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

By Gbatchin Kochoni

May 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 offer the opportunity to develop predictive models that can understand the multiple factors influencing real estate prices, thereby enabling more informed 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 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 distribution of sale prices, identified correlations between variables, and visualized the 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

8.1. Models Evaluated

This report compares multiple regression models applied to housing data in Washington, D.C. The models include:

- Linear Regression,
- Ridge,
- Lasso,
- Decision Tree,
- Random Forest, and
- XGBoost.

The goal is to determine the best predictive model for house prices.

8.2. Model Evaluation and Performance Comparison

The models were evaluated using the following metrics:

- MAE (Mean Absolute Error)
- MSE (Mean Squared Error)
- RMSE (Root Mean Squared Error)
- R^2 (Coefficient of Determination)
- Adjusted R^2
- MAPE (Mean Absolute Percentage Error)

The model evaluation results are shown in the table below:

Table 1: Model Evaluation Results

Model	MAE	MSE	RMSE	R2	Adjusted R2	MAPE	Selected
Linear Regression	233660.419	160696147420.317	400869.240	0.495	0.494	4.100	No
Ridge Regression	233681.885	160751190817.963	400937.889	0.495	0.494	4.099	No
Lasso Regression	233664.600	160707221483.153	400883.052	0.495	0.494	4.100	No
Decision Tree	276701.549	219143598020.019	468127.758	0.312	0.310	3.576	No
Random Forest	**225166.535**	**141419401110.085**	**376057.710**	**0.556**	**0.555**	**3.754**	**Yes**
XGBoost	**209933.807**	**139808073385.339**	**373909.178**	**0.561**	**0.560**	**5.681**	**No**
Linear Regression (Not Tuned)	233660.419	160696147420.317	400869.240	0.495	0.494	4.100	No
Ridge Regression (Tuned)	233810.927	161022573043.344	401276.181	0.494	0.493	4.090	No
Lasso Regression (Tuned)	233695.582	160778855542.942	400972.388	0.495	0.494	4.096	No
Decision Tree Regressor (Tuned)	217235.697	145453246116.003	381383.332	0.543	0.542	5.799	No
Random Forest Regressor (Tuned)	**212598.031**	**126857626955.353**	**356170.783**	**0.602**	**0.601**	**5.796**	**Yes**
XGBoost Regressor (Tuned)	**209672.114**	**143213755107.841**	**378435.933**	**0.550**	**0.549**	**5.809**	**No**

Based on the evaluation of multiple regression models, the **Random Forest** stands out as the most accurate and reliable performer after tuning. It achieved the **lowest RMSE (356,171)**, **highest R^2 (0.602)**, and **lowest MSE**, indicating its superior ability to capture data variance and minimize prediction errors.

Although the **XGBoost** initially showed the **lowest Mean Absolute Error (MAE) at 209,933**, it was ultimately outperformed by the tuned Random Forest model. This highlights the importance

of hyperparameter tuning, which notably improved model performance, not only for Random Forest but also for the Decision Tree. Conversely, **linear models** (Linear, Ridge, Lasso) showed **minimal gains** from tuning. They plateaued at an R^2 of around 0.495, revealing their limited capacity to model non-linear and complex patterns typical of real estate data.

This aligns with the model performance comparison shown in Figures 1 and 2 (below), which display R^2 and RMSE scores by model.

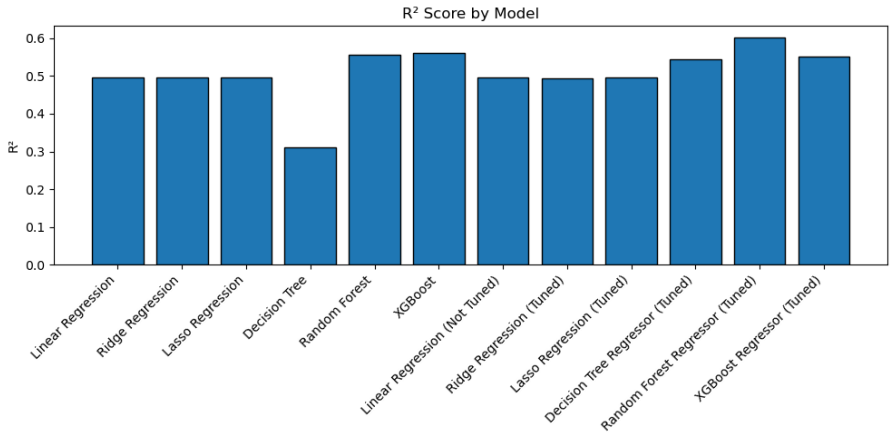

Figure 1: R^2 Score by Model

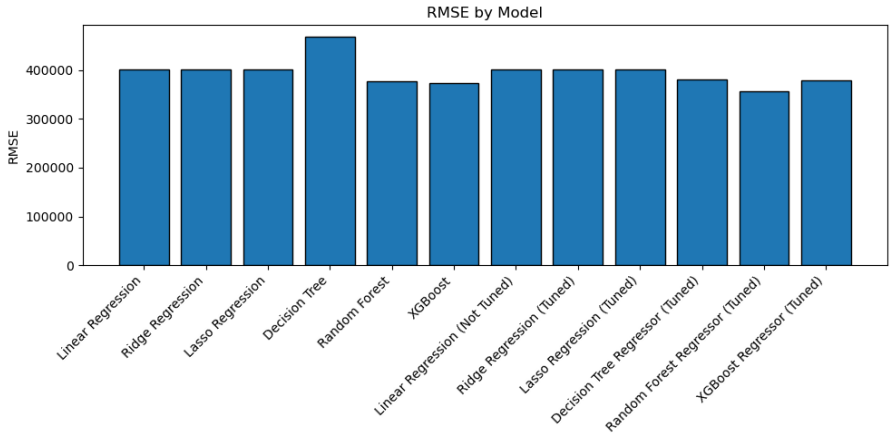

Figure 2: RMSE by Model

8.3 Variable Importance Analysis from Random Forest and XGBoost

We analyzed the feature importances computed during training to gain a deeper understanding of the model's internal decision-making process.

The importance score assigned to each feature reflects how often and significantly it is used to split data across the trees. Features with higher importance values have a greater impact on the model's predictions.

Figures 3 and 4 below show horizontal bar charts of the top contributing variables for Random Forest and XGBoost, respectively.

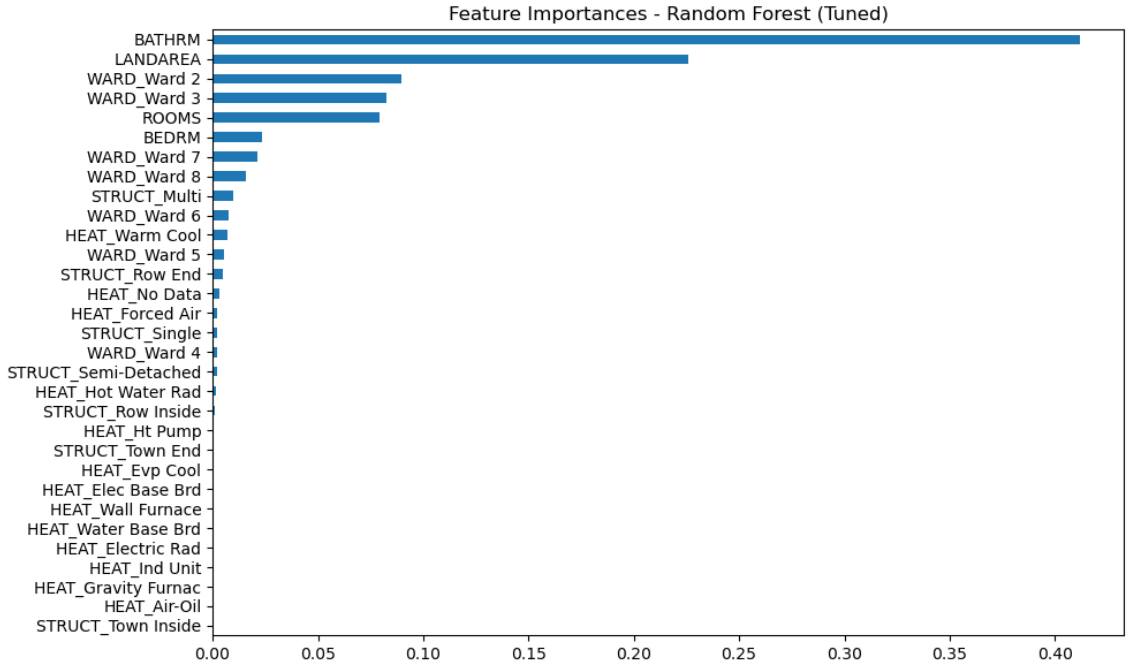

Figure 3: Feature importances from the tuned Random Forest model.

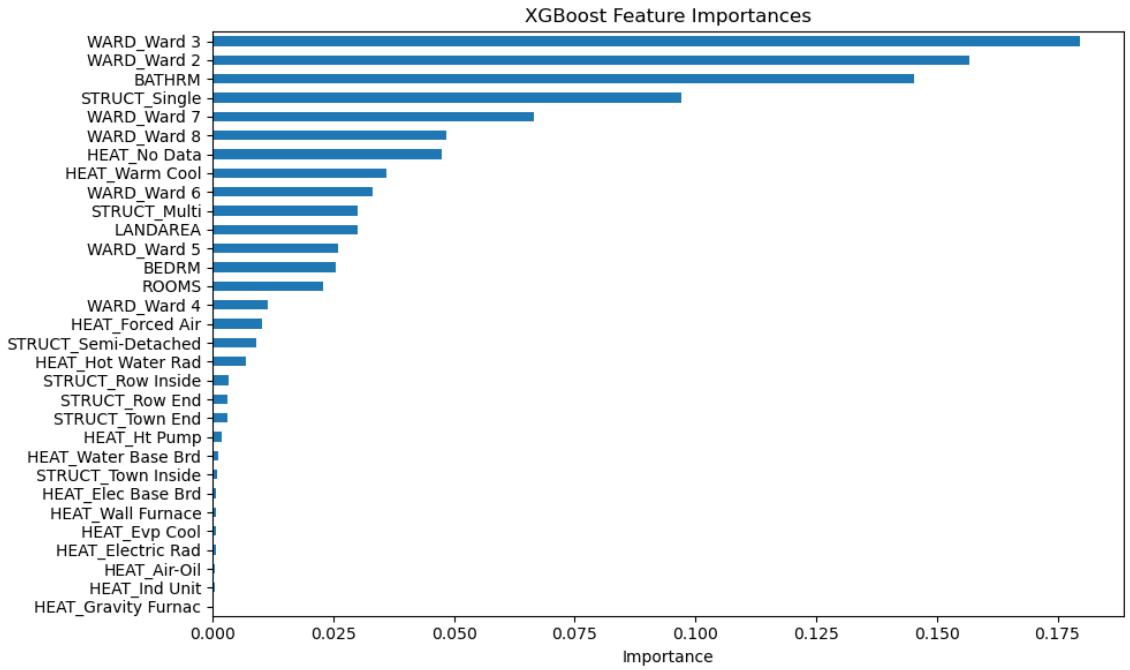

Figure 4: Feature importances from the tuned XGBoost model.

As shown in the figures, the five most influential predictors that strongly correlate with housing prices include:

- For the Random Forest model: BATHRM, LANDAREA, WARD_Ward 3, WARD_Ward 2, and ROOMS.
- For the XGBoost model: WARD_Ward 3, WARD_Ward 2, BATHRM, STRUCT_Single, and WARD_Ward 7.

Figures 5 and 6 illustrate, respectively, the relationship between actual and predicted housing prices using the tuned Random Forest and XGBoost models. The closer the data points are to the diagonal line, the more accurate the model's predictions.

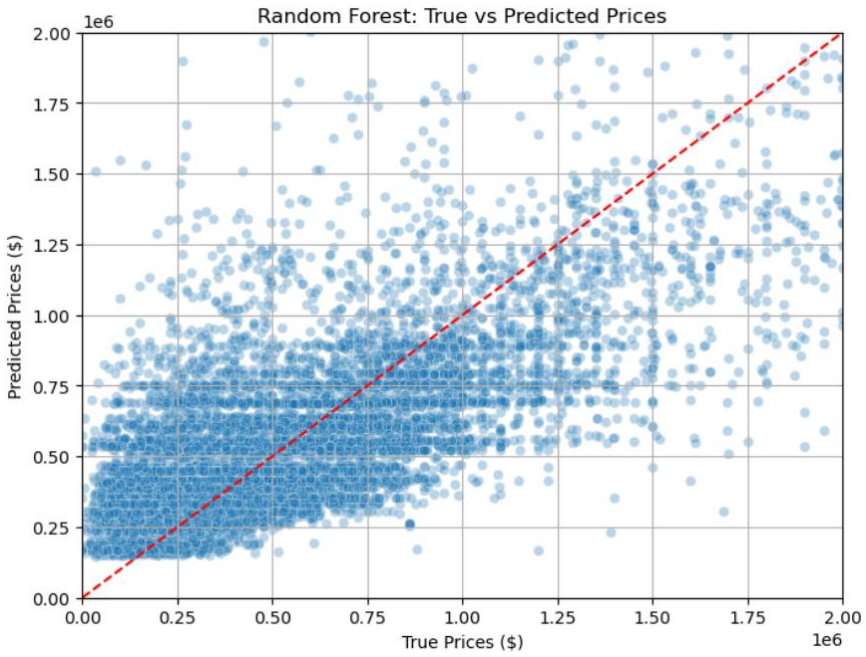

Figure 5: Comparison of proper versus predicted housing prices using the tuned Random Forest model. The red diagonal represents perfect prediction alignment.

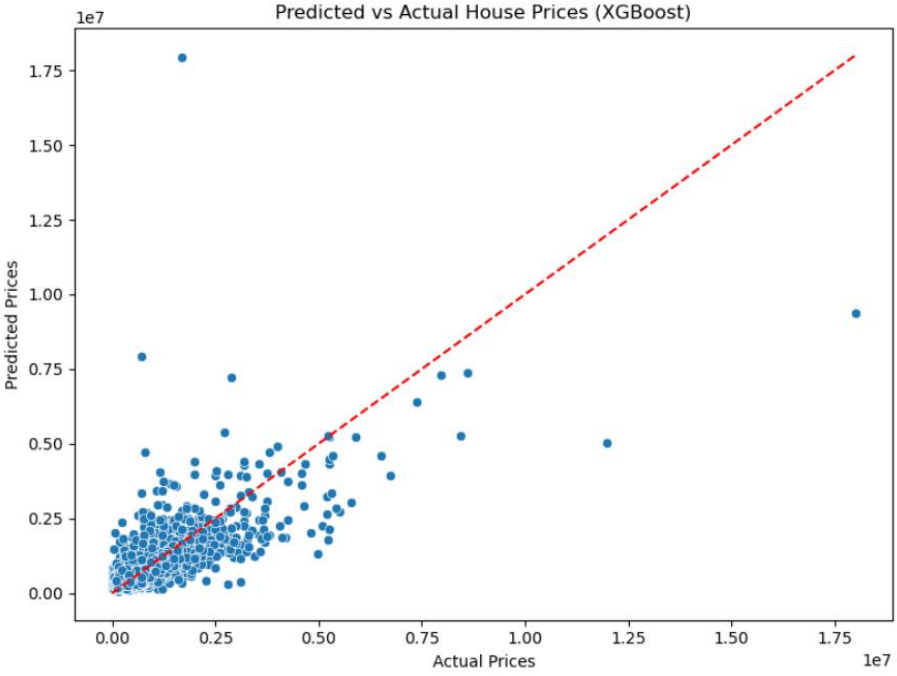

Figure 6: Comparison of actual versus predicted housing prices using the XGBoost model. The red diagonal represents perfect prediction alignment.

9. Conclusion and Perspective

This project successfully predicted residential property prices, with the **tuned Random Forest model emerging as the top performer**. Its superior accuracy and ability to capture non-linear relationships make it the most suitable candidate for deployment.

Here are the key takeaways:

- **Tuned Random Forest** consistently outperforms other models, including linear regressions and decision trees.
- **XGBoost** also demonstrated strong potential and may surpass Random Forest with further optimization.
- While simple and interpretable, **linear models** fail to handle the complex patterns inherent in real estate data.

To proceed with the following steps, we need to:

- Integrate **external data sources** such as crime rates, school quality, and amenities to enrich model features.
- **Deploy the model** using a web application framework such as Flask or through an API for real-time predictions.

In conclusion, this analysis confirms that **tree-based ensemble methods,** particularly the optimized Random Forest, are the most robust, flexible, and accurate for predicting housing prices in a complex urban environment like Washington, D.C. **XGBoost remains a competitive and scalable alternative**, especially for future iterations that require fine-tuning and faster inference.