
INNOVATION IN HOUSE PRICE PREDICTION (PHASE 2)

BY:
G.KANIMOSHI
MAIL.ID:kanimoshigopal2004@gmail.com
NM ID:au51132104040
BE-CSE 3RD YEAR

INNOVATION IN HOUSE PRICE PREDICTION (PHASE 2)

INTRODUCTION

- We will begin by revisiting the problem definition.
- Our goal is to predict house prices using machine learning

PROBLEM DEFINITION:

- The problem is to predict house prices using machine learning techniques. The objective is to develop a model that accurately predicts the prices of houses based on a set of features such as location, square footage, number of bedrooms and bathrooms, and other relevant factors. This project involves data preprocessing, feature engineering, model selection, training, and evaluation.
- In Phase 1, we designed a framework involving data preprocessing, feature engineering, model selection, training, and evaluation

RECAP OF PHASE 1

- We have already seen in the previous presentation that we selected our dataset, cleaned and preprocessed the data, handled missing values, and converted categorical features into numerical representations.
- We have performed feature selection, choosing the most relevant features, and selected a suitable regression algorithm for prediction.
- Our model has been trained using the preprocessed data, and we have evaluated its performance using metrics like MAE, RMSE, and R-squared.

DESIGN THINKING - PHASE 1

- **Objective:** Accurate house price prediction using machine learning.
- **Phases:**
 - Data Preprocessing
 - Feature Selection
 - Model Selection
 - Model Training
 - Evaluation
- **Tools/Modules:** Pandas, NumPy, Scikit-Learn, Matplotlib for data manipulation, modeling, and evaluation.
- **Key Deliverables:** Cleaned dataset, selected features, trained models, evaluation metrics (MAE, RMSE, R-squared).
- **Outcomes:** Baseline models for house price prediction.

DATA PREPROCESSING

- Data preprocessing will be the initial step in the house price prediction project. It will involve cleaning and preparing the raw dataset for modeling. This will include handling missing values, dealing with outliers, and transforming data into a suitable format. For example, converting categorical features like location etc. into numerical representations and normalizing numerical variables.
- **Importance** : Proper data preprocessing will be critical as it will ensure the dataset is clean and ready for analysis, which will be essential for building accurate predictive models.

FEATURE SELECTION

- Feature selection will be the process of choosing a subset of the most relevant features (attributes) from the dataset. In the context of house price prediction, features might include square footage, the number of bedrooms, location, etc. Feature selection will help in reducing the dimensionality of the dataset and will focus on the attributes that have the most impact on predicting house prices.
- **Importance** : Selecting the right features will not only simplify the model but will also improve its predictive accuracy by focusing on the most influential factors.

MODEL SELECTION

- Model selection will involve choosing an appropriate machine learning algorithm for the task. In the case of house price prediction, regression algorithms will typically be used. Common choices will include Linear Regression, Decision Trees, Random Forests, Gradient Boosting, and XGBoost. The selection will be based on the characteristics of the dataset and the problem at hand.
- **Importance** : The choice of the model will significantly impact prediction accuracy. Selecting the right model will be crucial to obtaining accurate house price estimates

MODEL TRAINING

- Model training will be the process of fitting the chosen machine learning algorithm to the preprocessed dataset. During this phase, the model will learn from the data by adjusting its parameters to minimize the prediction errors. The training will involve splitting the dataset into a training set and a testing set for validation.
- **Importance** : Proper model training will be necessary for the model to learn from the data and make accurate predictions. It will be a crucial step in building an effective house price prediction model.

EVALUATION

- Evaluation will be the phase where the performance of the trained model will be assessed using appropriate metrics. Common evaluation metrics for house price prediction will include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. The evaluation results will provide insights into how well the model predicts house prices.
- **Importance** : Evaluation will help measure the model's accuracy and effectiveness. It will determine whether the model can be trusted for real-world applications, such as pricing properties in the housing market.
- These steps, when executed effectively, will form a structured and logical process for the house price prediction project. They will ensure that the data is properly handled, the model is selected and trained appropriately, and its performance is rigorously assessed, resulting in reliable and accurate price predictions.

THE NEED FOR INNOVATION

- As we embark on Phase 2 of our house price prediction project, it's essential to recognize that the real estate market is constantly evolving.
- To stay competitive and provide accurate pricing predictions, we must continuously explore innovative techniques in the field of machine learning.
- Innovation is the key to gaining a competitive edge and meeting the ever-changing demands of the real estate industry.
- In this phase, we will not only seek improved predictive accuracy but also aim to enhance our understanding of the underlying dynamics that influence house prices.

ADVANCED REGRESSION TECHNIQUES

- We understand that traditional linear regression models, while valuable, may not capture the complexity of real estate markets.
- Therefore, we will delve into advanced regression techniques that have proven effective in handling intricate relationships within housing data.
- Prominent methods we will explore are Gradient Boosting , XGBoost etc., which belong to the ensemble learning family.
- Ensemble methods combine multiple models to make collective predictions, often outperforming individual models.

THE POWER OF ENSEMBLE LEARNING

- Ensemble learning is a powerful concept in machine learning, which leverages the wisdom of crowds.
- By combining multiple models, we aim to mitigate individual model weaknesses and achieve superior predictive accuracy.
- Ensemble techniques, like Gradient Boosting and XGBoost, excel in capturing nuances within data, including non-linearity and interactions among features.
- Through this approach, we will harness the collective intelligence of multiple models to make more accurate house price predictions

EXPECTATIONS FROM ADVANCED TECHNIQUES

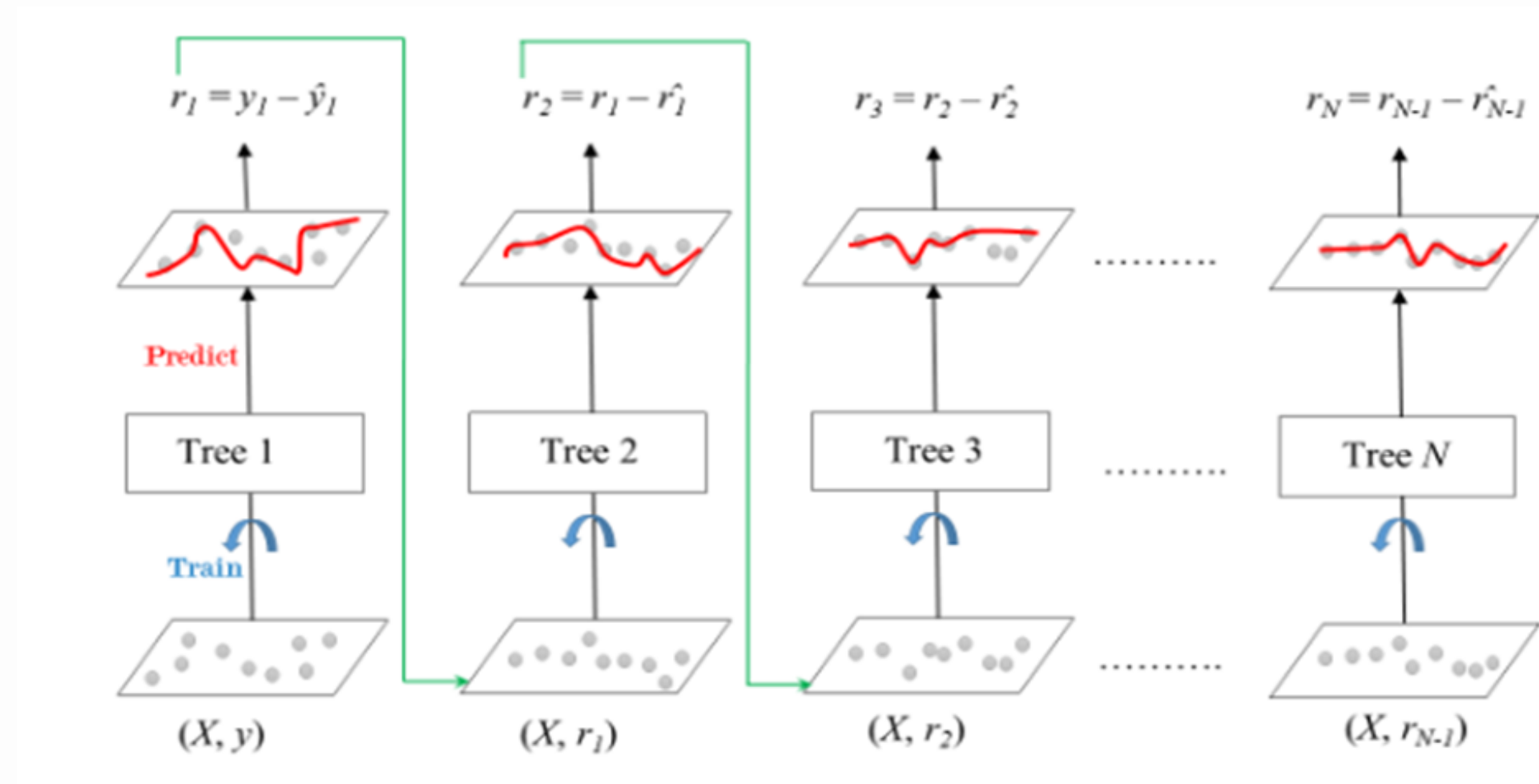
- Our expectations from exploring advanced regression techniques include:
- Increased precision in predicting house prices, reducing the margin of error.
- Improved ability to identify complex patterns, such as local market trends and outliers.
- Enhanced model robustness, making our predictions more reliable in varying real estate market conditions.
- Greater insight into the factors that most strongly influence house prices

PREPARING FOR THE PROJECT

- As we embark on this innovative journey, we must ensure we have a strong foundation
- High-quality, well-preprocessed data.
- A clear understanding of our evaluation metrics (e.g., MAE, RMSE, R-squared).
- A structured approach to model selection and training.
- A willingness to iterate and refine our models based on results.

1. GRADIENT BOOSTING

- Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.
- In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of the predecessor as labels. There is a technique called the **Gradient Boosted Trees** whose base learner is CART (Classification and Regression Trees). The below diagram explains how gradient-boosted trees are trained for regression problems



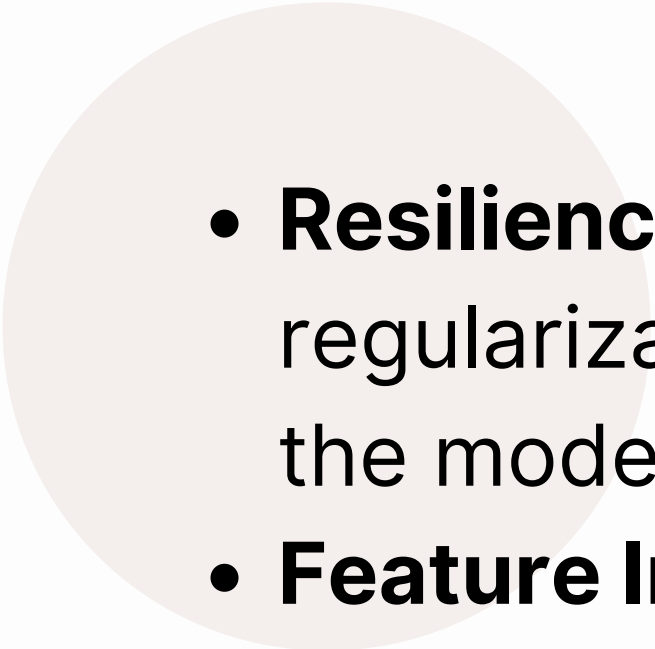
The ensemble consists of M trees. Tree1 is trained using the feature matrix X and the labels y . The predictions labeled \hat{y}_1 are used to determine the training set residual errors r_1 . Tree2 is then trained using the feature matrix X and the residual errors r_1 of Tree1 as labels. The predicted results \hat{r}_1 are then used to determine the residual r_2 . The process is repeated until all the M trees forming the ensemble are trained. There is an important parameter used in this technique known as **Shrinkage**.

SHRINKAGE IN GRADIENT BOOSTING

- **Shrinkage** refers to the fact that the prediction of each tree in the ensemble is shrunk after it is multiplied by the learning rate (eta) which ranges between 0 to 1. There is a trade-off between eta and the number of estimators, decreasing learning rate needs to be compensated with increasing estimators in order to reach certain model performance. Since all trees are trained now, predictions can be made. Each tree predicts a label and the final prediction is given by the formula,
 - $y(\text{pred}) = y_1 + (\text{eta} * r_1) + (\text{eta} * r_2) + \dots + (\text{eta} * r_N)$

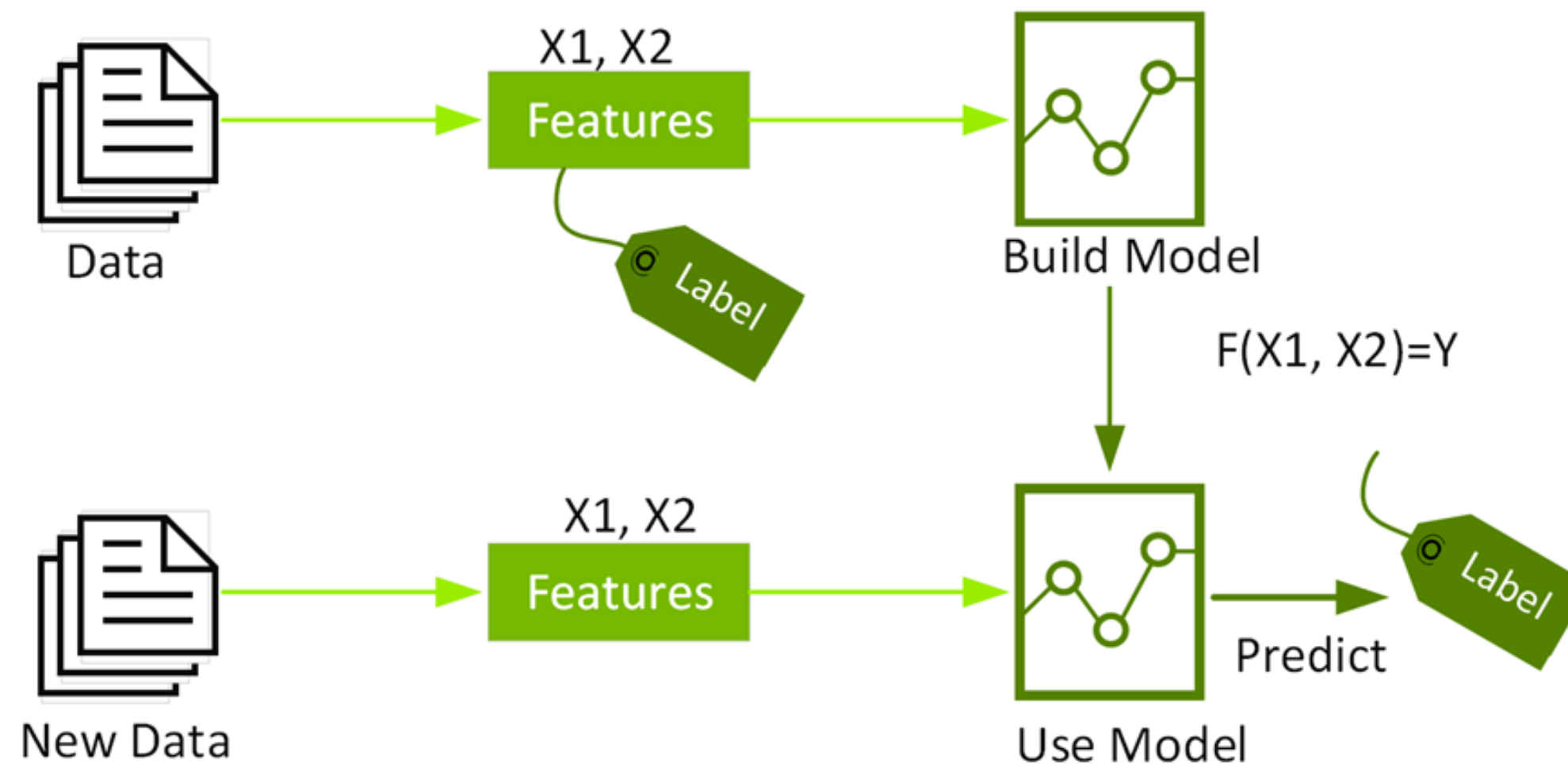
BENEFITS OF GRADIENT BOOSTING.

- **Improved Accuracy:** Gradient Boosting is known for its impressive predictive accuracy. By iteratively correcting errors in predictions, it fine-tunes the model and brings it closer to the true relationship between features and the target variable.
- **Handling Complex Relationships:** Real-world data often contains intricate and nonlinear relationships. Gradient Boosting is well-equipped to capture these complexities, making it suitable for the multifaceted nature of house price prediction. It excels in modeling intricate market dynamics and local trends.

-
- 
- **Resilience to Over fitting:** Gradient Boosting incorporates techniques like regularization and shrinkage, which help prevent over fitting. This ensures that the model generalizes well to unseen data, enhancing its reliability.
 - **Feature Importance:** Gradient Boosting provides insights into feature importance. We can assess which features have the most significant impact on house prices, aiding in feature selection and understanding market dynamics.
 - **Robustness:** It is robust against outliers and noisy data, making it a robust choice for real-world datasets, which often have inconsistencies.
 - Gradient Boosting is a powerful ensemble method that sequentially combines weak learners to create a robust and highly accurate predictive model. Its ability to handle complex relationships and deliver superior performance makes it a valuable tool for house price prediction in dynamic real estate markets

2. XGBOOST (EXTREME GRADIENT BOOSTING)

- XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.
- It's vital to an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon: supervised machine learning, decision trees, ensemble learning, and gradient boosting.
- Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset's features



XGBoost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energizing machine learning model performance and computational speed. With XGBoost, trees are built in parallel, instead of sequentially like GBDT (Gradient Boosting Decision Trees). It follows a level-wise strategy, scanning across gradient values and using these partial sums to evaluate the quality of splits at every possible split in the training set.

ADVANTAGES OF XGBOOST

- **Efficiency and Scalability:** XGBoost is highly efficient and scalable. It is optimized for speed and can handle large datasets with ease. This efficiency makes it a valuable choice when dealing with substantial real estate datasets containing numerous properties and features.
- **Regularization Techniques:** XGBoost incorporates L1 (Lasso) and L2 (Ridge) regularization techniques. These regularization methods help prevent over fitting by adding penalty terms to the loss function. This makes the model less likely to fit noise in the data, enhancing its generalization ability.
- **High Performance:** XGBoost consistently delivers high predictive performance. It often outperforms other algorithms in various machine learning tasks, including house price prediction. This means that we can expect more accurate price estimates when using XGBoost.

-
- **Robustness to Overfitting:** Thanks to its regularization techniques and built-in capabilities, XGBoost is robust against over fitting, even in the presence of noisy or incomplete data. This ensures that our predictions maintain accuracy and reliability.
 - **Feature Importance:** XGBoost provides a valuable feature importance score, allowing us to understand which features contribute the most to house price predictions. This insight aids in feature selection and market analysis.
 - **Flexibility:** XGBoost can be used for both regression and classification tasks. Its flexibility allows us to adapt it to various aspects of real estate prediction beyond price forecasting, such as property classification or market trend analysis.
 - XGBoost is a versatile and efficient algorithm known for its high performance, scalability, and robustness to over fitting. Its incorporation of regularization techniques makes it particularly well-suited for complex real estate datasets, where accurate predictions and generalization are paramount.

3. ADABOOST

- AdaBoost, short for Adaptive Boosting, is a versatile ensemble learning algorithm that focuses on improving predictive performance by strategically combining multiple weak models.
- **What is AdaBoost?**
- AdaBoost begins with a base model (often a simple decision tree), and it sequentially builds a series of models, each aiming to correct the errors made by the previous ones.
- During each iteration, AdaBoost assigns higher weights to the data points that were misclassified by the previous model, effectively shifting the focus toward difficult-to-learn instances.
- It then combines the predictions of these models, ultimately creating a strong ensemble learner.
- The term "adaptive" reflects its ability to adapt to the changing needs of the problem, emphasizing problematic areas and continuously enhancing the model's predictive power.

ADVANTAGES OF ADABOOST

- **Efficiency and Scalability:** AdaBoost is computationally efficient and scales well to large datasets.
- **Regularization Techniques:** AdaBoost incorporates regularization techniques to prevent over fitting.
- **High Performance:** AdaBoost often leads to high predictive performance.
- **Robustness to Over fitting:** AdaBoost is robust against over fitting.
- **Weighted Model Aggregation:** AdaBoost assigns different weights to individual models based on their performance.
- **Adaptive Learning:** AdaBoost can adapt to difficult-to-learn data points.
- **Versatility:** AdaBoost is a versatile algorithm that can be used with various base learners.
- AdaBoost's adaptability and sequential learning process make it a valuable tool in improving predictive accuracy, particularly in domains where the data landscape is intricate and ever-evolving, such as house price prediction.

4. NEURAL NETWORKS(DEEP LEARNING)

- Neural Networks, a subset of deep learning, are a class of algorithms inspired by the human brain's structure and function. They offer unique capabilities that are advantageous for house price prediction.
- **Key Features of Neural Networks (Deep Learning):**
 - **Complex Pattern Recognition:** Neural networks excel at capturing intricate and non-linear patterns within the data, making them well-suited for understanding the nuances of real estate markets.
 - **Feature Extraction:** Neural networks automatically extract relevant features from diverse data types, facilitating improved prediction by reducing the need for extensive feature engineering.
 - **Adaptability:** Neural networks can process various data types, including images and structured data, providing a holistic approach to data analysis.
 - **Ensemble Integration:** They can be seamlessly integrated into ensemble techniques, enhancing overall prediction accuracy by combining their strengths with other models.
 - **Complexity Management:** Careful model architecture design and hyper parameter tuning are essential to efficiently manage the complexity of neural networks for accurate predictions in real estate scenarios.

ADVANTAGES OF NEURAL NETWORKS

- **Pattern Recognition:** They excel at capturing complex and non-linear patterns in real estate data, allowing us to identify intricate relationships and market trends.
- **Feature Extraction:** Neural networks automatically extract relevant features, reducing the need for extensive manual feature engineering. This ability is particularly valuable in the diverse data landscape of real estate, encompassing images, text descriptions, and structured data.
- **Data Adaptability:** They can process various data types, making them versatile for different aspects of real estate prediction. For example, Convolutional Neural Networks (CNNs) are well-suited for property image analysis, while Recurrent Neural Networks (RNNs) can handle sequential data, such as time series of house prices.
- **Ensemble Capability:** Neural networks integrate seamlessly with ensemble techniques, enhancing overall prediction accuracy by combining their strengths with other models. This versatility allows us to harness the collective intelligence of multiple models.
- **Challenges in Model Complexity:** Managing model complexity and computational resources is crucial for efficient and accurate predictions in real estate scenarios.

5. LASSO REGRESSION

- Lasso Regression, an abbreviation for "Least Absolute Shrinkage and Selection Operator," is a versatile and powerful regression technique that is particularly well-suited for house price prediction tasks. It has several distinctive features that set it apart from other regression methods.
- **Key Features of Lasso Regression:**
- **Sparsity and Feature Selection:** Lasso encourages some coefficients to be exactly zero, automatically selecting the most influential features, making it well-suited for real estate datasets with numerous attributes.
- **Regularization for Enhanced Generalization:** Lasso employs L1 regularization to prevent overfitting and ensure the model generalizes effectively to unseen property data.
- **Balancing Complexity:** Lasso strikes a balance between model complexity and accuracy, creating a more interpretable yet effective representation of the relationship between features and house prices.
- **Handling Multi collinearity:** Lasso addresses multi collinearity by selecting the most relevant features, reducing redundancy, and maintaining model integrity.
- **Interpretability and Transparency:** Lasso's simplicity results in highly interpretable models, providing insights into the factors influencing property prices in the real estate market.

ADVANTAGES OF LASSO REGRESSION

- **Efficiency and Scalability:** Lasso Regression is computationally efficient and scales well, making it suitable for large datasets.
- **Regularization Techniques:** Lasso incorporates L1 regularization, effectively promoting feature selection and mitigating over fitting.
- **High Performance:** It focuses on relevant features, often leading to improved predictive accuracy for house prices.
- **Interpretability:** Lasso produces sparse models, making it easier to interpret the impact of individual features on property prices.
- Lasso Regression efficiently combines regularization with high performance, aiding in feature selection and enhancing model interpretability.



THANK YOU!