# Real Estate Price Prediction Model

## Data-Driven Insights for Property Valuation

By Ehab Henein

# Introduction

This project uses machine learning techniques to develop a predictive model for real estate prices. The model, which leverages property features and historical sales data, will help buyers, sellers, and investors make informed decisions. This study involves data preprocessing, feature engineering, model selection, and evaluation to ensure accurate predictions.

# Data Collection and Preprocessing

The dataset is sourced from multiple real estate listings and structured into:

- **Train.csv**: Contains property features such as size, location, room count, and historical prices.
- **Data Cleaning**: Missing values were handled via imputation, normalization, and outlier removal.
- **Feature Engineering**: New features such as price per square foot and neighborhood price trends were created.

# Feature Engineering

Created additional features, including:

- **Neighborhood average price**
- **Historical price growth rate**
- **Rolling statistics (mean, standard deviation) to capture market trends**
- **Encoded categorical features using one-hot encoding and transformed numerical variables.**

# Model Development

Various machine learning algorithms were explored to predict property prices:


## Regression Analysis:

Linear, Ridge, and Lasso Regression were tested.


## Advanced Algorithms

Random Forest, Gradient Boosting, and XGBoost were implemented for improved accuracy.

## Feature Selection

Applied Lasso Regression to identify the most influential features.

# Model Prediction Function

Developed a function to predict property prices based on input features.


**Validate property details**

**Uses trained models to estimate price**


**Returns predicted price with confidence intervals**

# Model Evaluation

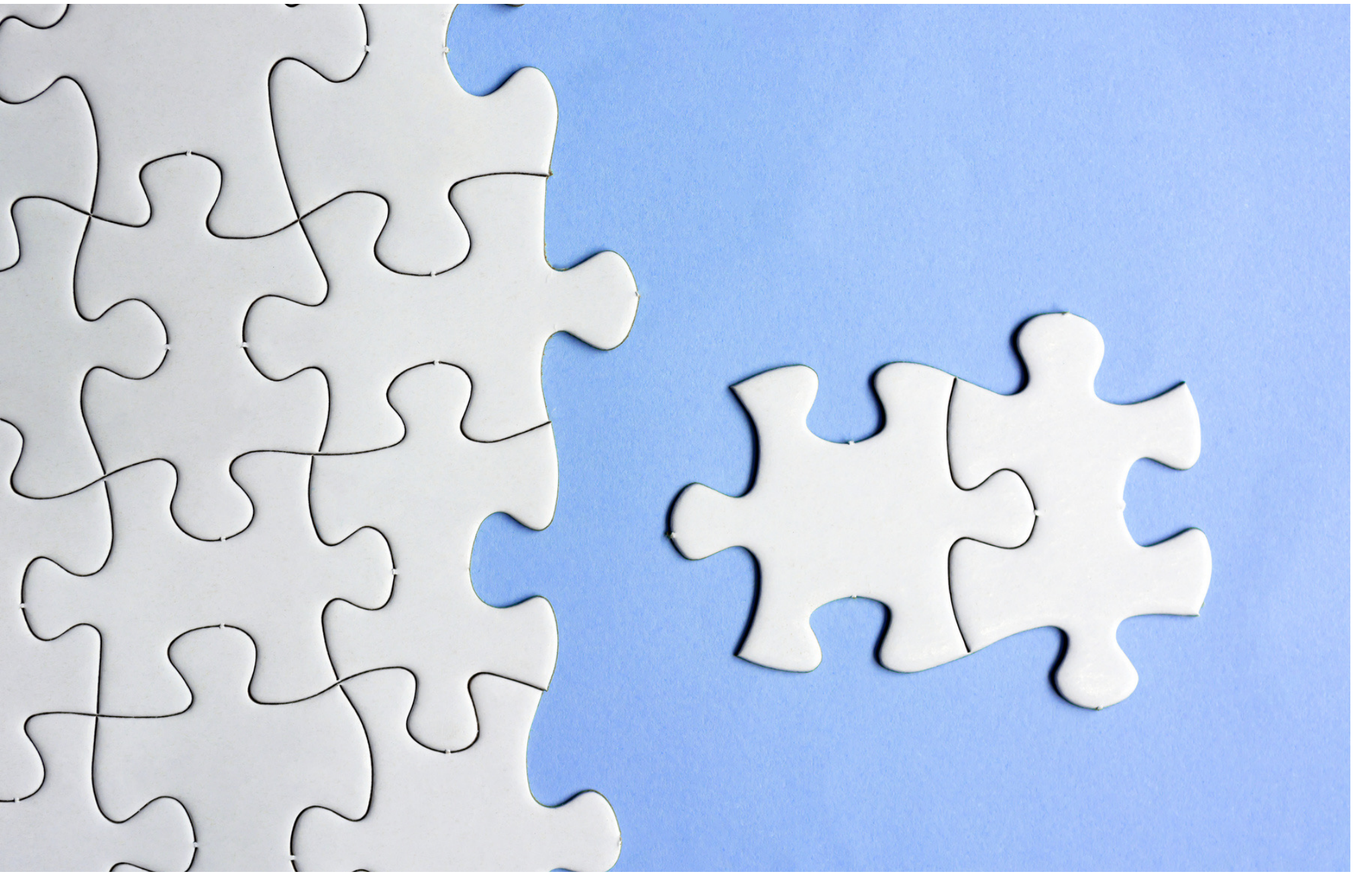
Performance metrics used:

- ***Mean Absolute Error (MAE):*** *Measures average prediction error.*
- ***Root Mean Squared Error (RMSE)****: Assesses prediction accuracy.*
- ***R-Squared (R2)****: Evaluates variance explained by the model.*
- ***Cross-Validation:*** *Ensured robustness and generalizability.*

# Results and Findings


- Achieved an R2 value of 0.742, indicating a strong correlation between predicted and actual prices.
- RMSE of 0.280, suggesting an acceptable error margin.
- PCA was applied to retain 90% variance while reducing dimensionality.
- Linear regression performed well, but advanced models like Gradient Boosting showed superior results.

# CHALLENGES AND LIMITATIONS



- **Data Limitations**: Incomplete records and missing values impact accuracy.
- **Feature Importance**: Some variables contribute minimally to prediction.
- **Market Fluctuations**: The model does not account for external economic factors affecting real estate trends.
- **Overfitting Risk**: Certain models may perform well on training data but not unseen data.

# Recommendations

**To enhance model performance and real-world application:**

- **Incorporate More Data**: Use additional sources such as economic indicators, interest rates, and seasonal trends.
- **Improve Feature Engineering**: Introduce more granular location-based features and economic variables.
- **Test Alternative Models**: Explore deep learning techniques for improved predictions.
- **Deploy the Model**: Build an interactive tool for real estate professionals to use in market analysis.

# Business Impact and Future Opportunities

- **Market Application**: The model can help investors, buyers, and real estate agents make data-driven decisions.
- **Scalability**: The model can be expanded to multiple markets with more data sources.
- **Integration Possibilities**: This approach can be integrated with online real estate platforms for automated valuation services.


# Conclusion

The real estate price prediction model demonstrates the potential of machine learning in estimating property values. While the model performs well, improvements in data quality, feature engineering, and model selection can further enhance accuracy. Future work should focus on integrating external economic data and automating predictions for real-time use.


Open floor for questions and discussions.