Toronto Real Estate Price Prediction Report

Natalia George, Meer Sisodia, Vraj Shah Raghav Bhatia

March 2, 2025

1 Introduction

This report details a project focused on predicting real estate prices in Toronto. The project utilizes a comprehensive dataset of housing features to perform exploratory data analysis, rigorous preprocessing, and feature engineering. We employ and evaluate several machine learning regression models, including Random Forest, Extra Trees, XGBoost, and a Feed Forward Neural Network, alongside an ensemble approach to optimize prediction accuracy. The primary goal is to develop a robust model capable of accurately estimating Toronto housing prices based on property characteristics.

2 Methodology

The project follows a structured methodology encompassing:

1. **Exploratory Data Analysis (EDA):** Initial data inspection, price distribution visualization, correlation analysis, and missing value identification were performed to understand the dataset's characteristics. Geographic price visualization on a Toronto map was also conducted.
2. **Data Preprocessing:** A data pipeline was implemented to clean and preprocess the data. This includes:
    - Removal of irrelevant columns and rows with missing price values.
    - Encoding of categorical features (size, exposure, DEN, parking, ward) into numerical formats.
    - Imputation of missing values using KNN-based imputation for categorical features and MICE (IterativeImputer) for numerical features.
3. **Feature Engineering:** A key feature, `local_avg_price`, was engineered to capture spatial price dynamics. This feature calculates the average price of neighboring houses within a 700-meter radius, considering properties with the same number of bedrooms, leveraging a BallTree for efficient spatial queries.
4. **Model Fitting and Evaluation:** The dataset was split into training and testing sets (80/20). Four regression models and an ensemble model were trained:
    - Random Forest Regressor
    - Extra Trees Regressor
    - XGBoost Regressor
    - Feed Forward Neural Network (Keras)
    - Ensemble Model (Averaging Predictions)

    Model performance was evaluated using Root Mean Squared Error (RMSE) and a custom accuracy metric (predictions within 15% of the actual price).

3 Models and Evaluation Metrics

The following models were implemented and evaluated:

- **Random Forest Regressor**
- **Extra Trees Regressor**
- **XGBoost Regressor**
- **Feed Forward Neural Network (Keras)**
- **Ensemble Model** (Mean of the above models)

Model performance was assessed using:

- **Root Mean Squared Error (RMSE):** A standard metric for regression models, quantifying the difference between predicted and actual values.
- **Custom Accuracy Metric:** Percentage of predictions falling within 15% of the actual house price, providing a practical measure of prediction accuracy.

4 Results

Model performance, measured by RMSE, is summarized in Table 1. The Ensemble model achieved the lowest RMSE, indicating superior performance in terms of this metric. However, under the custom accuracy metric, the Random Forest Regressor demonstrated competitive performance, achieving a custom accuracy of approximately 70%.

Model	**RMSE**
Random Forest	165249.268160
Extra Trees	174942.773153
XGBoost	172803.486092
Neural Network	165809.889018
Ensemble (Mean)	164624.098482

Table 1: Model Performance Comparison (RMSE)

Visualizations of predicted vs. actual prices and error distributions for each model were generated to further analyze model behavior and prediction accuracy (see plots in `datathon.py` output).

5 Conclusion

This project successfully developed and evaluated several machine learning models for predicting Toronto real estate prices. The Ensemble model and Random Forest Regressor demonstrated strong predictive capabilities. The engineered `local_avg_price` feature, capturing local market dynamics, likely contributed significantly to model performance. The project provides a valuable framework for real estate price prediction and insights into key factors influencing housing valuation in Toronto.

Note: Artificial Intelligence tools were utilized during code development to enhance efficiency and streamline certain processes, particularly in data handling and model implementation.

6 Dependencies

The following Python libraries are required to run the `datathon.py` script:

- pandas
- math

- numpy
- matplotlib
- seaborn
- scipy
- scikit-learn (sklearn)
- xgboost
- keras
- plotly_express