House Price Predictor DATA 100

Fall 2024 Ajay Sharma

Remark

If you would like to see the project source code (Jupyter Notebook), please contact me at ajay.sharma@berkeley.edu.

Overview

This project aimed to develop a predictive model for housing prices in Cook County, Illinois. By leveraging a robust dataset, the goal was to identify key features driving housing prices and build an accurate model with minimal root mean square error (RMSE). The project provided actionable insights into the dynamics of the real estate market.

Dataset Description

The analysis utilized the Cook County Assessor's Office (CCAO) dataset with over 100,000 residential properties. Key attributes included:

- **Property Features:** Number of bedrooms, bathrooms, square footage, and lot size.
- **Location Data:** Neighborhood identifiers, geographic coordinates, and proximity to amenities.
- **Market Trends:** Historical sale prices and tax assessment records.

Methodology

- **Exploratory Data Analysis (EDA):** Used `pandas`, `matplotlib`, and `seaborn` to analyze trends and visualize relationships between features and sale prices.
- **Data Cleaning:** Addressed missing values (imputation), removed outliers, and normalized features.
- **Feature Engineering:** One-hot encoded categorical variables, normalized numerical features, and removed multicollinear predictors.
- **Model Training:** Trained multiple regression models, tuned hyperparameters, and validated performance with cross-validation.
- **Evaluation:** Assessed model performance using RMSE and residual analysis.

Results and Insights

The final model achieved an RMSE $<$ 240K, ranking top 70 out of a class of 1200. Key findings included:

- **Key Predictors:** Property size, number of bedrooms, and proximity to urban centers were the strongest drivers of price.
- **Market Trends:** Significant variability in pricing across neighborhoods highlighted the critical role of location.