

Team Name: House for \$900???

Group Submission: Wuji Mike Chen, Sharang Mantri, Song Yi Qiu

Case Overview

The Cook County Assessor's Office (CCAO) is responsible for assigning fair market values to houses in order to determine the property tax bills. However, manually evaluating each one would be costly and impractical. As such, this project seeks to find a data-driven approach to evaluate properties in an efficient manner.

In this case, we were provided with a dataset of residential property transactions with features such as observed sale prices and detailed property characteristics (e.g., location, lot size, building attributes, and condition). Our objectives as data scientists were to clean and explore the data to understand key drivers of sale price, develop and compare predictive models for sale price, evaluate model performance, and select a final model that we will use to predict the prices for 10,000 unlabeled properties.

Methodology

Data Cleaning and Processing

Data quality was paramount for model performance. We adopted a tailored imputation strategy: missing values for **amenities** (e.g., AC, fireplace, basement) were set to 0 (treated as "not present"), while skewed numeric features (e.g., building size, age, income) were imputed with the **median** to prevent outliers from distorting the typical property profile.

To capture location effects, we engineered a high-impact feature, **avg_district_price**, by calculating the **median sale price** for each elementary school district. This target encoding transformed high-cardinality categorical data into a powerful numerical signal representing

neighborhood value. Additionally, we combined full and half baths into a single **char_baths** feature to align with buyer valuation metrics.

Finally, we addressed **non-arms-length transactions** (often nominal sales, e.g., \$907). Instead of discarding these outliers, we utilized **One-Hot Encoding** to convert the status into a binary numeric feature (ind_arms_lengthTRUE). This allowed the model to explicitly learn the systematic price discount associated with non-market transfers, thereby maximizing data utilization without confusing the model's market-value logic.

Modeling

After data processing was complete, we then began modelling. Our data (50,000 samples) was split into a training set (40,000 samples) and validation set (10,000 samples). We trained our models on the training set and compared their performance on the validation set to prevent any risk of overfitting. The target we model is the log of sale price, rather than the raw price, for the statistical models (linear regression and random forest). Taking logs reduces skewness, stabilizes the variance across cheap and expensive homes, and helps the model capture proportional differences in value. Predictions are then exponentiated back to the dollar scale for reporting and error calculation.

To reduce training time, model complexity and improve model transparency, we chose 15 features that combine location, physical attributes, socio-economic context, and structural details (see appendix for list).

Model evaluation and selection

For all three models, performance was evaluated on the cleaned training set using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Our overall results were as follows:

Model	RMSE	MAPE
Linear Regression	237938.83	45.06 %
Neural Network	114122.63	34.71 %
Random Forest	108596.91	32.98 %

The non-linear models substantially reduced error relative to the linear baseline. Between the neural network and the random forest, the random forest achieved lower percentage error and provides clearer diagnostics (OOB learning curve, feature importance plots).

Given this combination of accuracy, robustness, and interpretability, we selected Model C, the enhanced random forest, as the champion model for generating assessed values.

Application to scoring data

Finally, the champion Random Forest model was deployed to predict `log(sale_price)` for the 10,000 properties in the **prediction dataset**, strictly adhering to the same cleaning and feature-engineering pipeline used for training to ensure consistency. These predictions were converted back to dollar values via an exponential transformation.

The final output was exported as **assessed_value.csv**, containing the required **pid** and **assessed_value columns**. These predicted fair market values can now serve as a benchmark for the CCAO to identify potential assessment discrepancies and improve valuation equity.

Conclusion

Results and Assessment Distribution The final output is reported in **assessed_value.csv**, containing predicted market values for all **10,000** properties. The distribution ranges from a **minimum of \$16,507** to a **maximum of \$2,308,127**, with a **mean of \$291,701** and a **median of \$226,887**. The interquartile range lies between **\$130,158** and **\$354,974**.

Model Performance Our Random Forest model achieves an **RMSE of \$108,597** and an **R² of ~78%**. Feature importance analysis identifies our engineered **avg_district_price** as the dominant predictor. **Demographic and socio-economic factors also rank as top drivers, alongside Building Size (char_bldg_sf)**.

Recommendations

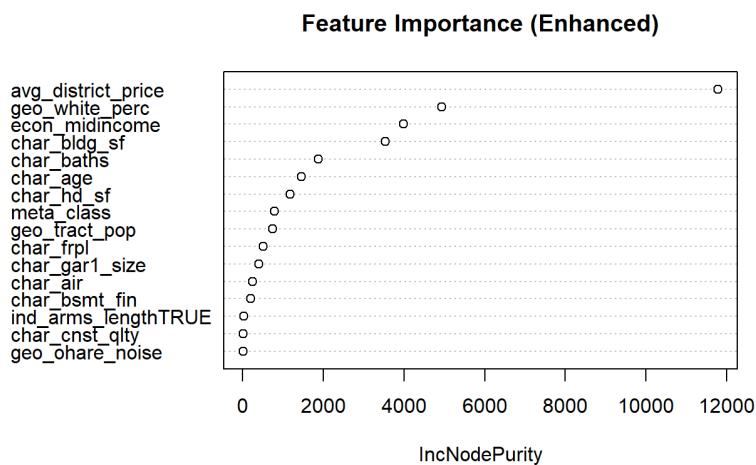
1. **Institutionalize Target Encoding:** Formalize median-based district pricing to systematically capture location premiums.
2. **Prioritize Structural Audits:** Focus field audits on verifying **Square Footage** and **Construction Quality**, as data accuracy here yields the highest valuation precision.
3. **Segregate Non-Market Transactions:** Maintain a separate workflow for non-arms-length transactions to prevent nominal transfers from distorting fair market valuations.

Appendix:

Figure A1: Feature List

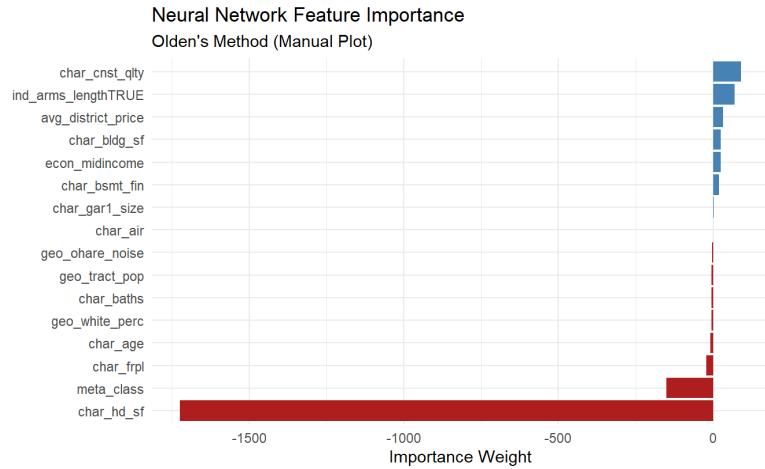
- Location: avg_district_price, geo_ohare_noise
 - Physical characteristics: char_bldg_sf, char_hd_sf, char_age, char_baths, char_gar1_size, char_air, char_frpl, char_cnst_qty
 - Socio-economic and geographic context: econ_midincome, geo_tract_pop, geo_white_perc
- Structural/administrative: char_bsmt_fin, meta_class, ind_arms_lengthTRUE

Figure A2: Random Forest Feature Importance



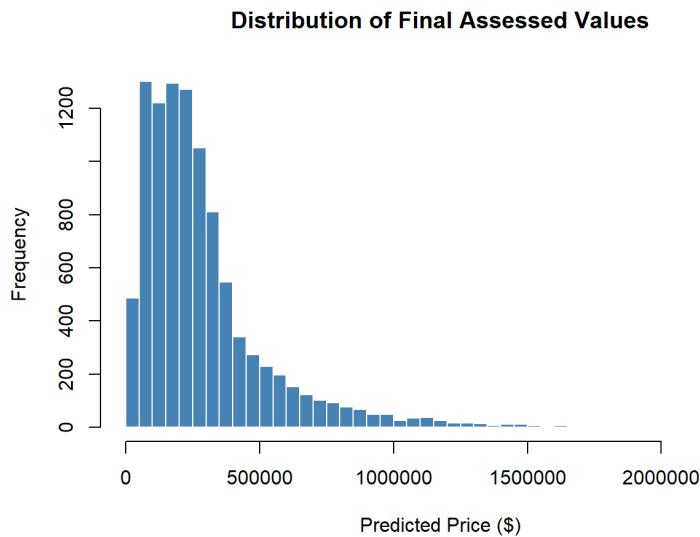
This plot illustrates the relative predictive power of each variable. Note that the engineered feature **avg_district_price** is the dominant predictor, confirming the effectiveness of our target encoding strategy.

Figure A3: Neural Network Feature Importance



This plot confirms our findings from the Random Forest model. Specifically, variables like **avg_district_price** and **char_bldg_sf** show the strongest influence on price, validating that the neural network is learning meaningful economic relationships rather than just memorizing noise.

Figure A4: Distribution of Final Assessed Values (*Histogram showing the spread of predicted market values for the 10,000 test properties.*)



The distribution of assessed values is **right-skewed**, which is consistent with typical real estate market data. The values are centered around a median of approximately **\$226,887**, with a long tail extending towards higher-value properties (max ~\$2.3M). Crucially, **all predicted values are positive**, meeting the CCAO's submission requirements.

A.5 Feature Engineering Logic: Average District Price Formula

$$AvgDistrictPrice_k = Median(\{SalePrice_i \mid District_i = k, i \in TrainingSet\})$$