Using Machine Learning to Predict Housing Prices in DC

Chuyuan Zhong

PPOL 565 Final Project

April 23, 2023

Outline

- Research question and background
- Data sources, target, and important features
- Parametric and non-parametric techniques applied
- Performance and interpretation
- Conclusion and limitations

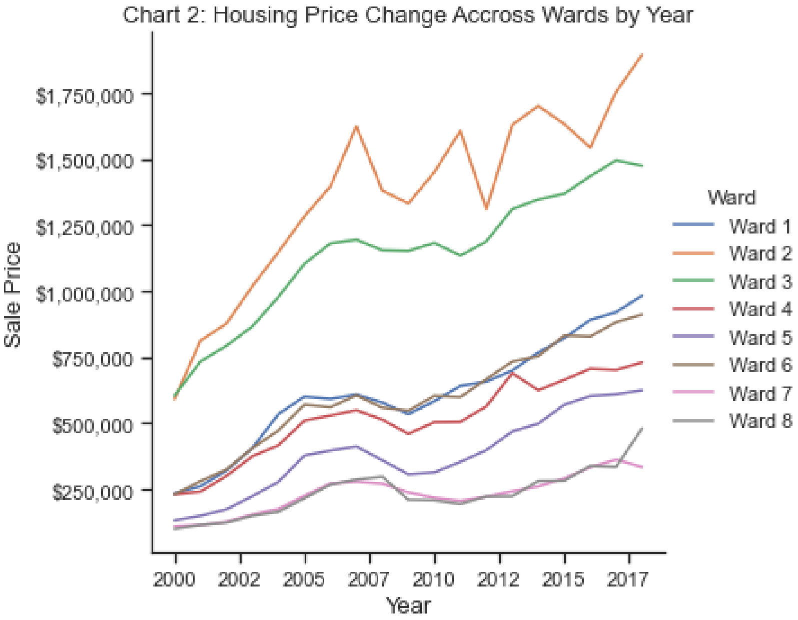
Research Question

What are the key factors that impact housing prices in the Washington DC area?

Background

Housing is an essential element for individuals and the development of a city. In Washington DC, the political and economic hub of the country, high housing prices are a significant issue. In the Washington metro area, the median sales price has risen for the past nine years, increasing by 50% since April 2013. This surge in housing prices has put pressure on low and middle-income households, who spend 30% of their income on housing. The housing market is affected by various factors, including housing and building characteristics, economic conditions, and demographic changes caused by gentrification.

Housing prices across wards



Data sources

- DC_Proporties.csv, obtained from Kaggle, originally from Open Data DC. The dataset contains 158957 observations and 49 features.
- Race_ethnicity of total population by ward.xlsx, provided by DC Action, originally from U.S. Census Bureau.
- Median income of families with children by ward.xlsx, provided by DC Action, originally from U.S. Census Bureau.
- The merged dataset contains 50,983 observations of sales recorded between 2000 and 2018.

Data Preparation

- Target variable: *PRICE* (After log transformation)
- Important features:
  - Internal features: *ROOMS*, *BEDRM*, *BATHRM_TOTAL*, *KITCHENS*, *FIREPLACES*, *SQUARE*, *HEAT*
  - Building characteristics: *Year_after_improved*, *LANDAREA*, *building_age*,*GBA*, *STYLE*, *STRUCT*, *CNDTN*, *EXTWALL*, *ROOF*, *INTWALL*
  - Demographics & Socioeconomic factors: *income($)*, *White*, *Black*, *Asian*, *Hispanic*
  - Others: *WARD*, *SALEDATE*

Method

- Parametric model:
  - LASSO
- Non-parametric models:
  - Decision Tree
  - Random Forest
  - XGBoost

Performance

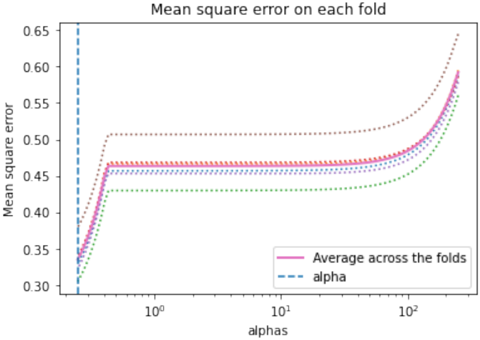


Figure 2: LASSO MSE Curve

Metrics	Score
MAE	0.427
MSE	0.324
RMSE	0.569
R^2	0.442

Table 1: LASSO Performance

Performance

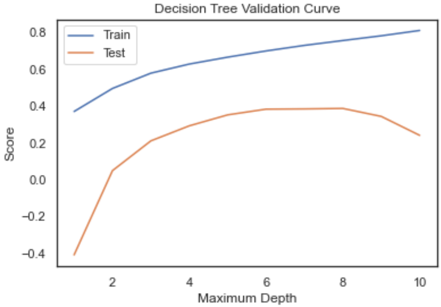


Figure 3: Decision Tree Validation Curve

Metrics	Score
MAE	0.280
MSE	0.172
RMSE	0.415
R^2	0.703

Table 2: Decision Tree Performance

Performance

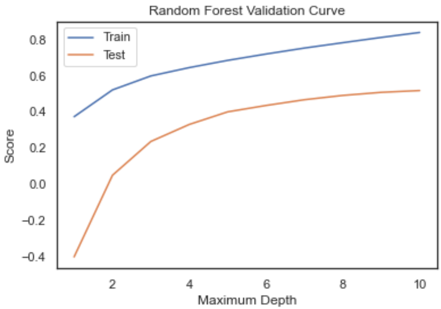


Figure 4: Random Forest Validation Curve

Metrics	Score
MAE	0.263
MSE	0.145
RMSE	0.381
R^2	0.737

Table 3: Random Forest Performance

Performance

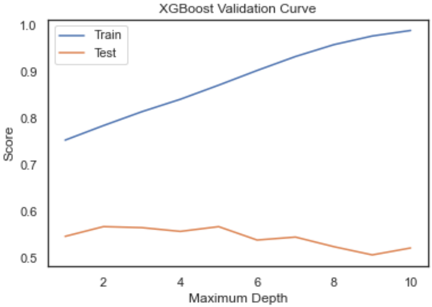


Figure 5: XGBoost Validation Curve

Metrics	Score
MAE	0.212
MSE	0.113
RMSE	0.336
R^2	0.807

Table 4: XGBoost Performance

Performance

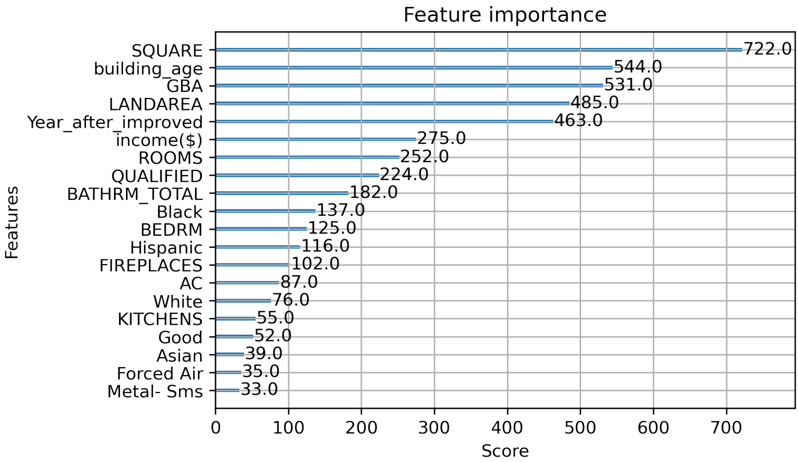


Figure 6: Feature Importance Ranking

Interpretation

- The optimal alpha value is 0.25 which achieved the lowest MSE of 0.32, indicating moderately regularized
- Decision Tree tends to overfit after a depth of 8
- XGBoost has the lowest MAE, MSE, RMSE, and the highest R^2, but it tends to overfit
- The Random Forest model seems to be the optimal model
- Important features are *SQUARE*, *building_age*, *GBA*, *LANDAREA*, *income($)*, *year_after_improved*

Conclusion

Housing prices in Washington, D.C. are largely influenced by internal features, such as house size, and external architectural characteristics, such as building footage, age, and year of renovation, as well as income level.

Limitations

- Highly right-skewed target variable may affect the accuracy of predictions
- Missing income and race data from 2001-2009 may limit the representativeness of the dataset
- The presence of many outliers in the dataset may negatively affect the performance of the models
- The limited number of features (such as poverty data, interest rate) may not fully capture the complexity of the housing market, omitted variable bias
- Regional differences