses support vector machines to find a hyperplane that best fits the data.
- Effective for small to medium-sized datasets with non-linear relationships.

Bayesian Regression:

- Applies Bayesian statistical techniques to regression analysis, allowing for uncertainty estimation in predictions.
- Useful when you want to incorporate prior beliefs into the regression model.

XG BOOST:

XGBoost, an abbreviation for Extreme Gradient Boosting, is an exceptionally powerful and versatile regression technique that can be highly effective for your house price prediction project. It's an ensemble learning method that sequentially builds decision trees, correcting errors of the previous trees.

Some commonly used regression algorithms are Linear Regression and DecisionTrees. There are several metrics involved in regression like root-mean-squared error (RMSE) and mean-squared-error (MAE). These are some key members of XGBoost models, each plays an important role.

RMSE: It is the square root of mean squared error (MSE).

MAE: It is an absolute sum of actual and predicted differences, but it lacks mathematically, that’s why it is rarely used, as compared to other metrics.

XGBoost is a powerful approach for building supervised regression models.The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners.

MODEL EVALUATION:

Splitting the Data:

Divide your dataset into training and testing sets or use techniques like cross-validation to ensure that your model's performance is assessed on unseen data. Common splits include 70/30 or 80/20 for training/testing.

Performance Metrics:

Choose appropriate evaluation metrics based on the nature of your regression task. Common metrics for house price prediction include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2). MAE and RMSE measure prediction accuracy, while R2 quantifies the variance explained by your model.

Overfitting and Underfitting:

Evaluate whether your model suffers from overfitting (fitting the training data too closely) or underfitting (oversimplification). Overfitting may be indicated by a significant gap between training and testing performance. Adjust model complexity and hyperparameters as needed to address these issues.

Visualization:

Visualize your model's predictions vs. actual values. Scatter plots, residual plots, and prediction vs. actual value plots can help you understand how well your model is performing and where it may be making errors.

Business Impact:

Consider the practical implications of your model's performance. How does its accuracy or error translate into real-world decision-making? Assess whether the model meets the requirements and expectations of your project stakeholders, such as homebuyers, sellers, or real estate professionals.

FACTORS THAT AFFECT HOUSE PRICING:

Unemployment:

Unemployment affects housing affordability. Rising unemployment reduces home affordability, and the fear of job loss can deter participation in the property market.

Economic growth:

Economic growth drives housing demand through income levels. When incomes rise during growth periods, people allocate more to housing, boosting demand and prices. Conversely, in recessions, falling incomes reduce affordability, potentially causing mortgage defaults and repossessions. Income elasticity impacts housing demand as income fluctuations affect housing expenditure ratios.

Market Trends:

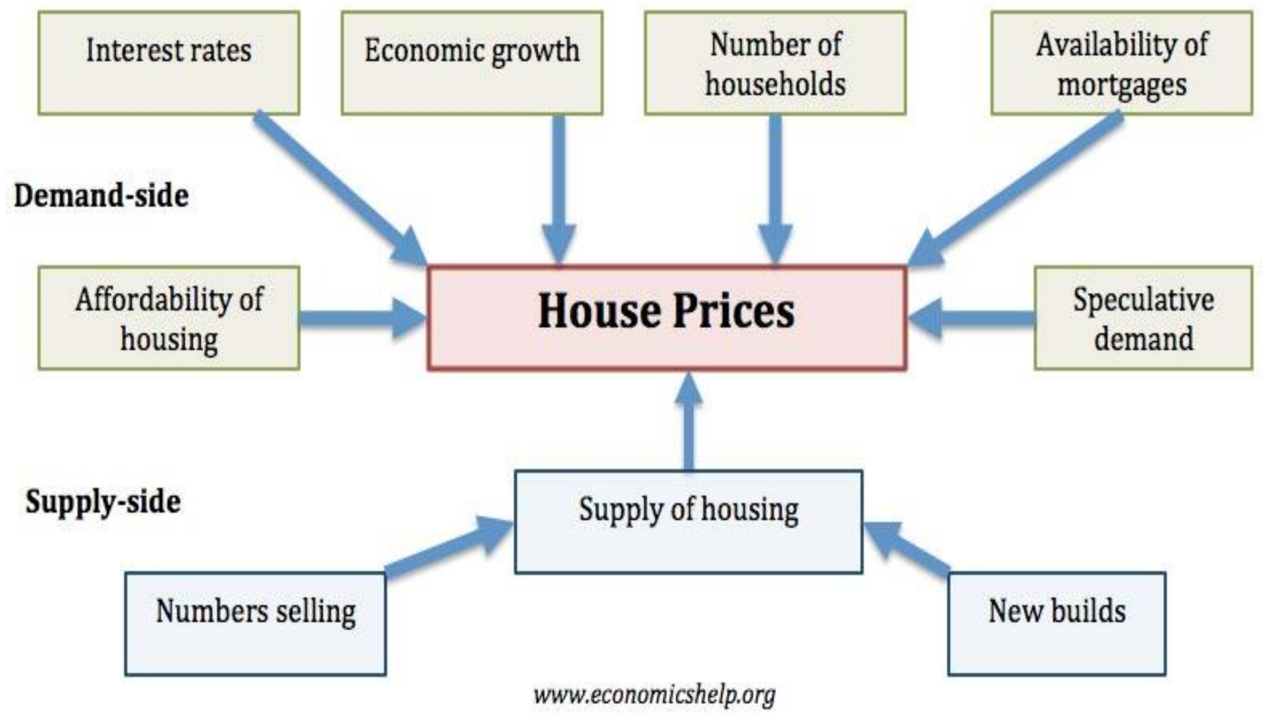
Real estate market conditions, such as supply and demand, influence house prices.

Historical Data:

Past property sales data can reveal trends and guide predictions.

Local Regulations:

Zoning laws, property taxes, and regulations can affect prices.



Interest rates:

Interest rates affect the cost of monthly mortgage payments. A period of high- interest rates will increase cost of mortgage payments and will cause lower demand for buying a house.

CONCLUSION:

We'll recap the significant findings and insights obtained through advanced regression techniques, emphasizing their contributions to enhancing house price prediction accuracy and reliability. We'll also explore potential future directions, including the integration of supplementary data sources like real-time economic indicators, the investigation of deep learning models for prediction, and the expansion of the project's scope.