Real Estate Price Prediction Report

Zixiang Huang, Zhongze Zheng, Zhengyu Wang, Yuqing Liu

March 2, 2025

1 Introduction

This report presents a detailed analysis and modeling approach for predicting real estate prices using a combination of geographic data, machine learning models, and advanced ensemble techniques.

2 Assumptions for Real Estate Price Prediction

1. **Timeframe of the Data**

   The dataset lacks an explicit collection year, but since real estate prices change gradually, we assume it represents a recent timeframe. Minor misalignment with external datasets (e.g., crime rates) is unlikely to affect model performance.

2. **Crime Data and Trend Consistency**

   Crime rates have shown a consistent yearly increase. Using the most recent crime data (2024) for all listings is reasonable, as past crime levels can be estimated based on proportional growth.

3. **External Data (Infrastructure Locations)**

   Infrastructure data (hospitals, police stations, subway stations) from OSM is assumed to be stable over time, making it a valid representation of public services' availability during the dataset's timeframe.

4. **Model Generalizability**

   The relationship between real estate prices and external factors (e.g., crime rate, infrastructure) is assumed to be stable across similar urban areas, ensuring the model remains applicable for future predictions under similar conditions.

5. **Model Selection (Decision Tree & XGBoost)**

   Housing prices follow a **non-linear** pattern, making Decision Trees more suitable than linear models. XGBoost improves prediction accuracy ($R^2 = 0.91$). We assume the cleaned dataset has no extreme outliers that could hinder model performance.

6. **Economic Factors and Market Stability**

   We assume economic factors (inflation, interest rates) have minimal impact, with housing prices primarily driven by geography and social conditions. However, extreme events (e.g., financial crises) may limit the model's applicability.

Field Name	Description	Example
id	Unique six-digit identifier for each listing	123456
ward	The ward in which the property is located	W10
beds	Number of bedrooms in the property	3
baths	Number of bathrooms in the property	2
DEN	Indicates if the property includes a den	Yes
size	Property size in square feet	500-999 sqft
parking	Indicates if the property includes a parking space	Yes
exposure	Direction the property faces	South
D mkt	Days on the market	30
building age	Age of the building in years	10
maint	Monthly maintenance fee (C$)	500
price	Listing price (C$)	750000
lt/lg	Latitude and longitude coordinates	(43.6532, -79.3832)
minHighwayDis	Minimum distance to the nearest highway	1.2 km
minMallDis	Minimum distance to the nearest mall	0.5 km
minParkDis	Minimum distance to the nearest park	0.3 km
minPoliceDis	Minimum distance to the nearest police station	2.1 km
minSchoolDis	Minimum distance to the nearest school	0.8 km
minStationDis	Minimum distance to the nearest subway station	1.0 km
hasHospital	Whether there is a hospital nearby	Yes
hasMall	Whether there is a mall nearby	Yes
hasPark	Whether there is a park nearby	Yes
hasPolice	Whether there is a police station nearby	Yes

Table 1: Input Features for Real Estate Price Prediction Model

3 Inputs

4 Mathematical Modeling

4.1 Step 1: Initial Data Cleaning and Visualization

We conducted preliminary data cleaning and visualization. Using K-means clustering, we segmented the data into different regions based on geographic coordinates. We observed distinct price patterns in different regions, prompting further exploration of additional influencing factors(Figure 1).

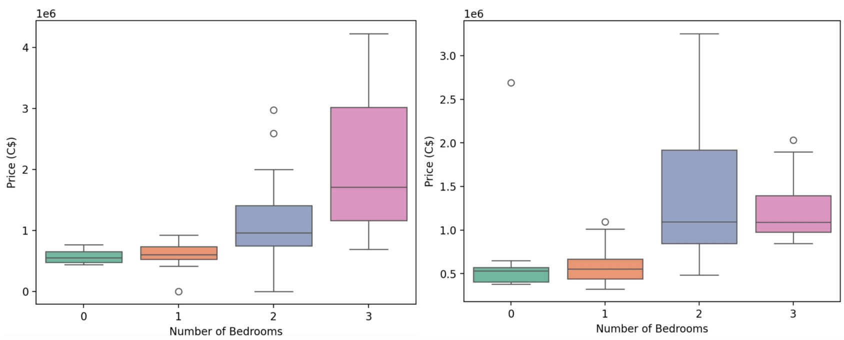

Figure 1: Comparison of the Price Relative to the Number of Bedrooms Based on the Region and Exposure of Houses(Left: region 15, north; Right: region 18, west)

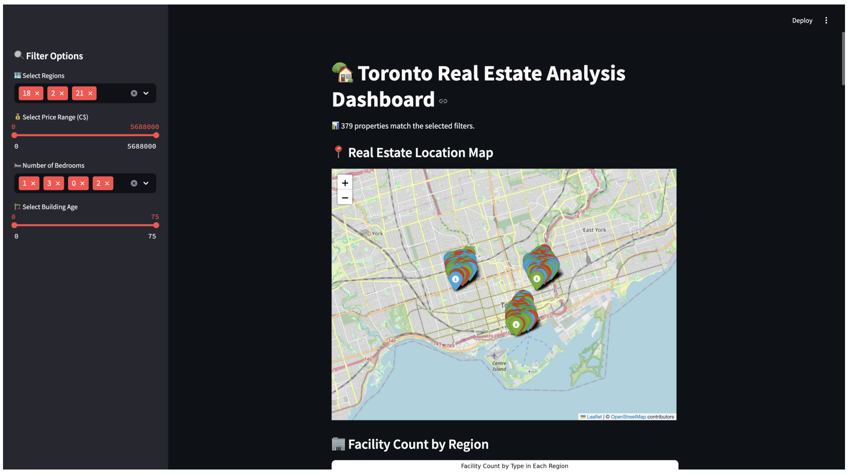

Figure 2: Dynamic Visualization of Housing Information

4.2 Step 2: Integration of Geographic and External Data

We utilized OpenStreetMap (OSM) to extract geographic data, including infrastructure such as hospitals, police stations, and subway stations (Figure 2). We also integrated crime data from Toronto Open Data, filtering by category (e.g., violent crimes, property crimes) and merging the crime rates into our dataset.

4.3 Step 3: Model Development

We trained and compared various models, including linear regression, lassoCV regression (Figure 3), neural networks, and Random Forest (Figure 4). After comparing accuracy, error metrics, and theoretical alignment, we found that the Decision Tree model best fit our case. We developed a idea of a regional ensemble approach using separate Decision Tree models for each region.

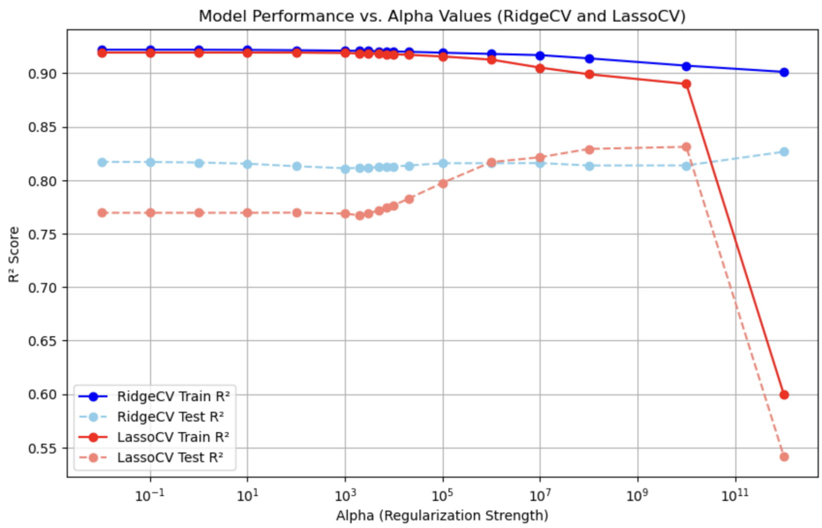

Figure 3: Error of lassoCV and RidgeCV Model with different alpha

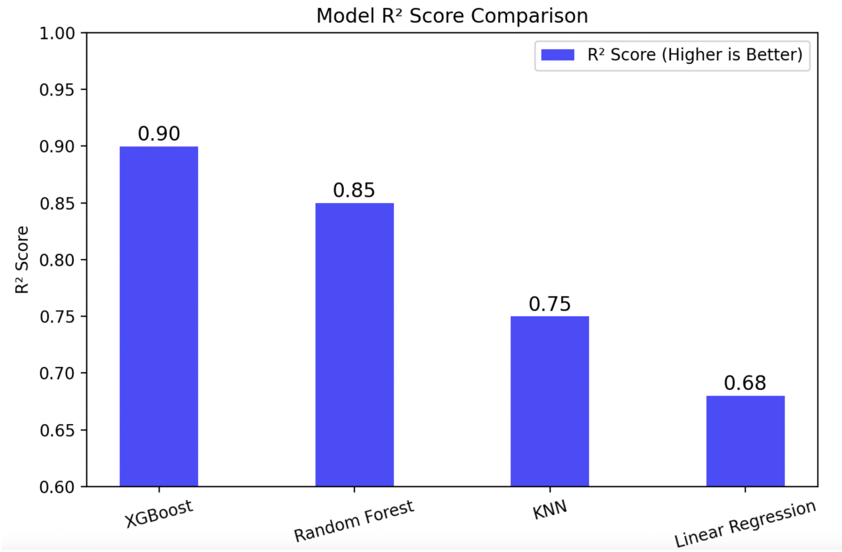

Figure 4: Comparison Between Different Types of Model

4.4 Step 4: Model Evaluation and Improvement

We compared a single Decision Tree model with our regional ensemble approach. (Figure 5, 6) The ensemble model outperformed the single model, aligning with XGBoost's principles. Our XGBoost model achieved a promising $\mathbf{R^2}$ score of 0.91 on the test set.

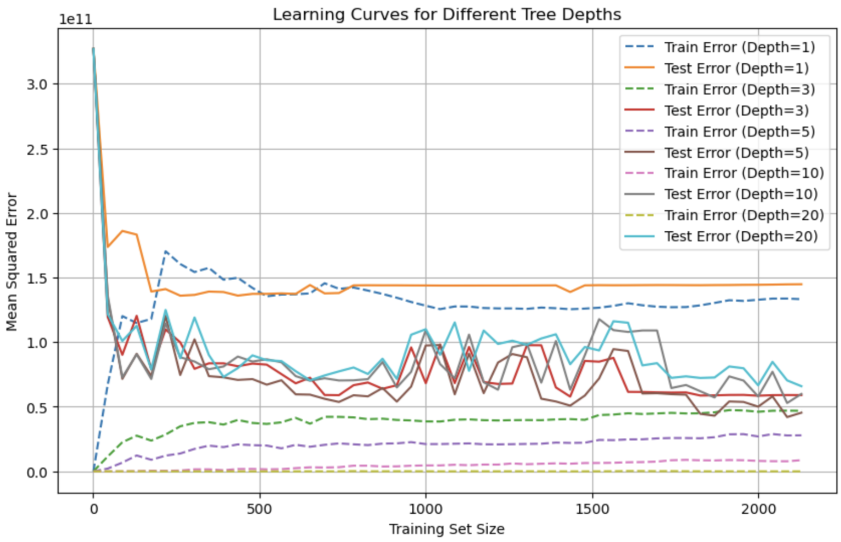

Figure 5: Learning Curve of Single Decision Tree With Different Depth

5 Future and Limitation

From the correlation table between price and all features, our model revealed a feature imbalance. To enhance model diversity and robustness, we propose using the Random Subspace Method for future adjustments, which could mitigate feature dependency and improve generalization.

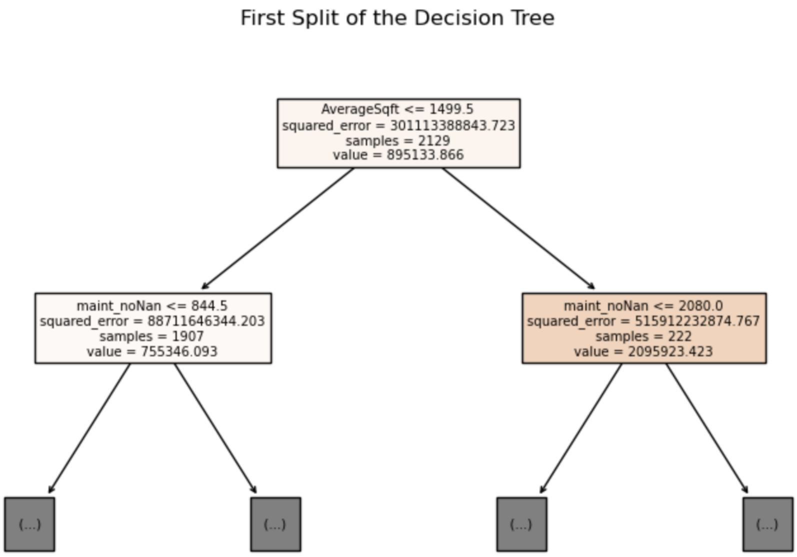

Figure 6: First Split of the Single Decision Tree

6 Business Analysis

Our analysis demonstrates the efficacy of using regional segmentation combined with XGBoost for predicting real estate prices. Future work will focus on diversifying the model using advanced ensemble methods like the Random Subspace Method to further optimize performance and stability.