Capstone Two – Final Project Report: Predicting Housing Prices in Washington, D.C.

By Gbatchin Kochoni

April 2025

Table of Contents

1. Context and Background

The Washington, D.C., real estate market is dynamic, complex, competitive, and constantly evolving. With economic growth, demographic shifts, and changing housing demand, accurately forecasting regional housing prices is essential for various stakeholders, including buyers, sellers, investors, and policymakers. Machine learning and data analytics provide the opportunity to develop predictive models capable of understanding the multiple factors that influence real estate prices, thereby enabling better decision-making.

2. Project Goal

The main objective is to train and evaluate several regression models to predict housing prices. We aim to identify the most influential features and optimize model performance for future deployment.

3. Data Sources

The main source of our dataset is DC_Properties.csv, which was sourced from Kaggle. It includes property characteristics such as:

- Number of rooms (ROOMS)
- Bedrooms (BEDRM), Bathrooms (BATHRM)
- Land Area (LANDAREA)
- Sale Price (PRICE)
- Location attributes (WARD, ZIPCODE, HEAT, STRUCT).

4. Data Wrangling

Data cleaning steps included:

- Removing irrelevant or empty columns
- Filtering records with non-positive prices (PRICE ≤ 0)
- removing duplicates
- dropping irrelevant columns.

5. Exploratory Data Analysis (EDA)

We explored the sale price distribution, identified correlations between variables, and visualized average prices by Ward. Key findings:

- Sale prices were right-skewed
- ROOMS, BATHRM, and LANDAREA positively correlate with PRICE
- WARD-based analysis revealed geographic price disparities.

6. Feature Engineering

One-hot encoding was applied to categorical variables (HEAT, STRUCT, WARD). Missing values were handled. Features like ROOMS, BATHRM, and LANDAREA were retained based on correlation. Variables with excessive missing values were excluded.

7. Pre-processing and Training Data Development

We applied one-hot encoding to categorical variables, standardized numerical features using StandardScaler, and split the dataset into 80% training and 20% testing subsets.

8. Modeling (with Hyperparameter Tuning)

Three models were applied:

1. Linear Regression (baseline)
2. Decision Tree Regressor (with GridSearchCV tuning)
3. Random Forest Regressor (with GridSearchCV tuning)

Table 1: Model Evaluation Results

Model	MAE ($)	RMSE ($)	R^2	Selected
Linear Regression	233,660.42	400,869.24	0.4954	No
Decision Tree (Tuned)	276,701.55	468,127.76	0.3119	No
Random Forest (Tuned)	225,166.54	376,057.71	0.5559	Yes

The Random Forest Regressor with tuned hyperparameters was the best-performing model and has been retained for final evaluation and prediction.

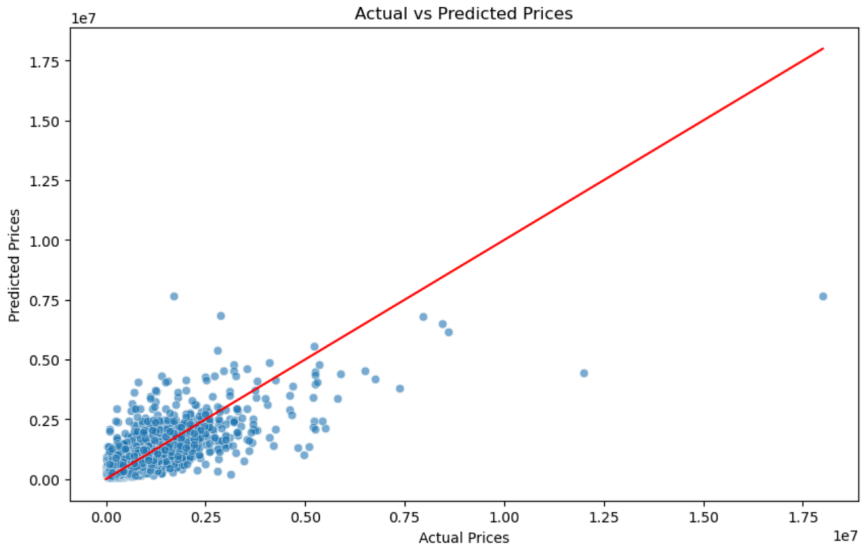

Figure 1: Random Forest Predictions vs. Actuals (placeholder for visual).

9. Evaluation Summary

Random Forest outperformed the other models in MAE, RMSE, and R^2. It was selected as the final model.

10. Conclusion

The project successfully predicted residential property prices with strong accuracy using Random Forest, as it is the outperformed model. Therefore, Random Forest is selected for deployment. In the following steps, we suggest:

- Adding external data (e.g., crime, schools)
- Testing ensemble methods like XGBoost or Gradient Boosting
- Deploying the model as a web app or via Flask API