House Price Prediction Model

This project was about predicting house prices based on several property features. I worked with a public dataset from Kaggle containing around 12,000 records of housing data. The main goal was to understand which factors have the biggest impact on prices and to build a model that can predict them accurately. I used Python for data preparation and machine learning, and Power BI to turn the results into a clean and interactive dashboard.

Tools and Technologies

Python

Core programming language for data processing and model development

Pandas & NumPy

Data manipulation and numerical computing libraries

Matplotlib & Seaborn

Data visualization and exploratory analysis tools

Scikit-Learn

Machine learning framework for model building and evaluation

Data Description

The dataset included features like total square feet, number of bedrooms, bathrooms, balconies, location, and the actual price. Some columns had missing values, and a few contained inconsistent data like text mixed with numbers. I handled these issues during the cleaning stage — replaced missing values where possible, removed duplicates, and fixed wrong entries. After that, the data looked structured and ready for analysis.

Dataset Size

~12,000 records of housing data from Kaggle

Key Features

- Total square feet
- Number of bedrooms
- Number of bathrooms
- Balconies
- Location
- Actual price

Exploratory Data Analysis (EDA)

EDA was the first real step in understanding how the data behaves. I explored how each variable affects the price and checked for outliers or patterns. Larger houses and better locations clearly pushed prices higher. I plotted correlations, boxplots, and scatter plots to confirm this. One key observation was that **price per square foot varied a lot between locations**, which later helped during feature engineering.

To make insights more visual, I built quick visuals in Power BI — showing average price by location, area vs. price comparisons, and how bathrooms or bedrooms affected cost.

Feature Engineering

I created new features and cleaned the existing ones for better model performance.

01

Categorical Encoding

Converted categorical variables like location into numerical form using encoding.

02

Location Impact

Used average price per location to capture location impact.

03

Log Transformation

Applied a log transformation to the target variable (price) to reduce skewness and stabilize variance.

04

Outlier Handling

Handled outliers using the IQR method and capped extreme values.

After all transformations, the dataset was much more stable for training machine learning models.

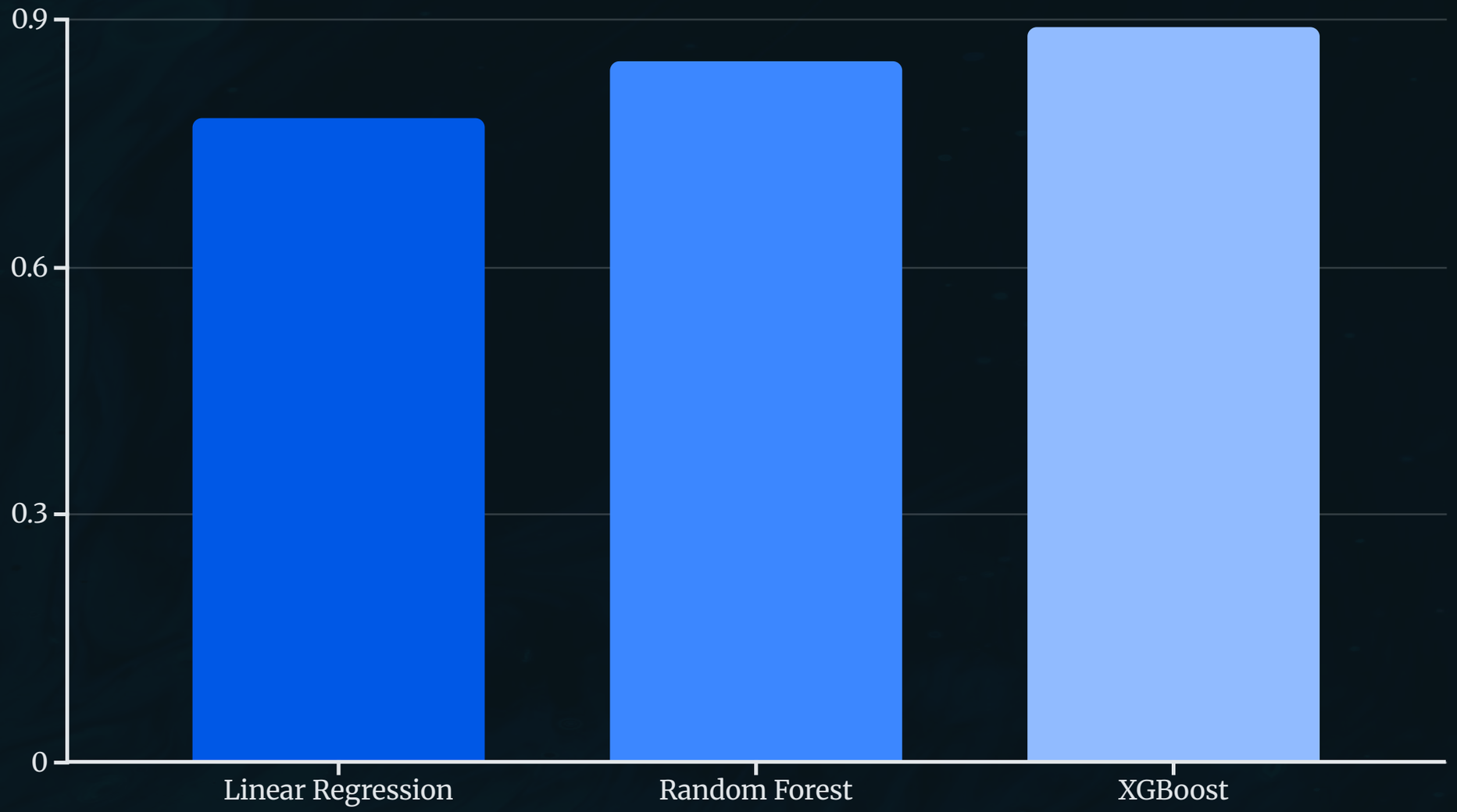
Model Building

I trained three main algorithms: **Linear Regression, Random Forest, and XGBoost**. The data was split in an 80:20 ratio for training and testing. Performance was evaluated using R^2 score and RMSE.

Model	R^2 Score	RMSE	Remarks
Linear Regression	0.78	0.42	Baseline model
Random Forest	0.85	0.35	Better generalization
XGBoost	**0.89**	**0.28**	Best performing model

XGBoost turned out to be the most consistent and accurate model, so I selected it as the final one for predictions.

Model Performance Comparison



The chart clearly shows the progression in model accuracy. **XGBoost achieved an R^2 score of 0.89**, significantly outperforming both Linear Regression and Random Forest. This superior performance, combined with the lowest RMSE of 0.28, made XGBoost the clear choice for final deployment.

Key Insights

Location is King

Location plays the most critical role in pricing, followed by square footage and number of bathrooms.

Premium Properties

Houses with 3+ bedrooms in premium locations had a significantly higher price-per-sqft range.

Regional Anomalies

A few regions showed unusual outliers, possibly due to inconsistent local pricing or unrecorded amenities.

Practical Application

Price prediction accuracy was solid enough to use this model for property valuation support or price negotiation insights.

Impact of Key Features on Price

The analysis revealed that while all features contribute to house pricing, their impact varies significantly. **Location emerged as the dominant factor**, often accounting for price variations even when other features remained constant. The interplay between these factors creates the complex pricing landscape that our XGBoost model successfully learned to predict.

Conclusion

This project covered the full cycle — from raw data to a working, interpretable machine learning model and visual dashboard. **XGBoost delivered the best accuracy.**

Project Achievements

- Successfully processed and cleaned 12,000+ housing records
- Engineered meaningful features that improved model performance
- Achieved 89% R^2 score with XGBoost model
- Created interactive Power BI dashboard for insights visualization
- Delivered actionable insights for property valuation

0.89

R^2 Score

Final model accuracy

0.28

RMSE

Prediction error

12K

Records

Dataset size