Predicting and Grouping Apartment Building Evaluation Scores in Toronto: A Linear Regression and Decision Tree Approach

Ali Sheikh Rabiei

March 3, 2025

1 Introduction

Evaluating apartment buildings is essential for ensuring compliance with safety, maintenance, and livability standards. This project explores the relationships between building characteristics (e.g., age, size, amenities, location) and their evaluation scores using two machine learning approaches:

- **Linear regression** to predict evaluation scores based on building features.
- **Decision tree modeling** to categorize buildings and identify key factors influencing score variations.

By identifying key factors influencing evaluation scores, this analysis can empower policymakers to allocate resources effectively, guide property managers in prioritizing renovations, and ensure safer living conditions for residents.

2 Research Questions

1. Which building features (e.g., year built, number of units, amenities) correlate most strongly with evaluation scores?
2. Can a linear regression model accurately predict evaluation scores based on these features?
3. How does a decision tree model categorize buildings, and what insights do these groupings provide about feature importance?
4. Does proximity to public transit or green spaces correlate with higher evaluation scores?

3 Dataset

Source: Open Data Toronto – Apartment Building Evaluations
Key Variables:

- `SCORE` (**Target**): Continuous evaluation score (0–100).
- `YEAR_BUILT`: Year of construction.
- `NUMBER_OF_STOREYS`: Building height.
- `NUMBER_OF_UNITS`: Total residential units.
- `PARKING_SPACES`: Number of parking spots.
- `ELEVATOR`: Binary (Yes/No).

Extended Variables (Optional):

- `DISTANCE_TO_SUBWAY`: Distance to nearest subway station (from TTC Subway Data).
- `PARK_PROXIMITY`: Distance to nearest public park (from Toronto Parks Data).

4 Methodology

4.1 Data Preparation

- Handle missing values (e.g., impute `YEAR_BUILT` using the median).
- Convert categorical variables (e.g., `ELEVATOR` to binary 0/1).
- Geocode building addresses to calculate proximity to subway stations and parks.

4.2 Exploratory Data Analysis (EDA)

- Visualize distributions using histograms and assess correlations using heatmaps.
- Identify nonlinear relationships (e.g., age vs. score decay) using scatterplots.
- Plot proximity variables against scores to test location-based hypotheses.

4.3 Modeling

- **Linear Regression**: Predict `SCORE` using scikit-learn's `LinearRegression`. Validate with R^2, MSE, and residual plots.
- **Decision Tree**: Train a regression tree (`DecisionTreeRegressor`) and analyze feature importance via SHAP values for interpretability.

5 Potential Challenges

- Missing data in critical fields (e.g., `YEAR_BUILT`).
- Geocoding inaccuracies when merging subway/parks datasets.
- Computational demands of SHAP analysis for large decision trees.

6 Libraries

- `pandas`, `numpy`: Data cleaning and manipulation.
- `scikit-learn`: Machine learning models and cross-validation.
- `shap`: Explainable AI for interpreting decision trees.
- `geopandas`: Geospatial data processing (for proximity variables).

7 Conclusion

This study aims to provide actionable insights into factors influencing apartment building evaluation scores, including potential location-based effects. By leveraging linear regression for prediction and SHAP-enhanced decision trees for interpretability, the project bridges technical analysis with practical urban planning needs. Future work could integrate satellite imagery or tenant surveys to capture qualitative factors.

Approval Checklist

- ■ Dataset is publicly available on Open Data Toronto.
- ■ Methods align with course requirements (linear regression and tree models).
- ■ Analysis includes prediction and feature interpretation.

Acknowledgments

After completing the initial draft, I used Grammarly to refine grammar, formatting, and wording errors. Additionally, I utilized ChatGPT to debug LaTeX-related issues and brainstorm creative analytical approaches.