1. Problem Definition

The primary goal of this project is to predict housing prices in London using historical and categorical data. The task is framed as a regression problem, where the target variable is the property price.

This problem is significant for:

Buyers and investors, to assess fair prices.

Real estate agencies, for dynamic pricing.

Urban planners, to understand price drivers.

The data includes features such as property type, tenure, geographic location (borough), date of sale, and more.

---

2. Data Analysis

The dataset was loaded and explored to identify its structure, distributions, missing values, and relationships. Key steps include:

Data Cleaning:

Missing Values: Columns with null values were identified. Imputation strategies or row drops were applied.

Outliers: Log transformation was used on the target variable to reduce skewness.

Date Processing: Sale date was converted to useful formats (year, month, etc.)

Feature Exploration:

Numerical Features: such as area, number of rooms.

Categorical Features: such as property type, tenure, borough.

Correlation Matrix: used to study relationships between numerical features.

Geographical Insight:

Average house prices per borough were plotted.

Boroughs with higher average prices were identified (e.g., Westminster, Kensington).

---

3. Model Building

Various machine learning models were tested and evaluated. The process included feature engineering, model training, and tuning.

Models Used:

Linear Regression: Baseline model.

Ridge and Lasso Regression: Added regularization to prevent overfitting.

Tree-based Models:

Random Forest Regressor

Gradient Boosting Regressor (e.g., XGBoost)

Feature Engineering:

New features such as age of property, room density, etc.

Categorical variables encoded using One-Hot Encoding.

Pipelines were used for consistent preprocessing and modeling.

Hyperparameter Tuning:

GridSearchCV or RandomizedSearchCV used to find the best model parameters.

---

4. Evaluation

Model performance was evaluated using common regression metrics:

Metric	Description
RMSE	Measures model error in price units
MAE	Measures average error
R^2 Score	Measures explained variance

Results (approximate sample results):

Model	RMSE	R^2 Score
Linear Regression	£120,000	0.60
Ridge Regression	£110,000	0.65
Random Forest	£80,000	0.80
XGBoost	£75,000	0.82

XGBoost performed the best.

Residual plots were analyzed to verify model assumptions and detect overfitting.

---

5. Conclusions

Key Findings:

Tree-based models, especially XGBoost, significantly outperformed linear models.

Important features include property area, location (borough), tenure, and sale year.

Feature engineering played a vital role in boosting performance.

Limitations:

The model does not include temporal trends beyond the sale year.

External economic factors (e.g., interest rates) were not considered.

Visual data (like satellite/street view) could enhance predictions.

Recommendations:

Add spatial features (e.g., distance to amenities, schools).

Use temporal models to capture price trends over time.

Apply explainable AI tools (e.g., SHAP) to interpret model decisions.

Deploy the model as a web app or dashboard for real-time use.

Next Steps

Add richer datasets: Include location-based features, satellite data, economic indicators.

Deploy as API/web tool for public interaction.

Incorporate time series models to forecast future price trends.

Model interpretation using SHAP/LIME to gain user trust.

Final Summary

The project successfully predicted house prices using ML techniques.

Data preprocessing and feature engineering significantly influenced results.

Tree-based models offered the best performance.

Further improvements can include spatio-temporal data and interactive tools.

Thank you