Data Science Midterm Project

JAFARNEJAD, DIBA & WHITE, KORINNE ANGELA

Project/Goals

1. Load, preprocess, and explore housing sales data.
2. Train, evaluate, and optimize supervised learning models.
3. Fine-tune the best model and implement a prediction pipeline.

Process

I. Data Cleaning, Exploration and Visualization

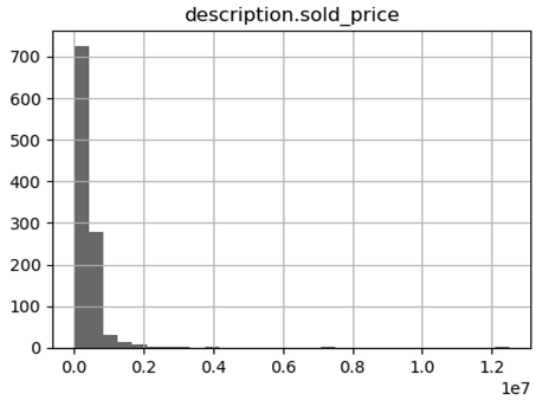
II. Model Selection

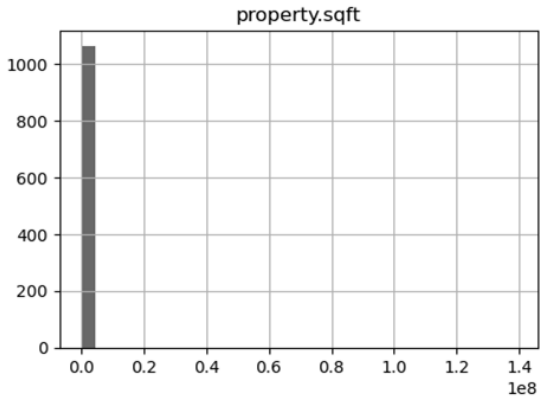
III. Finetuning

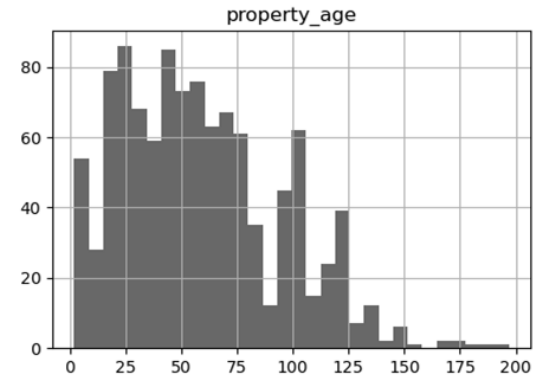
IV. Model Evaluation & Results

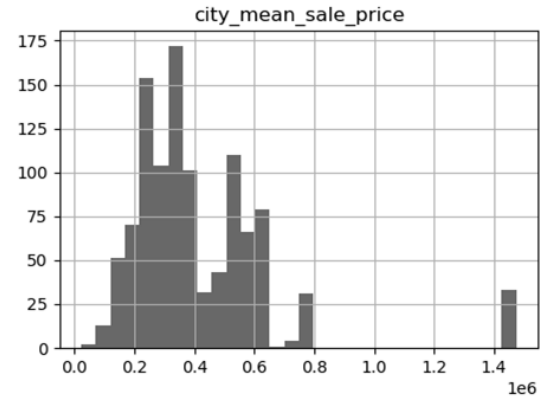
I. Data Cleaning, Exploration and Visualization

Before Log Transformation

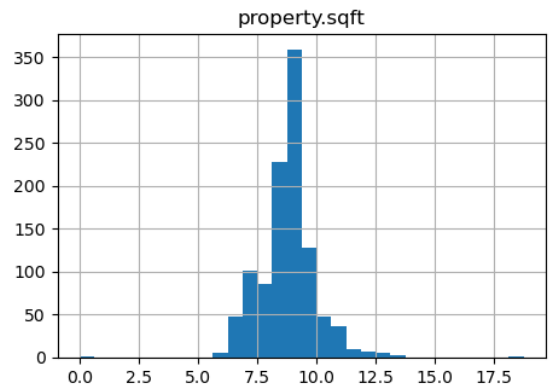


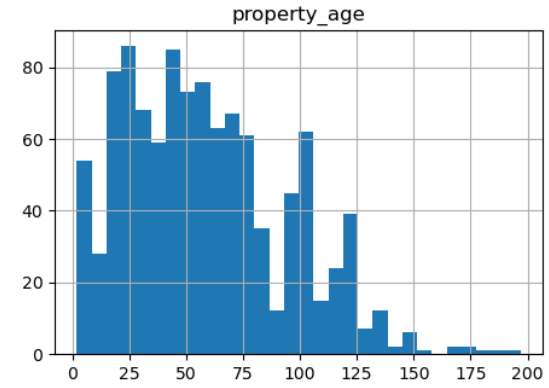


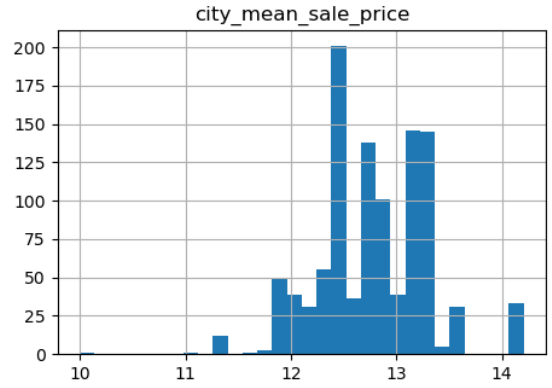




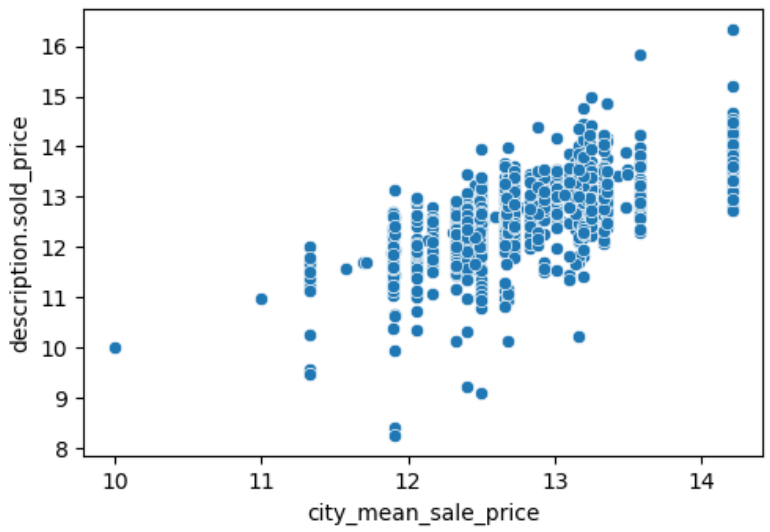
After Log Transformation

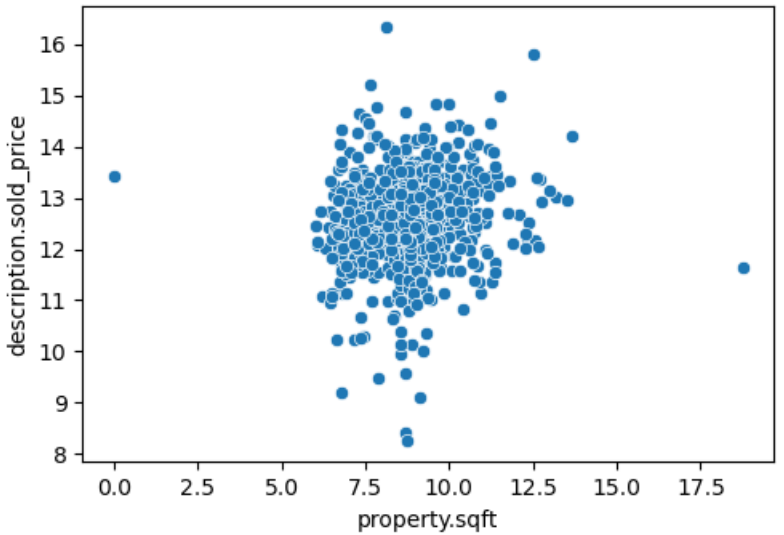


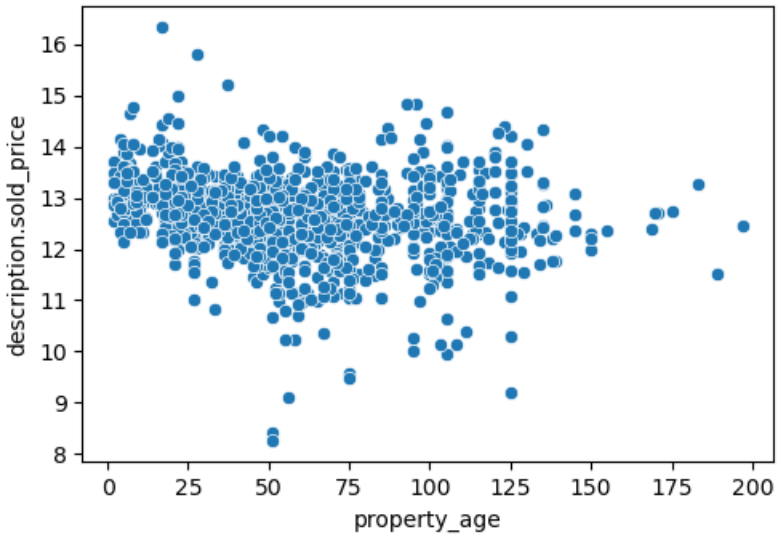




I. Data Cleaning, Exploration and Visualization





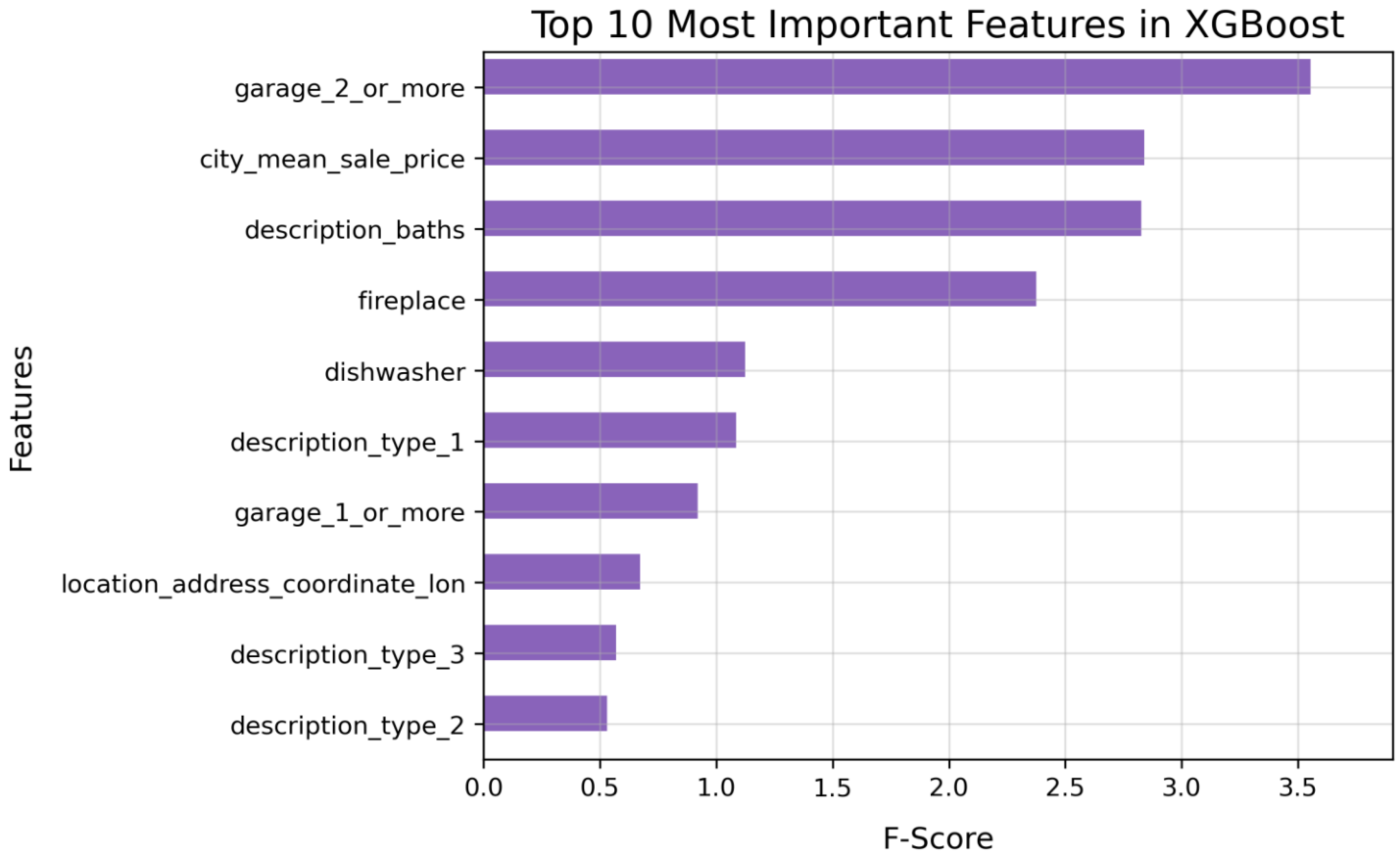


II. Model Selection: XGBoost

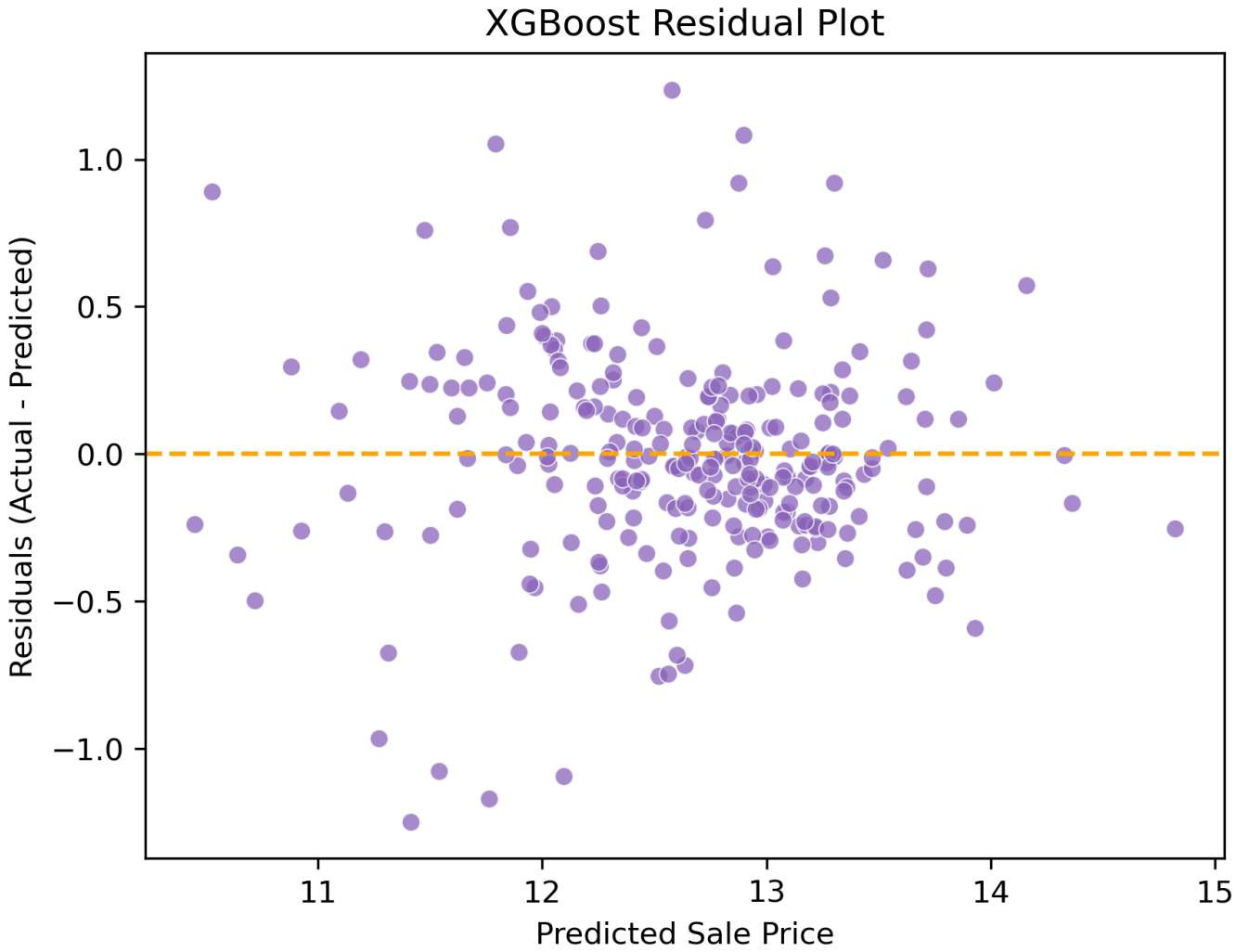
Model	Train MAE	Test MAE	Train RMSE	Test RMSE	Train R^2	Test R^2
Linear Regression	0.30	0.30	0.43	0.39	0.6905	0.7447
SVR	0.15	0.26	0.31	0.37	0.8466	0.7683
Random Forest	0.21	0.28	0.33	0.40	0.8169	0.7293
XGBoost	0.20	0.27	0.30	0.37	0.8494	0.7758

III. Fine-Tuned XGBoost's Performance:

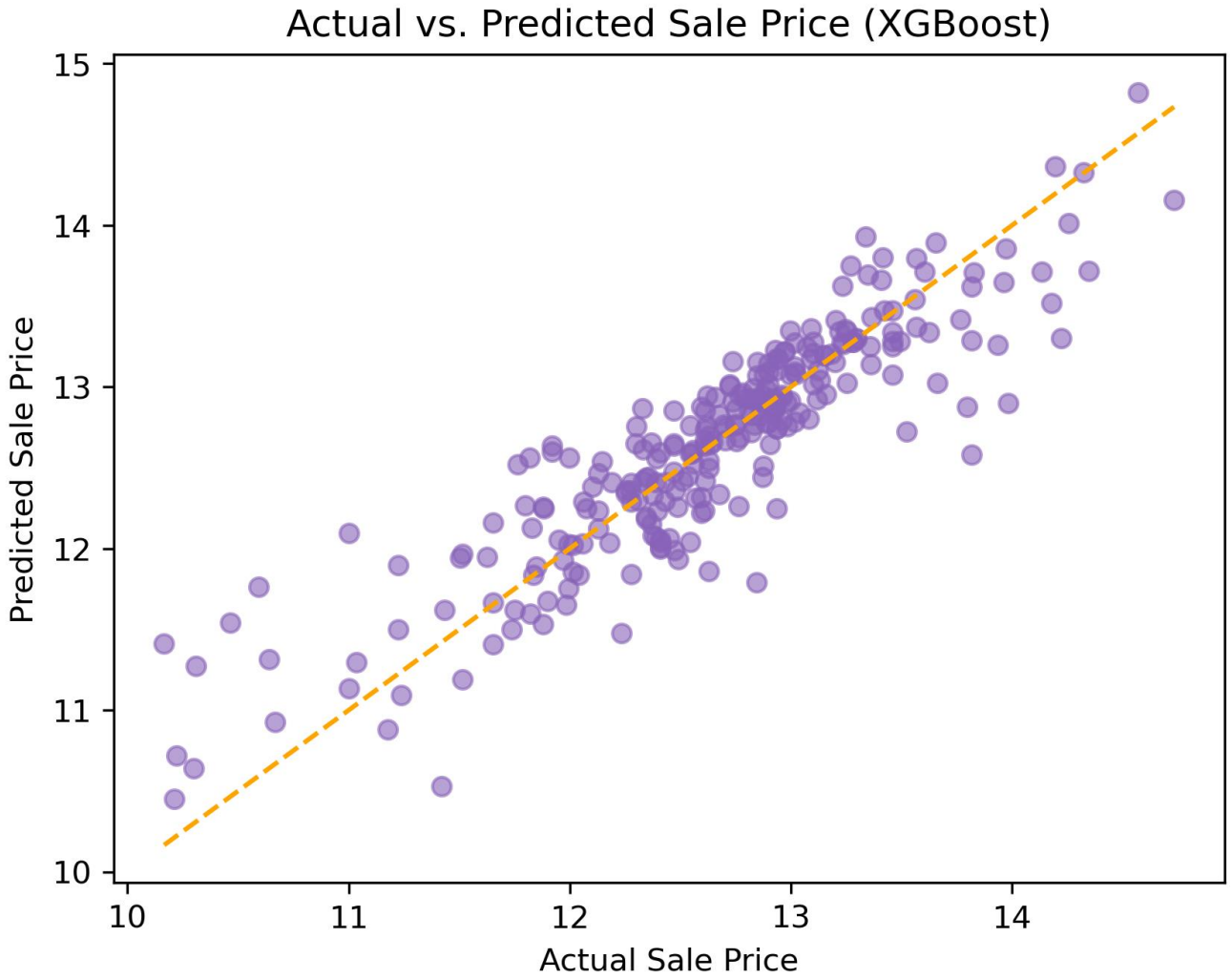
Metric	Before (XGBoost)	After (Fine-Tuning)	Change
Test MAE	0.2700	**0.2536**	**-6.1%**
Test RMSE	0.3700	**0.3525**	**-4.7%**
Test R^2 Score	0.7758	**0.7945**	**+1.9%**



IV. Model Evaluation & Results



IV. Model Evaluation & Results



IV. Model Evaluation & Results

Conclusion

Challenges

1. Iterative Data Cleaning & Wrangling
2. Handling List-Type Tags in String Format

Future Goals

1. More advanced feature selection
2. Train separate models for high vs low priced homes
3. Address potential multicollinearity
4. Incorporate more location-based features