**PREDICTING HOUSE PRICE USING MACHINE LEARNING**

**REG NO:420721104059 NAME : YOGESWARI E**

**Phase 2 Submission Document**

**ABSTRACT:**

Machine learning plays a major role from past years in image detection, Spam recognition, normal speech command, product recommendation and medical diagnosis along it provides better customer service and safer automobile systems. This shows that ML is trend in almost all fields so we try to coined up ML in our project for betterment. Nowadays, people looking to buy a new home tend to be more conservative with their budgets and market strategies. The current systems main disadvantage is that the calculation of house prices are done without the necessary prediction about future market trends and price increase. The goal of the project is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. In the present paper we discuss about the prediction of future housing prices that is generated by machine learning algorithm. In-order to select the prediction methods we compare and explore various prediction methods. To predict the future price, the previous market trends, price ranges and also upcoming development will be analyzed. Every year House prices increase , so there is a need for a system to predict house prices in the future. We create a housing cost prediction model in view of Machine Learning algorithm models such as Lasso Regression, Ridge Regression, Ada-Boost Regression, XGBoost Regression, Decision Tree Regression, Random Forest Regression. House price prediction on a data set has been done by using all the above mentioned techniques to find out the best among them. The developer and customer will be benefited by this model on determining the selling price of a house and helps the latter to arrange the right time to purchase a house.

**Keywords:** House Price Prediction, Machine Learning, Regression



**INTRODUCTION:**

* The real estate market is one of the most exciting and lucrative industries, with housing prices continually varying depending on location, size, amenities, and economic situations, among other variables. For both buyers and sellers, being able to estimate house prices accurately is essential since it can aid in making decisions about purchasing, selling, or investing in real estate.
* Traditional linear regression models are frequently used to predict home prices. However, they might not fully account for intricate connections between predictors and the target variable, producing predictions that are less than ideal. In order to improve the reliability and accuracy of house price prediction models, we will investigate advanced regression approaches in this project.
* Provide a succinct overview of the real estate market and the value of precise house price forecasting.

Highlight the limitations of traditional linear regression models in capturing complex relationships

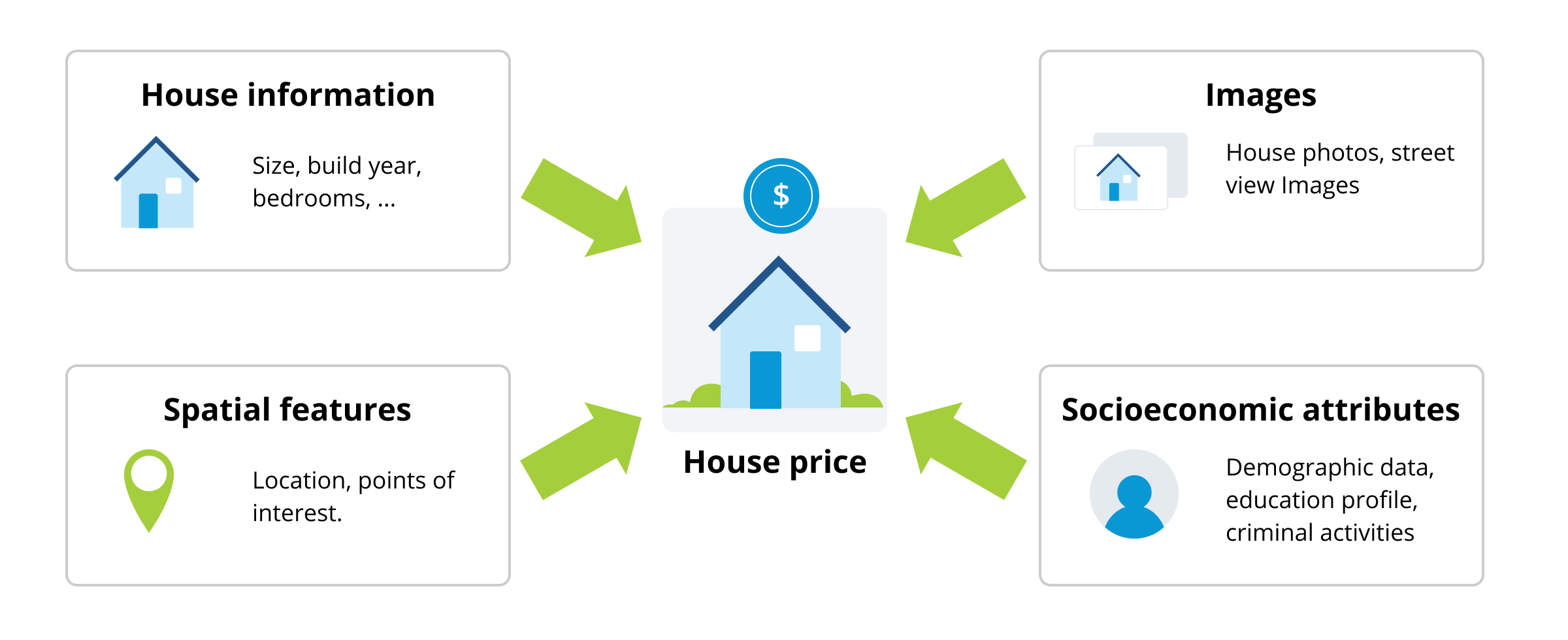
* Emphasize the necessity for sophisticated regression approaches like Gradient Boosting and XGBoost to boost prediction accuracy.

**DATA SOURCE:**

A good data source for house price prediction using machine learning should be Accurate , Complete ,Covering the geographic area of interest, Accessible.

Dataset Link :([**https://www.kaggle.com/datasets/vedavyasv/usa-housing**](https://www.kaggle.com/datasets/vedavyasv/usa-housing))

**REAL ESTATE PRICING WITH MACHINE LEARNING:**



**DATA COLLECTION AND PREPROCESSING:**

* **Importing the dataset:** Obtain a comprehensive dataset containing relevant features such as square footage, number of bedrooms, location , amenities , etc.
* **Data preprocessing :** Clean the data by handling missing values , outliers , and categorical variables. Standardize or normalize numerical features.

**EXPLORATORY DATA ANALYSIS (EDA):**

* Visualize and analyze the dataset to learn more about the connections between the variables.
* Identify connections and trends that can guide feature engineering and selection.
* Present various data visualizations to gain insights into the dataset.
* Investigate the relationships between characteristics and the goal variable (house prices).
* Discuss any important results from the EDA phase that have an impact on feature choice.

**FEATURE ENGINEERING:**

* To collect important data, develop new features or alter current ones.
* Utilize domain knowledge to engineer aspects, such as closeness to transportation, schools, or crime rates, that may have an impact on home pricing.
* Describe the steps involved in adding new features or changing current ones.
* Showcase the development of domain-specific features, like proximity scores or composite indicators.
* Emphasize the impact of engineered features on model performance.

**ADVANCED REGRESSION TECHNIQUES:**

**Ridge Regression:** Ridge regression, also known as L2 regularization or L2 regularization, is a linear regression technique used in statistics and machine learning. It is an extension of ordinary least squares (OLS) regression and is designed to address some of its limitations, particularly when dealing with multicollinearity, overfitting, and model instability.

**Lasso Regression:** Another widely used method in linear regression to handle multicollinearity and carry out feature selection is lasso regression, also known as "Least Absolute Shrinkage and Selection Operator" regression. It adds a regularization element to the linear regression cost function, similar to ridge regression, although it employs L1 regularization rather than L2.

**Elastic Net Regression:** Elastic Net regression is a linear regression technique that combines the properties of both Lasso (L1 regularization) and Ridge (L2 regularization) regression. It is designed to address some of the limitations of each of these techniques while incorporating their strengths. Elastic Net introduces a new hyperparameter, α (alpha), which allows you to control the mix of L1 and L2 regularization.

**Random Forest Regression:** Random Forest Regression is a machine learning technique that extends the concept of decision trees to perform regression tasks. It is based on the Random Forest algorithm, which is an ensemble learning method. While decision trees are used for both classification and regression, Random Forest Regression specifically focuses on predicting continuous numerical values rather than discrete classes.

**Gradient Boosting Regressors (e.g., XG Boost, Light GBM):** Gradient Boosting Regressors, such as XGBoost, LightGBM, and other variants, are powerful machine learning techniques that have gained popularity for their effectiveness in various predictive modeling tasks, including house price prediction. These algorithms belong to the gradient boosting family and are particularly well-suited for regression problems. Here's how you can use gradient boosting regressors in the context of house price prediction:

**1. Data Preprocessing:**

* Begin by preparing and cleaning your dataset. This includes handling missing values, encoding categorical features, and scaling numerical features.
* Split your data into training and testing sets. Cross-validation can also be useful for hyperparameter tuning.

**2. Feature Engineering:**

* Identify and create relevant features that may influence house prices, such as the number of bedrooms, square footage, neighborhood, proximity to amenities, and more.
* Interaction terms or transformations of features to capture non-linear relationships.

**3. Selecting the Algorithm:**

Choose a gradient boosting regressor algorithm such as XGBoost, LightGBM, or CatBoost. These libraries offer efficient and optimized implementations for gradient boosting.

**4. Hyperparameter Tuning:**

* Tune the hyperparameters of the chosen algorithm. Common hyperparameters include the learning rate, number of trees (boosting rounds), tree depth, and regularization terms. Hyperparameter tuning can be performed using techniques like grid search or random search.
* Ensure you optimize the parameters to avoid overfitting and underfitting. This often involves finding the right balance between model complexity and predictive accuracy.

**5. Model Training:**

Train the selected gradient boosting regressor on your training data using the optimal hyperparameters. The model will iteratively fit decision trees to the residuals to reduce prediction errors.

**6. Model Evaluation:**

Evaluate the model on the test data using appropriate regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), or R-squared (R²). These metrics provide insights into how well the model predicts house prices.

**7. Feature Importance Analysis:**

Many gradient boosting libraries offer feature importance scores, which can help you identify the most influential features in predicting house prices.

**8. Ensemble Methods:**

Consider using ensemble techniques, such as stacking or blending, to combine predictions from multiple gradient boosting regressors or other regression models. This can further enhance prediction accuracy.

**9. Regular Monitoring and Updating:**

House price prediction models should be regularly updated with new data. Housing markets can change over time, so it's crucial to ensure your model remains accurate and relevant.

Both XGBoost and LightGBM, as well as other gradient boosting implementations, are capable of handling complex, non-linear relationships, and they perform well with high-dimensional data. When applied correctly, these algorithms can provide highly accurate predictions for house prices and are widely used in real estate and finance industries.

**MODEL EVALUATION AND SELECTION:**

* Split the dataset into training and testing sets.
* Evaluate models using appropriate metrics (e.g., Mean Absolute Error, Mean Squared Error, R-squared) to assess their performance.
* Use cross-validation techniques to tune hyperparameters and ensure model stability. Compare the results with traditional linear regression models to highlight improvements.
* Select the best-performing model for further analysis.

**MODEL INTERPRETABILITY:**

* Describe the interpretation of feature importance in the Gradient Boosting and XG Boost models.
* Discuss the insights gained from feature importance analysis and their relevance to house price prediction.
* Interpret feature importance from ensemble models like Random Forest and Gradient Boosting to understand the factors influencing house prices.

**DEPLOYMENT AND PREDICTION:**

* Deploy the chosen regression model to predict house prices.
* Develop a user-friendly interface for users to input property features and receive price predictions.

**CONCLUSION:**

* We will highlight the major discoveries and critical takeaways from the advanced regression approaches in the conclusion of Phase 2. We will restate how these strategies have improved the reliability and accuracy of house price projections.
* **Further Work:** In this section, we'll talk about possible directions for further research, such adding other data sources (like real-time economic indicators), investigating deep learning models for prediction, or turning the project into a web application with greater features and interactivity**.**