HOUSE PRICE
PREDICTION

Marcela Caetano

GOAL OF THE PROJECT

- The goal of this project is to predict housing prices by analyzing historical data and property features.

- Deliver insights that support data-driven decision-making using machine learning regression algorithms

PROJECT IMPLEMENTATION
OVERVIEW

Business Understading

- Predict housing price trends over time by comparing fluctuations between actual and predicted prices.

- Identify the key features most strongly correlated with the target variable (price) to better understand the factors influencing housing price trend

CHALLENGES & SOLUTIONS

- Trend-Focused Approach: Emphasized price trend prediction and analyzed feature correlations.
- Model Optimization: Tested various models to find the most accurate and precise one for forecasting price trends.

- Complex Feature Selection: Numerous relevant features created challenges in determining which ones were most influential for accurate predictions.

DATASET
OVERVIEW

- **Shape**: The dataset contains 21,613 rows and 21 variables(columns). The majority of the variables are numeric.
- **File Format:** The dataset is provided in CSV format.
- **Dataset Source:** https://www.kaggle.com/datasets/minasameh55/king-country-houses-aa

PROJECT METHODOLOGY

ETL, Data Cleaning & Data Understading

- Initial examination of the dataset for missing values and duplicate rows.
- Descriptive statistics and visual analysis to understand the data ditributions.
- Removal of irrelevant features (ID variable).
- Outlier detection using interquartile ranges (IQR).

PROJECT METHODOLOGY

Data Preprocessing & Feature Engineering

- Putting the target (price) on the end for clear understading of correlations of the features with this target
- Dealing with Multicolinearity by examining the correlation matrix (using person method)
- Before going deep on the data Modelling step, aplied feature engineering for the features with lower correlaction with the target price.

PROJECT METHODOLOGY

Data Modeling

- Data Preparation: Split dataset into features (X) and target (price variable), followed by train-test split (70/30).
- Model Selection: Tested models (Linear Regression, Ridge, Lasso, Decision Tree, XGBoost). XGBoost performed best based on R^2, RMSE, MSE, and MAE.
- **How we will use XGBoost?** Defined 100 decision trees, the model will improve predictions by minimizing errors, making small adjustments after each tree.

BEST MODEL
XGBOOST

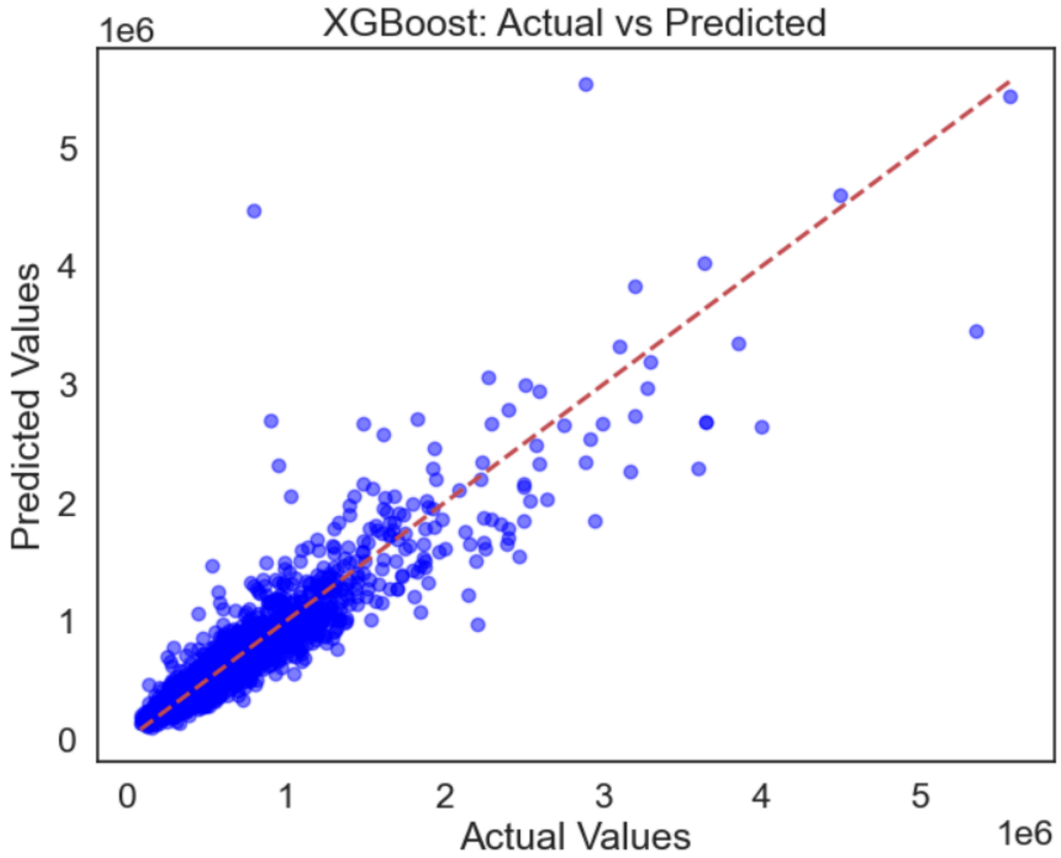
Model 1

- **R^2 Score: 0.8788**
- **MAE (Mean Absolute Error): 69,736.92**
- **RMSE (Root Mean Squared Error): 135,367.13**
- **MSE (Mean Squared Error): 18324260489.88**

- The model's R^2 score indicates it captures most of the underlying data patterns, but the MAE and RMSE show there is room for improvement, especially in reducing larger errors.

Model 2

- Performed Standard Scaller for a better scaler
- The results of the model were the same as the previous one.

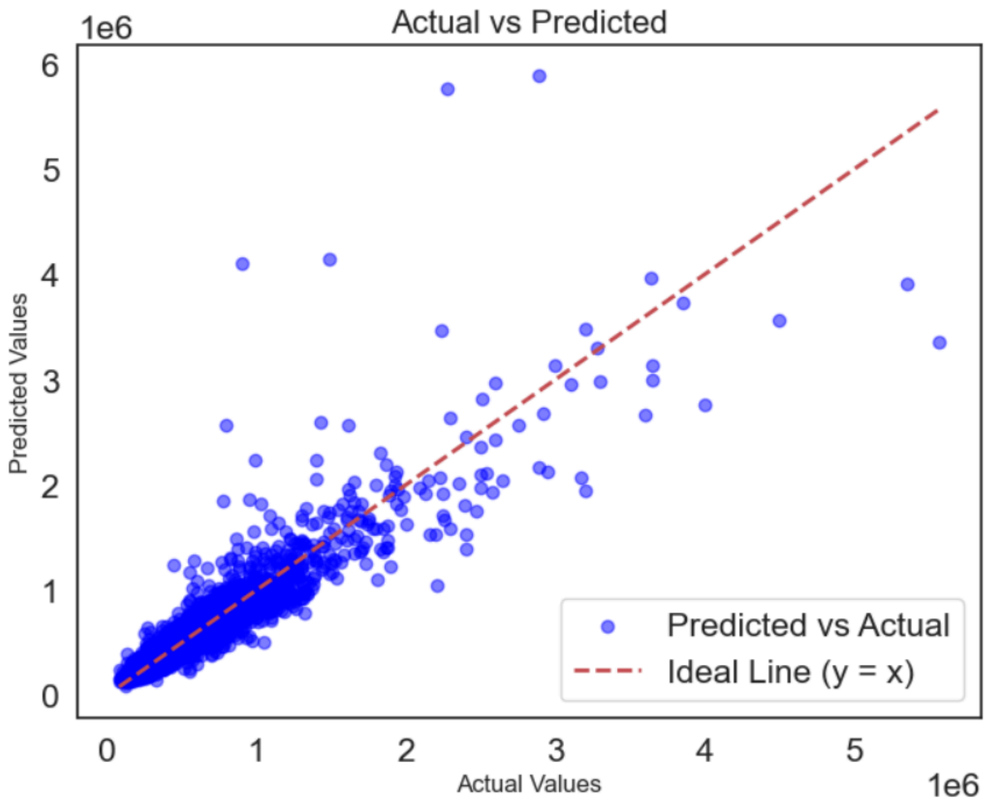


BEST MODEL
XGBOOST

Model 3 - Log Tranformation

- **R^2(log): 0.8211055248811276**
- **RMSE(log): 164452.6636737367**
- **MSE(log): 27044678589.387157**
- **MAE (log): 74518.51744123005**

- The model's R^2 score decreased after aplying log transformation.
- High RMSE and MSE indicates space for model improvement as the outliers or multicolinearity can be seems to be impacting the precision and the margin of error on the model



BEST MODEL
XGBOOST

Model 4 - Outliers Removal

- R²: 0.8736337280079476
- MAE: 40503.78607704978
- MSE: 3297593376.6072335
- RMSE: 57424.67567698779

Model 5 - Feature Engineering

- R²: 0.8749847011192959
- MAE: 70000.8703732264
- MSE: 18048025095.852814
- RMSE: 134342.93839220883

BEST MODEL SCEARIO

Model	R^2	MSE	RMSE	MAE
Model 1	0.88	18324260489.89	135367.13	69736.92
Model 2	0.88	18324260489.89	135367.13	69736.92
Model 3	0.82	27044678589.39	164452.66	74518.52
Model 4	0.87	3297593376.61	57424.68	40503.79
Model 5	0.87	18048025095.85	134342.94	70000.87

Model 4

- The model 4 (with all features and no outliers) showed better results than the previous scenarios, with lower margin of error and balanced variability.

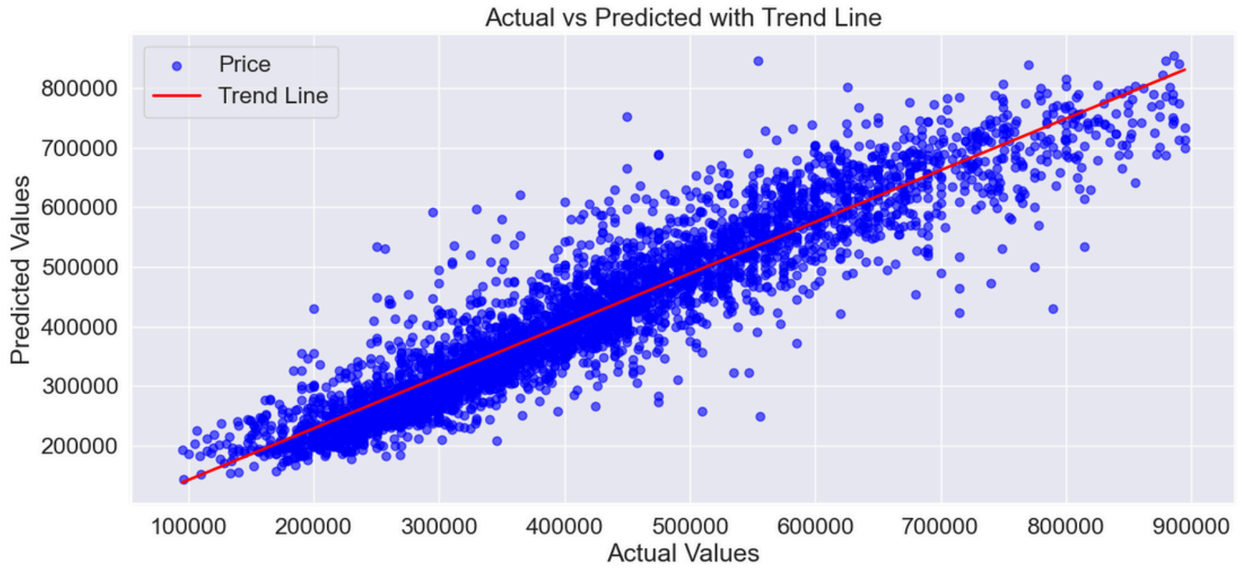
Next steps

- The next steps to improve the model would be fine tuning and checking the hyperparameters, PCA and perform cross validation.

DATA VISUALIZATION & KEY
INSIGHTS

Machine Learning

This is a machine learning model used to predict house prices



Built with Streamlit

Fullscreen


THANK
YOU