Forecasting Canadian Housing Starts

Executive Summary

Erdos Institute Data Science Bootcamp Fall 2025

Anwesha Basu, Debanjan Sarkar

This project explores how quarterly housing starts across Canadian provinces can be forecasted from 1990 to 2025. Housing starts — the number of new residential construction projects that begin within a given period — are one of the clearest indicators of housing supply and overall economic activity. In recent years, Canada's rapid population growth and rising affordability challenges have drawn attention to the country's housing supply gap. By forecasting housing starts, we aim to build simple, interpretable models that help identify upcoming supply trends and better understand how demographic and economic factors shape 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 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 added lag features such as the previous quarter's and previous year's housing starts.

Model Setup

We evaluated two forecasting horizons:

1. **Next quarter (t+1)** forecasts, which test short-term predictability.
2. **Same quarter next year (t+4)** forecasts, which test one-year-ahead seasonal patterns.

We used two validation strategies: a chronological split (training up to 2018 and testing on 2019–2025), and a rolling setup where models were retrained each quarter to simulate real-time forecasting.

Our baseline model was a **seasonal naïve forecast**, assuming that housing starts in a given quarter are similar to the same quarter last year. We compared this baseline with several machine learning and regression models using metrics such as MAE, RMSE, sMAPE, and MASE, with MASE serving as the key benchmark for evaluating model performance relative to the baseline.

Models Tested

We trained and compared four main models:

- **Linear Regression**, to capture direct relationships between housing starts and demographic trends.
- **Ridge Regression**, to stabilize coefficients in the presence of correlated lag features.
- **Random Forest** and **XGBoost**, to model nonlinear effects and interactions.

Each province was modeled separately to reflect differences in scale, growth, and volatility.

Findings

Across most provinces, the seasonal naïve model remained a very strong benchmark. Simple models like Linear and Ridge Regression performed slightly better in smaller provinces (such as Prince Edward Island and New Brunswick), while larger provinces like Ontario, Alberta, and British Columbia showed greater volatility and higher errors.

Visual inspection showed that **next-quarter (t+1) forecasts** followed the data more closely, but their MASE values were often above 1 because the naïve baseline was already very accurate at short horizons. By contrast, **t+4 forecasts** looked less precise but sometimes achieved MASE values below 1, meaning they performed better than the seasonal baseline numerically.

When examining the time-series plots, we noticed that model forecasts often **lagged behind the true values**, especially during sudden upswings or drops in housing starts. This suggests that while the models captured general trends and seasonality, they struggled to anticipate rapid short-term fluctuations driven by economic or policy shocks.

Adding population-based variables helped interpretability and long-term behavior, though the overall accuracy gains were modest. When we tried adding more complex features — such as additional lags, rolling averages, or quarter indicators — model errors actually increased due to overfitting and multicollinearity. The simplest feature set, combining lagged housing starts with population trends, turned out to be the most stable and consistent approach.

Housing Adequacy Index (HAI)

To link housing supply and demand, we defined a **Housing Adequacy Index (HAI)**:

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

An HAI value near 1 indicates that new housing supply keeps pace with population-driven demand, while lower values suggest growing shortages. We calculated the HAI historically but did not extend it to future projections yet, as the forecasted housing starts were not sufficiently accurate. Once model performance improves, we plan to revisit this framework using Statistics Canada's population projections to evaluate future housing adequacy scenarios.

Next Steps

To strengthen the forecasting framework, we plan to:

- Incorporate higher-frequency indicators such as building permits, construction employment, and mortgage data.
- Explore panel and hierarchical models to share information across provinces.
- Build probabilistic forecasts that include confidence intervals around future housing starts.

We also aim to extend the HAI analysis once forecast accuracy improves, combining population projections with predicted housing starts to estimate how well supply may meet future demand.

Conclusion

This work establishes a clear and reproducible foundation for forecasting housing starts across Canada using demographic data. Although current models perform close to the seasonal baseline, they provide valuable insight into how population growth and housing supply interact. With additional economic and policy features, this framework can evolve into a powerful tool for understanding and anticipating Canada's housing supply challenges in the years ahead.