ML-Powered House Price Prediction

A comprehensive machine learning solution for accurate real estate valuation, leveraging advanced regression techniques and feature engineering to transform property assessment.

OVERVIEW

Critical Challenges in Property Valuation

Revenue Leakage

Inaccurate property valuations lead to significant lost revenue or overpayment, impacting both buyers and sellers in the market.

Investment Risk

Real estate investors struggle to identify high-potential properties in competitive markets due to unreliable valuation methods.

Operational Inefficiency

Governments and tax authorities face costly, slow manual processes when valuing millions of properties for taxation purposes.

Lending Uncertainty

Banks and mortgage lenders face significant risks without accurate property valuations to inform lending decisions and risk assessment.

Machine Learning Objective & Success Metrics

Task Definition

Regression-based prediction of house prices using property amenities and characteristics as input features.

Primary Metric

R^2 score measuring model performance (target: closest to 1.0 indicates superior predictive accuracy).

Secondary Metric

RMSE (Root Mean Squared Error) for evaluating prediction precision and error magnitude.

Key Beneficiaries

- Banks and financial institutions making lending decisions
- Government agencies conducting property assessments
- Real estate companies optimising pricing strategies
- Individual investors evaluating market opportunities

Business Impact: Reduced estimation risk and processing time by up to 80%.

Project Workflow Architecture

Our systematic approach transforms raw property data into actionable predictions through a rigorous machine learning pipeline, encompassing data understanding, preprocessing, model selection, and deployment preparation.

Define Problem

Data Collection

Exploratory Data

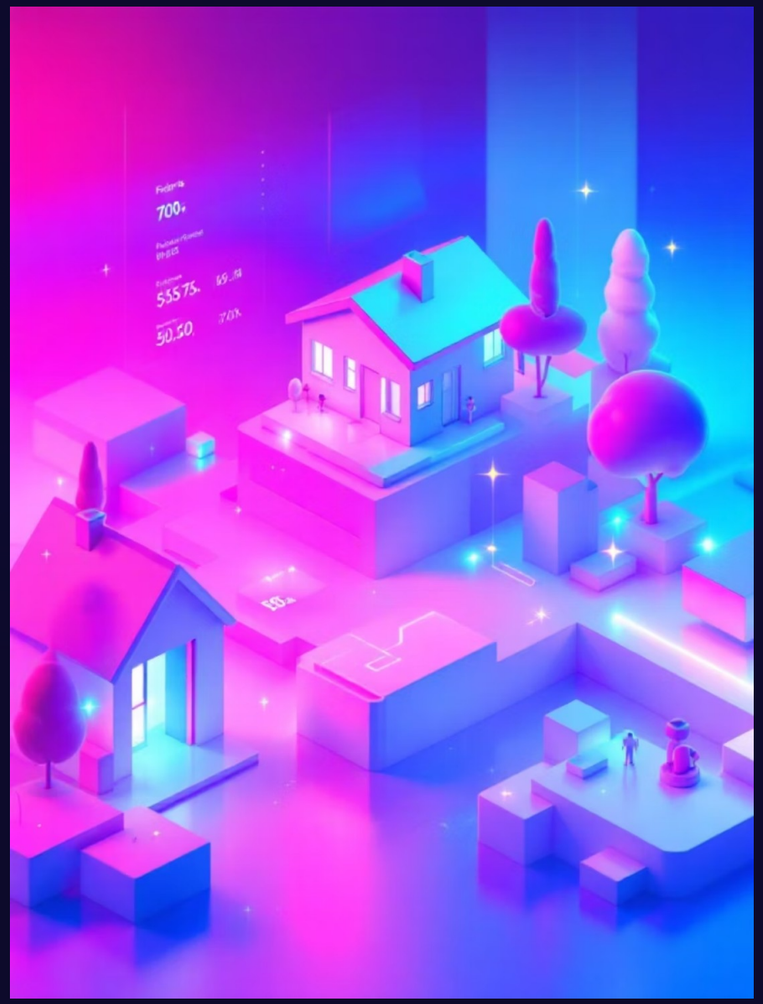
Preprocess & Features

Train & Select

This streamlined workflow ensures every stage—from initial problem framing to final model deployment—is optimised for accuracy and business value.

DATA FOUNDATION

Dataset Overview & Key Features



Dataset Composition

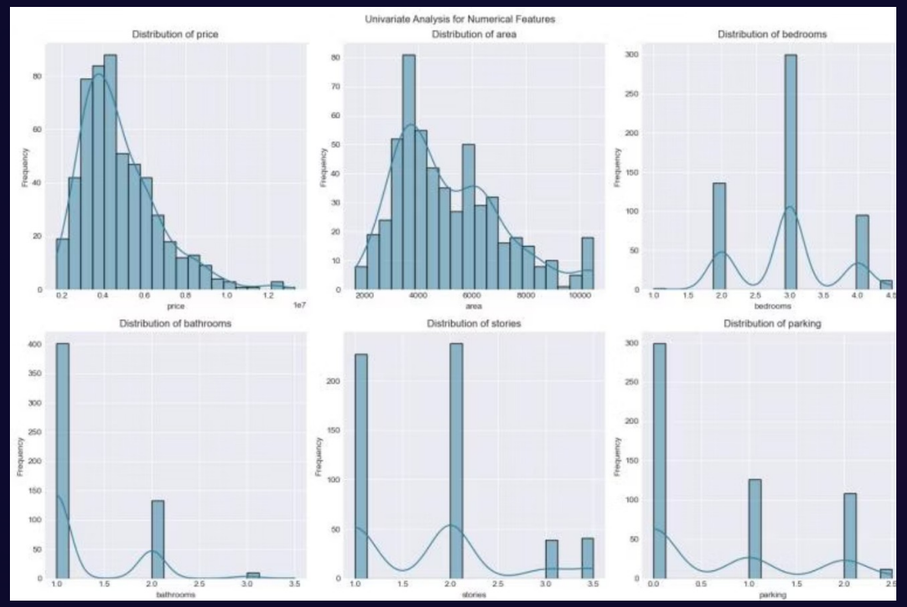
Our analysis leverages a comprehensive dataset comprising **545 property records** with **13 distinct features** capturing essential property characteristics and amenities.

Core Features Include:

- Physical attributes: area, bedrooms, bathrooms, stories
- Location factors: main road access, preferred area designation
- Amenities: guest room, basement, hot water heating, air conditioning
- Additional facilities: parking availability, furnishing status

Data Quality: Zero missing values, no duplicates—ensuring a pristine foundation for model training.

Exploratory Data Analysis: Distributions & Patterns



Key Distribution Insights

- **Price & Area Skew**

  Both exhibit positive skew, indicating concentration at lower values with high-value outliers—typical of real estate markets.

- **Bedroom Preference**

  Three-bedroom properties dominate the dataset (frequency ~300), reflecting market demand patterns.

- **Bathroom Distribution**

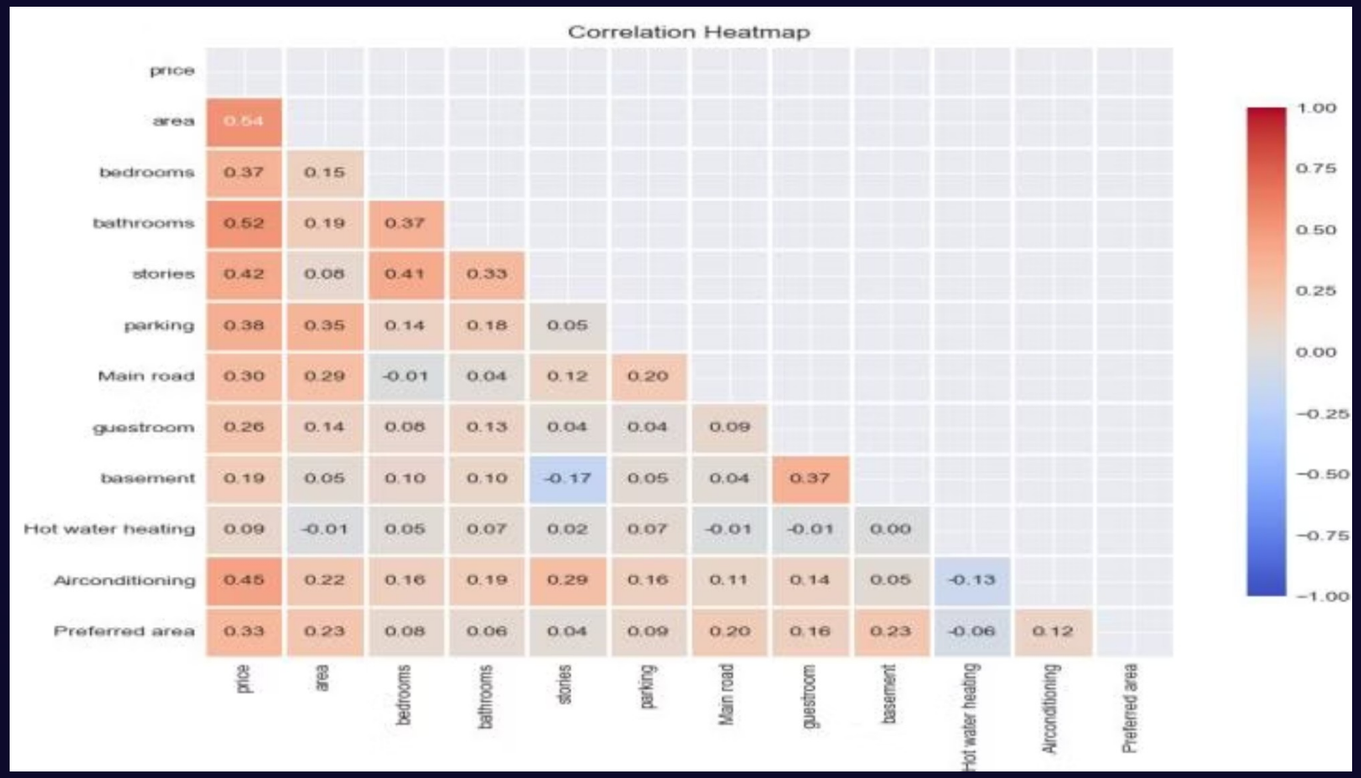
  Single-bathroom homes show highest frequency (~400), suggesting prevalence of compact properties.

- **Parking Availability**

  Properties without parking facilities are most common, potentially indicating urban density factors.

CORRELATION ANALYSIS

Feature Relationships & Price Drivers



Correlation Insights

Correlation analysis reveals critical relationships between property features and pricing, guiding feature engineering and model development strategies.

1 Primary Price Drivers

Area and **Price** show strongest positive correlation with price—key valuation factors.

2 Secondary Factors

Bathrooms and stories contribute moderately to price determination.

3 Feature Multicollinearity

Bathrooms and stories show high correlation with bedrooms, indicating potential redundancy requiring careful feature selection.

Data Preprocessing & Feature Engineering Excellence

1 Outlier Detection

Identified outliers across all features using IQR method: stories (41), area (12), bedrooms (12), bathrooms (9).

2 Outlier Treatment

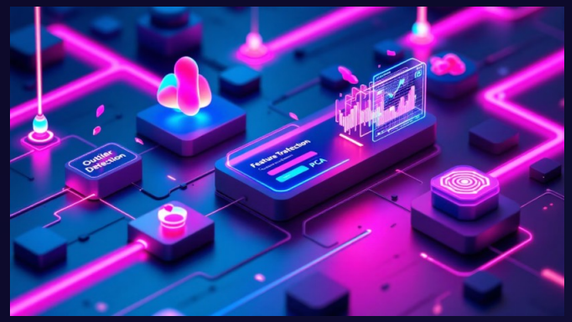
Applied quantile-based capping to preserve data integrity whilst reducing extreme value impact on model performance.

3 Feature Engineering

Created 23 engineered features including area ratios, amenity scores, luxury indices, and logarithmic transformations.

4 Scaling Strategy

Deployed RobustScaler to handle residual outliers using median and IQR, ensuring algorithm stability.



Notable Engineered Features: price_log, area_per_room, amenity_score, luxury_score, bed_bath_ratio, area_bin—each designed to capture complex property relationships.

MODEL PERFORMANCE

Comprehensive Model Evaluation Results

Seven regression algorithms were rigorously evaluated both with and without PCA dimensionality reduction, revealing optimal configuration for production deployment.

LightGBM (Winner)

R^2 Score: 0.862 | MAE: 485,159 | RMSE: 834,057

Superior performance without PCA, capturing complex non-linear relationships effectively.

Random Forest

R^2 Score: 0.848 | MAE: 482,886 | RMSE: 873,841

Strong ensemble performance, second-best overall accuracy without dimensionality reduction.

XGBoost

R^2 Score: 0.843 | MAE: 472,301 | RMSE: 888,730

Competitive gradient boosting performance with lowest MAE amongst top performers.

Linear Regression

R^2 Score: 0.822 | MAE: 556,605 | RMSE: 948,141

Solid baseline performance with interpretable coefficients for feature importance analysis.

Key Finding: Models without PCA consistently outperformed PCA-transformed variants, indicating that full feature set provides superior predictive power for this dataset.

Business Impact & Strategic Recommendations

Deployment Success:

Link to access the deployed work: https://house-price-prediction3.streamlit.app/

86.2%

Model Accuracy

LightGBM R^2 score demonstrates exceptional predictive performance for production deployment.

80%

Time Reduction

Automated valuation reduces manual assessment time, accelerating decision-making processes.

100%

Data Quality

Zero missing values and comprehensive feature engineering ensure robust predictions.

Strategic Next Steps

01

Enhanced Feature Collection

Incorporate additional amenities (smart home features, energy ratings, proximity metrics) to improve model granularity.

02

Cloud Deployment

Migrate from Streamlit Cloud to enterprise infrastructure (AWS/Azure/GCP) for production-scale operations and reliability.

03

Continuous Learning

Implement model retraining pipelines with fresh market data to maintain accuracy amidst evolving real estate trends.

Current Deployment: Fully functional Streamlit application enables real-time predictions through intuitive web interface, ready for stakeholder testing and feedback collection.