Predictive Modeling of Melbourne Housing Prices Using Machine Learning

Term Project – Machine Learning Models for Business Analytics

Presented by: Masfa and Sana

Date: 1st June 2025

Introduction

Objective:

- Predict housing prices in Melbourne using machine learning.

Why this matters:

- Accurate predictions help buyers, sellers, and investors.

Methods used:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor

Dataset Overview

Dataset source:

- Melbourne Housing Dataset (Kaggle)

Size and structure:

- ~1,350 rows, 21 columns
- Target variable: `Price`

Selected features:

- `Rooms`, `Bathroom`, `Landsize`, `BuildingArea`, etc.

Data Preprocessing

- Handled missing values through filtering or imputation.
- Selected relevant numerical features for modeling.
- Performed train-test split for model validation.
- Visualized distributions and correlations of features.

Models Used

- **Linear Regression**:

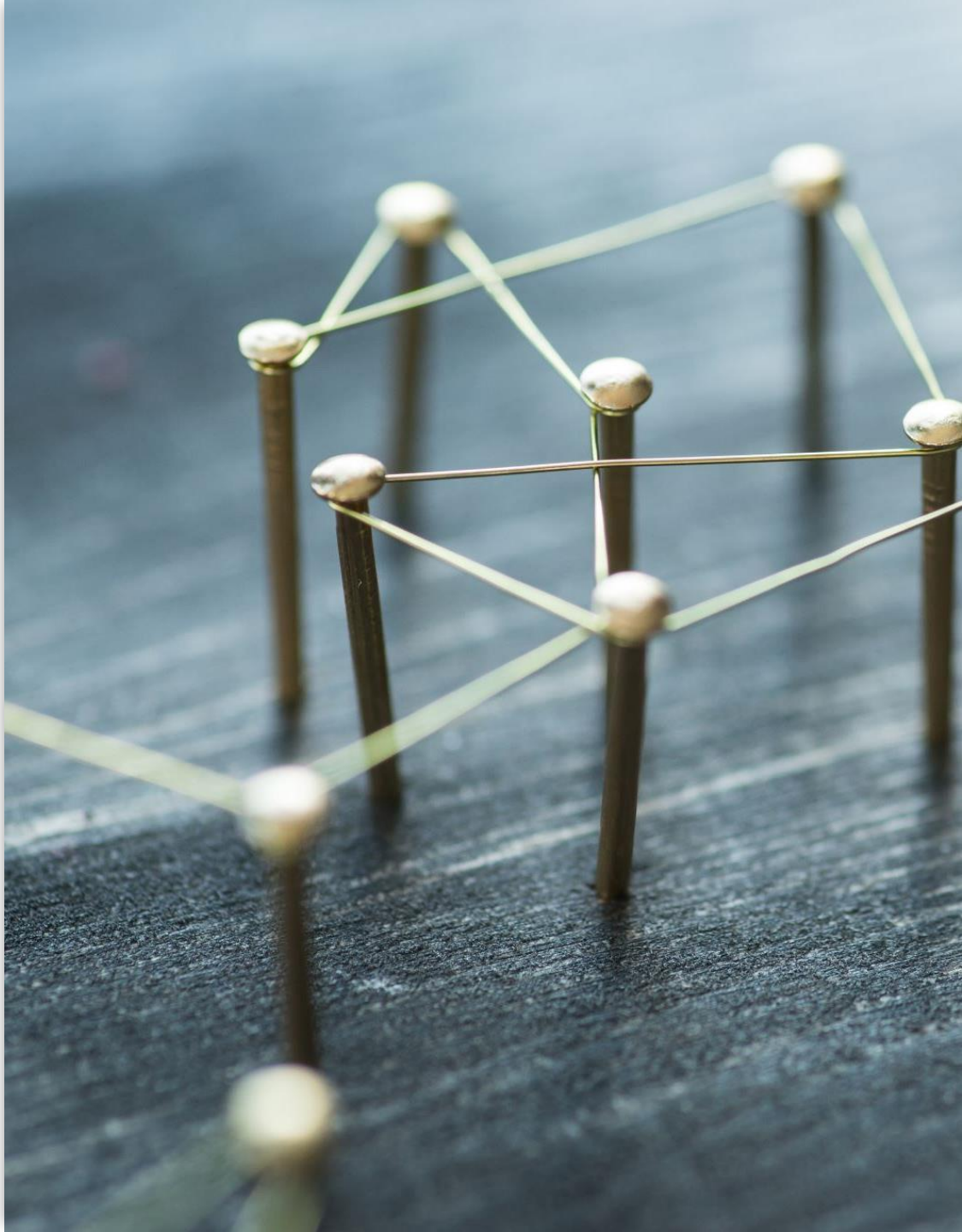
Simple, interpretable baseline model.

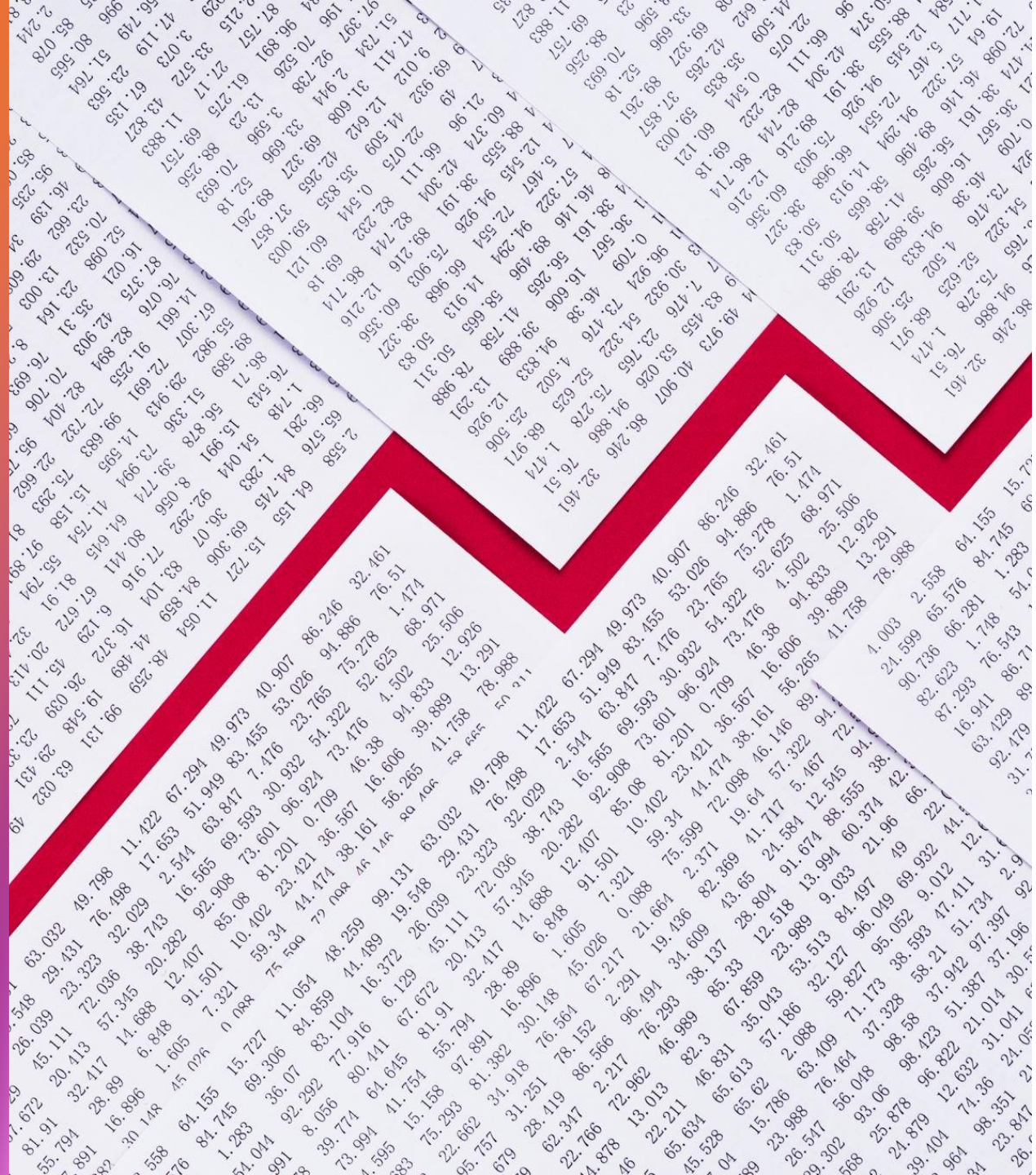
- **Decision Tree Regressor**:

Handles non-linear relationships and splits data into decision nodes.

- **Random Forest Regressor**:

Ensemble of decision trees for improved accuracy and generalization.





Evaluation Metric

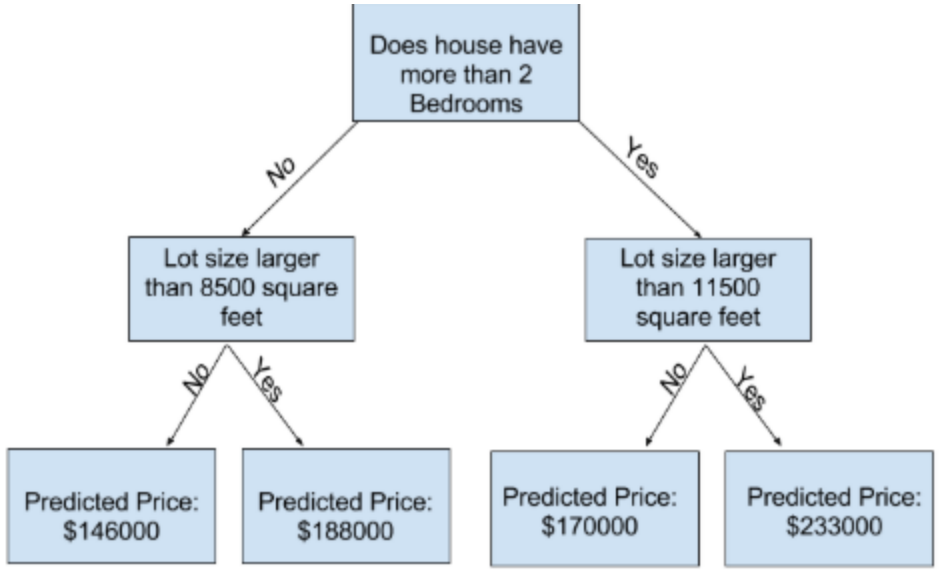
Metric: Mean Absolute Error (MAE)

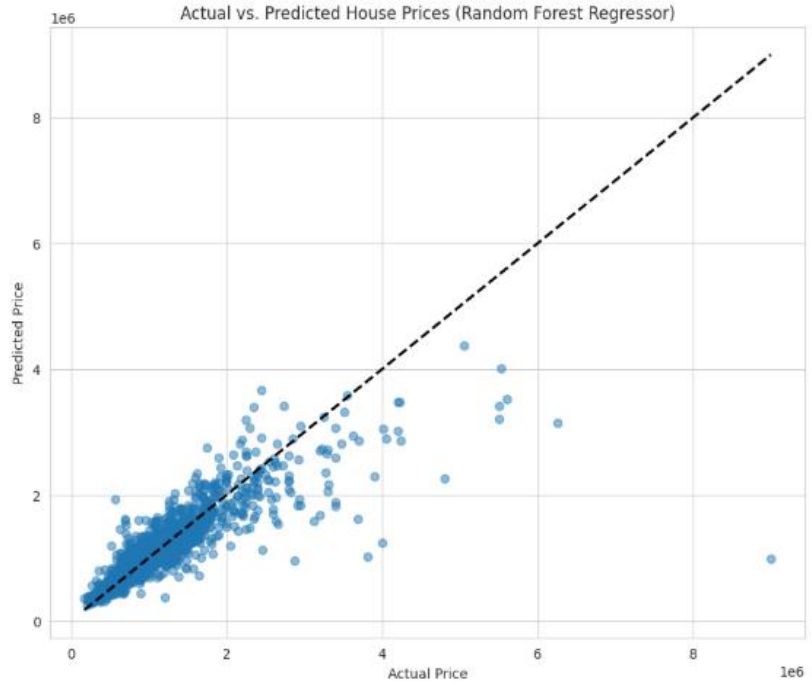
- Measures average magnitude of prediction errors.
- MAE = mean(|actual - predicted|)

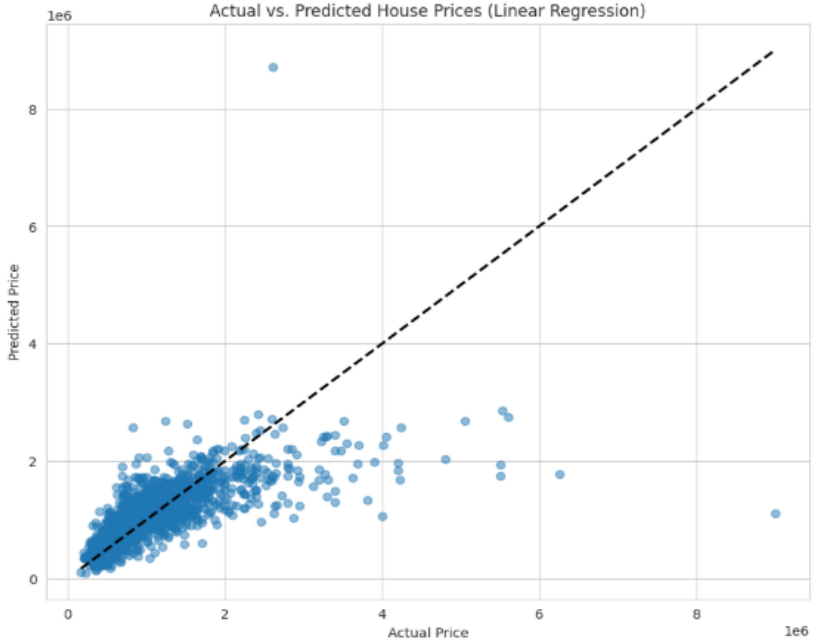
Why MAE?

- Intuitive and scale-sensitive.
- Easy to compare across models.

Results


Does house have more than 2 Bedrooms
No
Yes
Lot size larger than 8500 square feet
Lot size larger than 11500 square feet
No
Yes
No
Yes
Predicted Price: $146000
Predicted Price: $188000
Predicted Price: $170000
Predicted Price: $233000


Actual vs. Predicted House Prices (Random Forest Regressor)
1e6
Predicted Price
Actual Price
1e6


Actual vs. Predicted House Prices (Linear Regression)
1e6
Predicted Price
Actual Price
1e6

Feature Importance

Feature Insights from Random Forest:

- `Landsize`, `Rooms`, and `Bathroom` were most influential.
- Importance visualization highlights which features impact predictions the most.

Why it matters:

- Helps explain the model.
- Guides future data collection priorities.

Conclusion

Random Forest achieved the best predictive performance.

Linear Regression was useful as a baseline model.