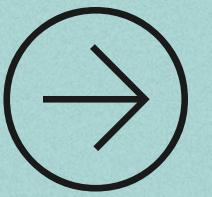




HOUSE PRICE PREDICTION

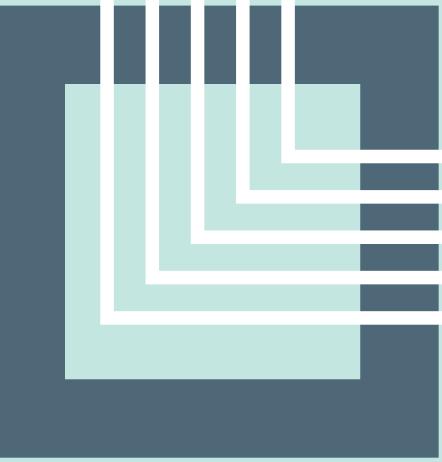
**DR: Essam El-Din El-Fakharany
ENG: Ebrahim Benbella**

Team Members



Mohamed Akram Mohamed	– import the data & presentation
Omar ahmed gaber	–data preparation
Husein Waleed Hussein	–Data Preprocessing
Yusef Atef Ahmed	–features engineering
Abdelrahman Rady Mahmoud	–XGBoost Model
Mohamed Osama Nabil	–Visualization

THE BIG IDEA



Our Main Goal:

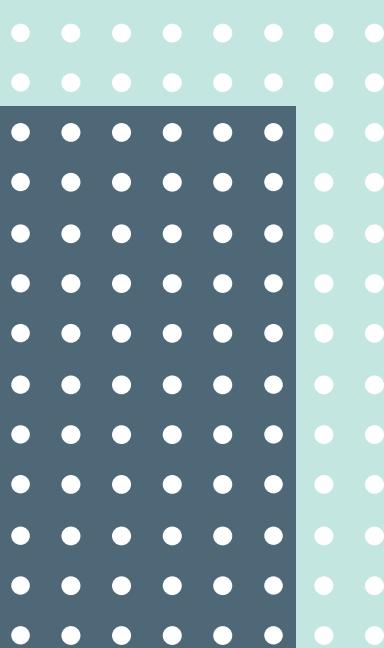
To build an intelligent system using Machine Learning that can accurately predict house prices in California.

Why Is This Important?

- **For Buyers & Sellers:** It helps them make informed decisions and determine fair market prices.
- **For Investors:** It provides valuable insights into the real estate market to maximize return on investment.

Our Methodology

We will use a Linear Regression model to analyze the relationship between different house characteristics (like median income, house age, and number of rooms) and their final price.



DATA & LIBRARIES USED

The Dataset: Ames Housing

- **Source:** Ames Housing Dataset? Ames, Iowa from 2006 to 2010.
Created by Prof. Dean DeCock, on Kaggle.
- **Structure:** It contains two main files: train.csv and test.csv.
 - Training Data: 1460 rows (houses) and 81 features.
 - Target Variable: SalePrice is the column we aim to predict.
- **Features:** A mix of numerical and categorical data describing every aspect of a home, from its size and location to the type of roof and basement quality.

Core Python Libraries

We are leveraging a standard stack of data science libraries to build our model:

- **Pandas:** The primary tool for loading, cleaning, and manipulating the dataset.
- **NumPy:** Used for efficient numerical computations and handling arrays.
- **Matplotlib & Seaborn:** Essential for creating visualizations, exploring data patterns, and analyzing feature relationships.
- **Scikit-learn:** The core machine learning library used for:
 - Splitting the data (train_test_split)
 - Preprocessing (StandardScaler)
 - Implementing the LinearRegression model.

DATA EXPLORATION & CLEANING

Initial Data Exploration & Cleaning

The dataset was explored to understand its structure and quality. The following steps were conducted:

Dataset Dimensions:

Using `.shape`, the number of rows and columns was identified.

This provides an overview of the dataset size and helps in planning preprocessing.

Dataset Information:

With `.info()`, data types of each column and non-null counts were displayed.

This step ensures understanding of numerical vs categorical features and checks for potential type issues.

Missing Values Check:

Using `.isnull().sum()`, missing values across all columns were detected.

Identifying missing data is critical for deciding whether to impute, drop, or otherwise handle them.

Duplicate Records:

The `.duplicated()` function was applied to check for repeated rows.

Removing duplicates prevents bias in model training.

Statistical Summary:

With `.describe()`, descriptive statistics such as mean, median, standard deviation, minimum, and maximum were reviewed.

This helped to spot potential outliers and understand the overall distribution of the data.



XGBOOST REGRESSION MODEL

- Implementing the XGBoost regression model to predict house prices.
- I trained the model on the processed training data and evaluated its performance using R^2 score and Mean Absolute Error (MAE) for both training and test sets.
- Additionally, I visualized the relationship between actual vs. predicted prices to better understand model accuracy
- This work ensured the project had a well-optimized prediction model and clear performance evaluation.



FEATURE RANKING

1-New features engineered:

GarageAge, HasGarage, TotalSF, PorchSF, BathPerRoom,
LivingAreaRatio, GaragePerCar, OverallScore, TotalBath, QualLivingInteraction.

2-Verification: code prints which ones were added successfully.

3-Ranking Approaches:

Correlation with SalePrice (linear insight).

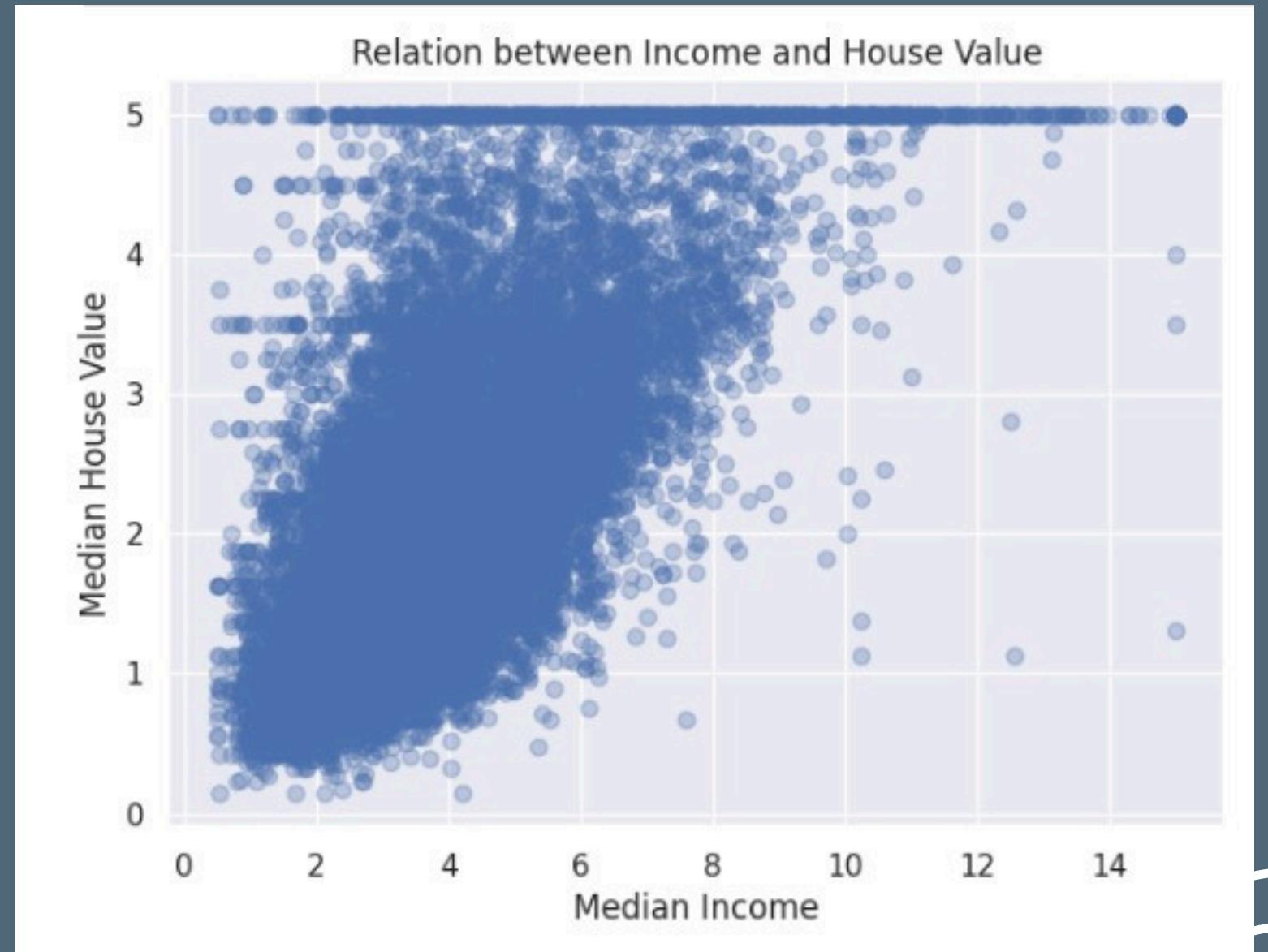
RandomForest feature importance (model insight, includes interactions & non-linear).

RELATION VISUALIZATION

The first

The scatter plot illustrates a clear positive relationship between median income and house values. Higher income levels are generally associated with higher housing prices.

The bar chart highlights that median income has the strongest positive correlation with house values. Other features like average rooms and house age also impact prices but less significantly.





THANK YOU