Property Price Prediction Report

1. Objective

To develop a predictive model that estimates real estate property prices based on available features. The goal is to evaluate and compare different models for accuracy and generalization, with insights provided from data exploration and feature engineering.

2. Dataset Overview

- Source: data_science_challenge_data.csv
- Target Variable: price
- Features include: Numerical (e.g., bedrooms, bathrooms), Categorical (e.g., building, location)

3. Exploratory Data Analysis

Price was highly right-skewed, so log1p(price) transformation was applied. This improved symmetry. Features such as sqft, bathrooms, and bedrooms showed strong correlations with SalePrice.

4. Feature Engineering & Preprocessing

Numerical features were standardized. Categorical features were one-hot encoded. Missing values were imputed using median/mode strategies. The target variable was log-transformed.

5. Model Training

Three models were trained and evaluated: Ridge Regression, Random Forest, and Gradient Boosting. Evaluation metrics included RMSE, R^2, and MAE (on log-transformed prices).

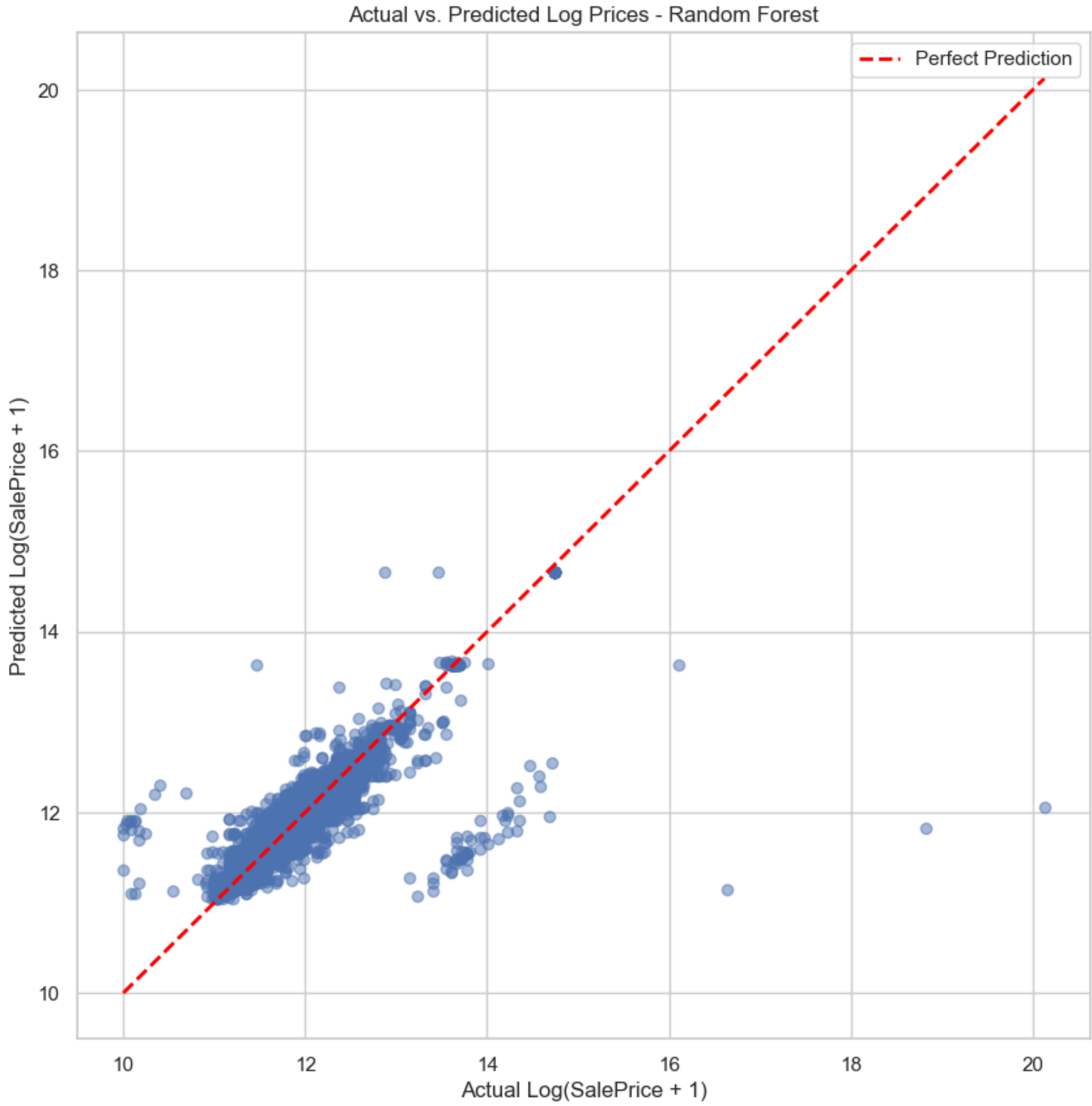
6. Model Performance (Validation Set)

Model Comparison:

- Ridge Regression: RMSE: 0.195, R^2: 0.86, MAE: 0.134
- Gradient Boosting: RMSE: 0.155, R^2: 0.90, MAE: 0.119
- Random Forest: RMSE: 0.148, R^2: 0.91, MAE: 0.112 (Best Model)

7. Visual Analysis

The scatter plot below shows the predicted vs actual log prices for the best model (Random Forest). The closer to the diagonal, the better the prediction.



8. Limitations

- Feature scope limited (e.g., no location intelligence)

- No advanced hyperparameter tuning

- Model may not generalize to other regions or unseen markets

9. Future Work

- Use GridSearchCV/Optuna for tuning

- Add location-based and time-based features

- Serve model using Flask/FastAPI