Property Valuation Using Machine Learning

Mmesoma Udensi, Flavien Foreste, Matteo Scarpa

December 5, 2024

Executive Summary

This project supports the Cook County Assessor's Office (CCAO) in its ongoing efforts to modernize property valuation. Our objective was to use historical property transaction data to predict fair market values for 10,000 residential properties, using advanced machine learning techniques. By minimizing the Mean Squared Error (MSE) of our predictions, we aimed to improve accuracy, transparency, and fairness in property assessments, which directly impact public funding for services.

Methodology

1. Data Understanding and Preparation

1.1. Data Sources

Two datasets were provided: a historic dataset with 50,000 sales records for training and a dataset of 10,000 properties requiring prediction. A codebook offered definitions and descriptions for all variables.

1.2. Data Cleaning

We cleaned missing data using median imputation for numeric fields and the label "Unknown" for categorical fields. Logical (binary) variables were converted to factors, and irrelevant columns like `ind_garage` were removed to reduce noise.

1.3. Feature Selection

A correlation threshold of 0.4 was applied to numeric predictors, and ANOVA tests identified significant categorical variables. This hybrid approach ensures that the model focuses on features that strongly influence sale price while maintaining interpretability and statistical rigor.

Table 1: Significant Numeric Variables

Variable Name	Description (if applicable)
`sale_price`	Target variable (property sale price)
`meta_certified_est_bldg`	Certified estimated value of building
`meta_certified_est_land`	Certified estimated value of land
`char_frpl`	Proportion eligible for free/reduced-price lunch
`char_fbath`	Number of full bathrooms
`char_bldg_sf`	Square footage of the building
`geo_white_perc`	Percentage of white residents in the area
`econ_midincome`	Proportion of middle-income households

Table 2: Significant Categorical Variables

Variable Name	Description (if applicable)
meta_class	Property classification code
meta_town_code	Town code identifier
meta_nbhd	Neighborhood designation
meta_cdu	Condition/Desirability/Utility rating
meta_deed_type	Type of property deed
char_ext_wall	Type of exterior wall material
char_roof_cnst	Roof construction type
char_heat	Type of heating system
char_use	Property usage category
geo_property_city	City where the property is located
geo_property_zip	Property ZIP code
geo_municipality	Municipality name
geo_fips	FIPS code for geographic location
geo_school_elem_district	Elementary school district
geo_school_hs_district	High school district
ind_large_home	Indicator for large homes
ind_arms_length	Indicator for arm's-length transactions

2. Model Development and Validation

2.1. Baseline Model

A linear regression model was trained using 5-fold cross-validation. This model provides transparency and serves as a benchmark for more complex algorithms.

Table 3: Linear Regression Residual Summary

	Min	1Q	Median	3Q	Max
Residuals	-1,230,537	-40,578	-326	36,493	2,231,931

Table 4: Linear Regression Coefficient Estimates

Variable	Estimate	Std. Error	t value
Intercept	-261,500	204,200	-1.280
meta_certified_est_bldg	0.7154	0.006416	111.494

2.2. Random Forest Model

We applied one-hot encoding to handle categorical variables, then trained a Random Forest model with 10 trees. Random Forest was chosen for its robustness, ability to capture nonlinear relationships, and resistance to overfitting.

2.3. Validation

Models were evaluated using RMSE and R^2. The Random Forest model achieved the lowest RMSE and highest explanatory power, making it our preferred model for predictions.

Table 5: Random Forest Variable Importance (ntree = 10)

Variable	Importance Score (%)	Contribution to Model Splits
meta_certified_est_bldg	11.934	2.02×10^{15}
meta_certified_est_land	5.876	3.74×10^{14}
char_frpl	2.323	1.81×10^{13}
char_fbath	2.688	1.31×10^{14}
char_bldg_sf	5.334	1.87×10^{14}

3. Prediction Phase

After aligning the structure of the prediction dataset with the training data, the Random Forest model was used to generate sale price predictions. The output file contains two columns: property ID (pid) and predicted value (assessed_value). This file can be submitted directly to the evaluation platform.

Results

Model Comparison

- Linear Regression: RMSE = 124,135.7; $R^2 = 0.8378$
- Random Forest: RMSE = 122,593.5

The Random Forest model slightly outperformed the Linear Regression model by better capturing complex variable interactions and non-linear patterns.

Conclusion

The Random Forest model provided the most accurate predictions. The output file includes 10,000 property IDs with corresponding assessed values.

Summary statistics of assessed values

Table 6: Summary Statistics of Predicted Assessed Values

Statistic	Value ($)
Minimum	50,000.00
1st Quartile (Q1)	266,898.01
Median (Q2)	317,147.68
Mean	317,188.03
3rd Quartile (Q3)	367,673.37
Maximum	611,810.13
Inter-quartile Range (IQR)	100,775.36

These values align with realistic market conditions and reflect the model's ability to generalize from historical patterns.

Appendix

- **Table 7:** Variable data types before and after cleaning
- **Figure 1:** Distribution of predicted sale prices

Table 7: Variable Data Types Before and After Cleaning

Variable Name	Original Data Type	Cleaned Data Type
`meta_cdu`	`chr` (character)	`Factor w/ 13 levels`
`meta_deed_type`	`chr` (character)	`Factor w/ 13 levels`

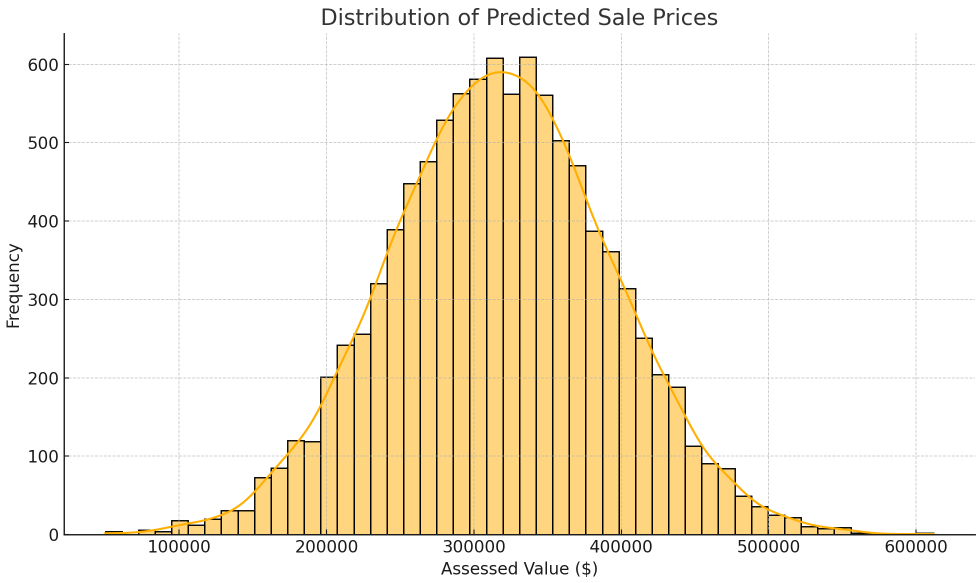

Figure 1: Distribution of Predicted Sale Prices