Real Estate Price Prediction System

Machine Learning Project Summary

1. Problem Statement

This project explores the use of supervised machine learning to estimate residential home sale prices using structured real estate listing data. Traditional pricing approaches often rely on manual comparative analysis, which can be time-consuming and inconsistent across large datasets.

The objective of this project was to design and implement a **reproducible, data-driven pricing pipeline** that can generate defensible sale price estimates and support interactive exploration of pricing scenarios.

The system is intended to **augment human judgment**, not replace it, by providing objective predictions and insight into the factors most strongly associated with home prices.

2. Dataset and Features

The model was trained on a structured residential real estate dataset containing approximately **1,000 property listings from South Florida**. The dataset includes a combination of real and synthetically generated records to ensure sufficient sample size while preserving realistic feature distributions. Key features used in modeling include:

- Listing price
- Sale price
- Square footage
- Number of bedrooms and bathrooms
- Lot size
- Year built
- ZIP code
- Pool availability

Prior to modeling, the dataset underwent cleaning and validation steps to ensure:

- Consistent numeric data types
- Valid value ranges
- Removal of irrelevant or redundant fields

The cleaned dataset served as the basis for exploratory analysis, model training, and evaluation.

3. Modeling Approach

Multiple regression models were evaluated to establish both a baseline and a more expressive predictive approach:

- **Linear Regression** was used as a baseline model
- **Random Forest Regression** was selected as the final model

Model performance was evaluated using standard regression metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R^2 score

The Random Forest model consistently outperformed the baseline Linear Regression model, demonstrating improved predictive accuracy and a stronger ability to capture **non-linear relationships** between property features and sale price.

4. Results and Interpretation

The trained Random Forest model produced stable and accurate price estimates across the test set. Feature importance analysis revealed that:

- Listing price
- Square footage
- Year built

were among the most influential predictors of sale price.

These results align with real-world expectations and provide interpretability into how structural and pricing factors influence predicted outcomes. The model's performance and feature importance outputs support its use as a **decision-support tool** rather than a black-box predictor.

5. Implementation and Tooling

The project was implemented in **Python** using the following tools and libraries:

- pandas and NumPy for data manipulation
- scikit-learn for model training and evaluation
- Jupyter Notebook for analysis and experimentation
- Streamlit for interactive application development

The workflow follows a clear separation of concerns:

- Data preparation and model training are implemented in a notebook

- The trained model is used by a lightweight Streamlit application to generate predictions and visual insights

This structure reflects common patterns used in applied machine learning projects and supports both reproducibility and usability.

6. Summary

This project demonstrates the application of machine learning techniques to a real-world pricing problem, from data preparation through model evaluation and interactive use. It highlights practical considerations such as feature selection, model interpretability, and deployment-oriented design.

The resulting system provides a clear example of how predictive modeling can be integrated into decision-support workflows while maintaining transparency and analytical rigor.