Are We Building Enough? Forecasting Canadian Housing Starts and Housing Adequacy

Executive Summary

Erdős Institute Data Science Bootcamp, Fall 2025

Anwesha Basu, Debanjan Sarkar

This project investigates how quarterly housing starts across Canadian provinces can be forecasted from 1990 to 2025. Housing starts — the number of new residential construction projects initiated within a given period — are a key indicator of housing supply and overall economic activity. In recent years, Canada's rapid population growth and persistent affordability challenges have underscored a widening housing supply gap. By forecasting housing starts, we aim to develop simple, interpretable, and reproducible models that identify emerging supply trends and reveal how demographic and macroeconomic factors drive construction activity.

Data and Features

Our main target variable is the number of **housing starts**, measured as seasonally adjusted annual rates (SAAR) for each province. To explain variations in housing starts, we used quarterly population data from Statistics Canada, focusing on two key features:

- **Population change** — the difference in total population between consecutive quarters.

- **Needed units** — an estimate of how many new homes are required to accommodate population growth.

We calculated needed units using an average household size (AHS) of 2.5 persons:

$$\text{Needed Units}_t = \frac{4 \times (\text{Population}_t - \text{Population}_{t-1})}{2.5}$$

This converts quarterly population growth into an annualized measure of housing demand. To capture persistence and seasonality, we also included lag features such as the previous quarter's and the previous year's housing starts.

Model Setup

We evaluated two forecasting horizons:

1. **Next quarter (t+1)** forecasts, capturing short-term predictability.

2. **Same quarter next year (t+4)** forecasts, representing one-year-ahead seasonal patterns.

Two validation strategies were used: (1) a chronological holdout split, with training up to 2018 and testing on 2019–2025; and (2) a rolling evaluation, where models were retrained each quarter to simulate real-time forecasting.

Our baseline model was a **seasonal naïve forecast**, which assumes that housing starts in a given quarter will match the same quarter of the previous year. We compared this baseline with several regression and ensemble models using standard forecast metrics (MAE, RMSE, sMAPE, and MASE), with MASE serving as our key benchmark to assess performance relative to the baseline.

Models Tested

We trained and compared six models representing a mix of linear and nonlinear approaches:

- **Seasonal Naïve Model**, as a benchmark based on last year's same-quarter value(for t+4)/ previous quarter's value(for t+1).

- **Linear Regression (LR)**, to capture direct relationships between housing starts and demographic variables.

- **Ridge Regression**, to stabilize coefficients in the presence of correlated lag features.

- **Random Forest (RF)**, to capture nonlinear patterns and interactions.

- **Extra Trees Regressor (ETR)**, to enhance variance reduction and ensemble robustness.

- **XGBoost (XGB)**, to model nonlinearities efficiently through gradient boosting.

All applicable models were **hyperparameter-tuned** using randomized search with cross-validation. Each province was modeled independently to reflect regional differences in population growth, scale, and housing market volatility.

Findings

Across most provinces, the seasonal naïve model remained a strong benchmark. Linear and Ridge Regressions performed slightly better in smaller provinces such as Prince Edward Island and New Brunswick, while larger provinces like Ontario, Alberta, and British Columbia exhibited greater volatility and higher forecast errors.

Visual inspection showed that **next-quarter (t+1) forecasts** closely tracked observed housing starts but rarely outperformed the baseline in MASE terms, since short-term fluctuations are largely seasonal and persistent. By contrast, **same-quarter-next-year (t+4)** forecasts appeared less visually aligned but sometimes achieved MASE values below 1, indicating improved performance over the seasonal benchmark.

In several cases, forecasts **lagged behind observed turning points**, especially during rapid upswings or declines. This suggests that while models captured long-term trends and cyclical structure, they struggled to anticipate abrupt changes driven by economic shocks, policy shifts, or external constraints.

Adding population-based variables improved interpretability and long-run behavior, though the overall accuracy gains were modest. When we introduced more complex features (additional lags, rolling means, or quarter dummies), performance degraded due to overfitting and multicollinearity. The most stable configuration combined lagged housing starts with population change as key predictors.

Housing Adequacy Index (HAI)

To link housing supply and demographic demand, we constructed the **Housing Adequacy Index (HAI)**:

$$\text{HAI}_t = \frac{\text{Housing Starts}_t}{\text{Needed Units}_t}$$

An HAI value near 1 implies that new housing supply roughly matches population-driven demand, while values below 1 suggest increasing shortages.

We computed historical HAI series for all provinces but did not extend them into forecasts because model precision on housing starts remains close to the seasonal baseline. Once predictive accuracy improves, we plan to use forecasted housing starts together with Statistics Canada's population projections to estimate future housing adequacy scenarios.

Next Steps

To strengthen this framework, we plan to:

- Incorporate higher-frequency and macroeconomic indicators such as building permits, construction employment, mortgage rates, and inflation.

- Explore hierarchical or panel-based models to share statistical strength across provinces.

- Develop probabilistic forecasts with confidence intervals for housing adequacy projections.

We also intend to extend the HAI forecasting framework by combining population projections with predicted housing starts to evaluate supply adequacy under alternative demographic growth scenarios.

Conclusion

This work establishes a transparent and reproducible foundation for forecasting housing starts and evaluating housing adequacy across Canada. Although current models perform close to the seasonal baseline, they offer valuable insights into the relationship between population growth and housing supply. With additional economic and policy variables, this framework can evolve into a practical decision-support tool for anticipating Canada's housing supply challenges in the coming years.