# House Price Prediction

## Capstone Presentation

**Prepared and Presented**
*Sudheendra K*
PGPDSBA Oct.B.22
sudhi0404@gmail.com

## Overview

- The problem revolves around predicting house prices accurately by considering various property features

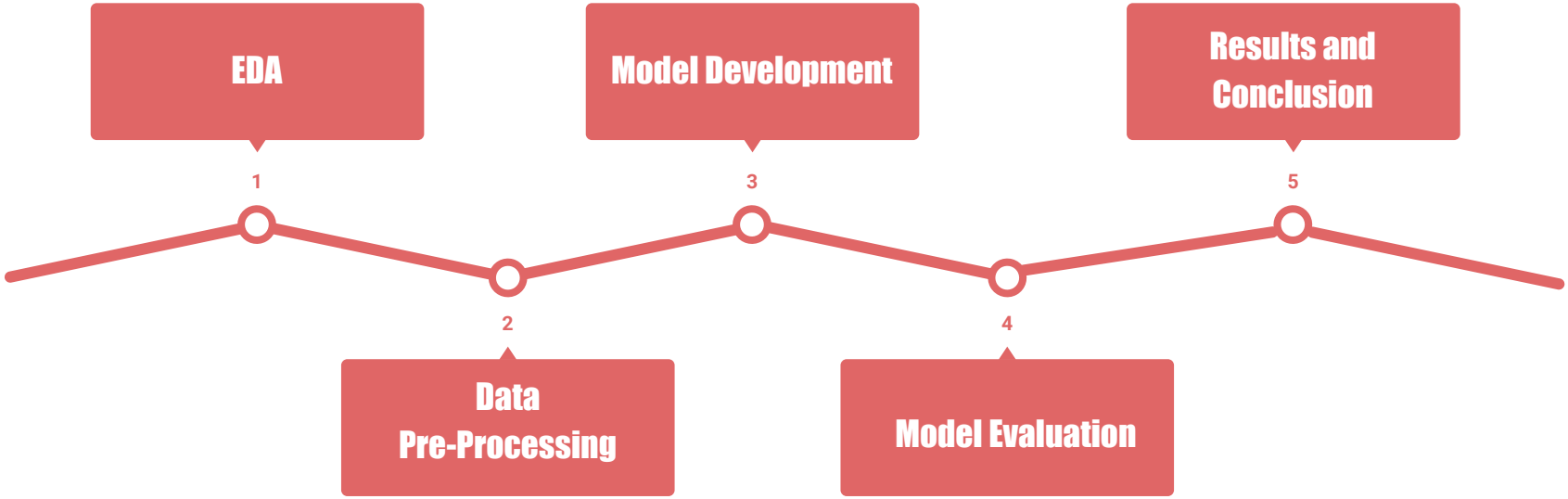
## Constraints

- Imbalanced Dataset:
- Historical Price Data Limitation:
- Limited Data Availability

## Scope and Objective

- The project will strive to maintain model interpretability, enabling users to understand the factors influencing price predictions.
- Develop a machine learning model that can accurately predict house prices based on a diverse set of features, providing precise valuations.

# Modelling Approach



# 1. EDA

### Univariate Analysis

examines a single variable's characteristics, such as its distribution, central tendency, and spread, to gain insights into its individual behavior.

### Bivariate Analysis

explores the relationships and interactions between two variables, aiming to uncover correlations, associations, or dependencies between them.

### Exploratory Data Analysis

These cross-tabulations and summary statistics is for valuable understanding of patterns and associations among different features, which can be useful for further analysis and decision making

# 2. Data Pre-Processing

### Data Cleaning

The raw data obtained from various sources is often incomplete or contain errors and this is the process to identify and correcting, ensuring the data is reliable for analysis

### Feature Engineering

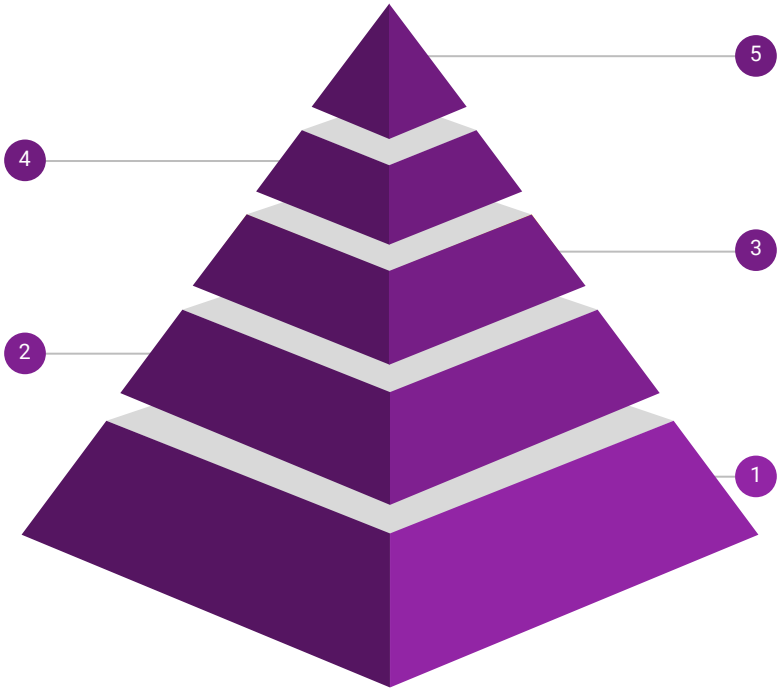
Its process of curating or selecting the most relevant features from the data which helps in reducing the curse of dimensionality using various techniques such PCA.

### Data Transformation

It is a process of converting the data into suitable format for further steps such as Normalization, standardization / Scaling

**Data Pre-Processing**

# 3. Model Development



5 Model Evaluation

Training and Testing the Model 4

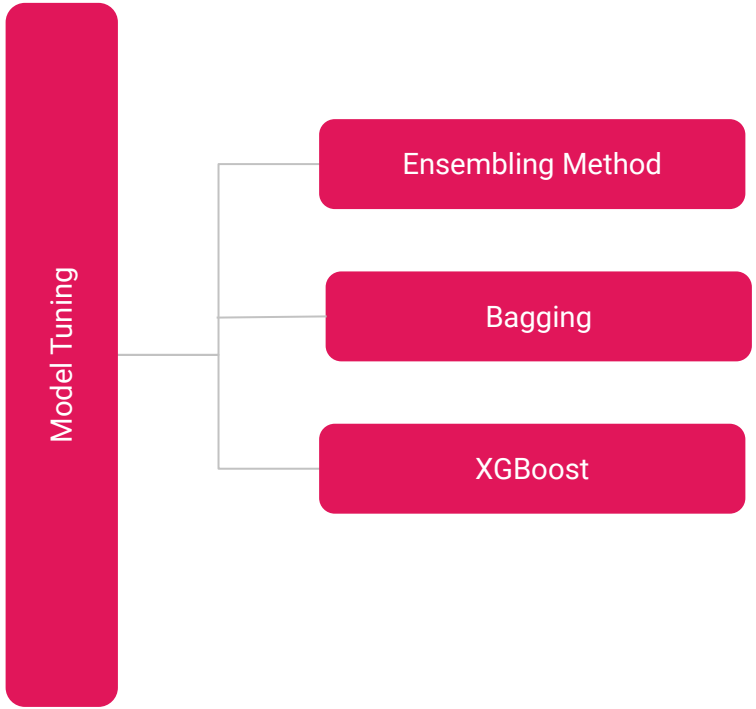
3 Model Selection

Data Splitting 2

1 Selecting target variable

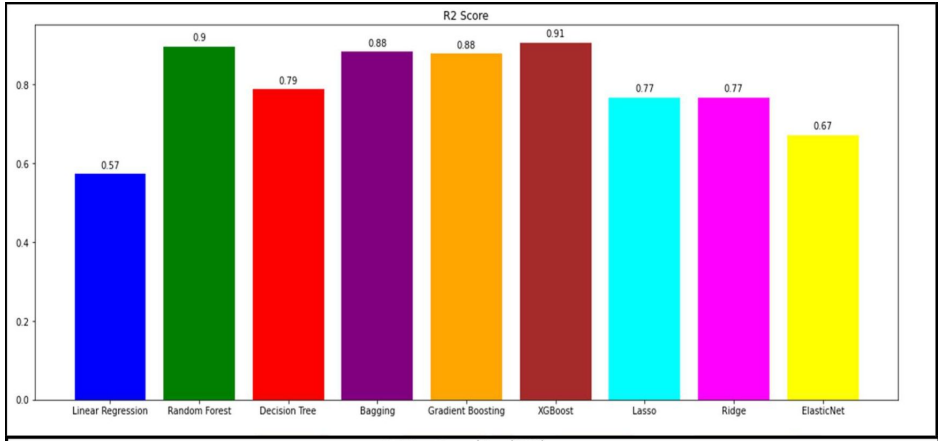
# 3. Model Selection and Tuning

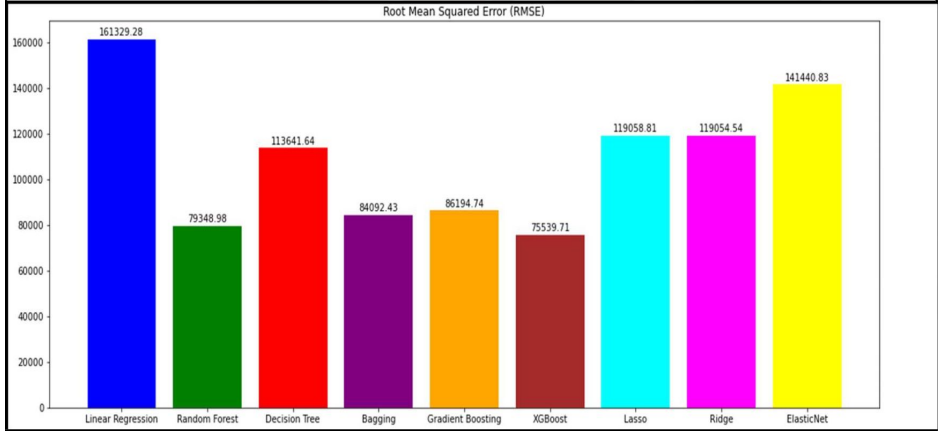
- Linear Regression
- Random Forest
- Decision Tree
- Lasso
- Ridge
- ElasticNet



# Model Evaluation

- XGBoost outperforms other models across MAE, MSE, RMSE, and R2 Score.
- Achieves low MAE and RMSE, crucial for precise house price predictions.
- Significantly captures house price variability with a low MSE.
- Strong explanatory power with an R2 score of 0.91.



# Insights from Analysis

- Comprehensive Analysis:Numerical and categorical attributes totaling 23 columns and 21,613 data points from 2014 to 2015
- Descriptive Statistics:Special focus on the 'price' attribute, revealing right-skewed distribution and outliers
- Some columns have missing values and anomalies
- There are extreme values or outliers in the response variable. Outliers can indicate data imbalance, especially if they are far from the majority of data points.
- The mean house price is approximately $540,182, while the median price is $450,000. This suggests that the distribution of house prices is right-skewed, as the mean is greater than the median.
- There is an extremely strong positive correlation (0.9998) between the total area (sum of living area and lot size) and the lot size. This is expected since the total area is the sum of its components.
- Notably, living area size, ceiling area size, living area size in 2015, and lot size appear to be positively correlated with house price, indicating their importance in influencing house prices.
- Additionally, the high correlation between living area size and ceiling area size suggests that these attributes are closely related.
- XGBoost outperforms other models across MAE, MSE, RMSE, and R2 Score achieves with strong explanatory power with an R2 score of 0.91.

# Recommendations

- Implement XGBoost for accurate house price predictions due to its superior performance.
- Enhance data quality continually to ensure accuracy and reliability.
- Update the model regularly to keep it relevant in the dynamic real estate market.
- Prioritize transparency and interpretability in the model for user understanding.
- Establish a feedback mechanism to gather user input for ongoing model improvement.
- Monitor the real estate market closely to adapt to changing trends.
- Incorporate Point of Interests (POIs) Data: Consider integrating data on the proximity of essential amenities such as malls, hospitals, transportation hubs, and the overall neighborhood quality. These factors can significantly impact house prices, and their inclusion can enhance prediction accuracy.
- Amenities and Neighborhood Insights: Gather more detailed information on the specific amenities available in the neighborhood and how they influence pricing. Analyze the relationship between amenities, neighborhood characteristics, and house values to enrich your predictive model.
-

# Recommendations

- Futuristic Considerations: Include data on ongoing or planned infrastructure developments in the area. This forward-looking approach can help anticipate potential changes in property values due to upcoming projects or improvements.
- Legal and Historical Data: Investigate the legal aspects of the properties, such as land titles and survey numbers. Additionally, historical price data for the buildings can offer valuable insights into price trends and market dynamics.
- Demographic and Economic Trends: Consider incorporating demographic data and economic indicators that may affect property values, such as population growth, employment rates, pollution metrics, and income levels in the area.