Canada Housing Prices Prediction

Lily Liu, Fay Yan

This report was written by us. We used ChatGPT or similar tools to revise grammar and sentence clarity only.

Abstract

This project aims to develop predictive machine learning models for housing prices in Canada using a property-level dataset sourced from Remax Canada. To address geographic features and market heterogeneity, we divided the data into four meaningful subsets: High Population, Resource-Rich, Smaller Population, and Luxury Market. We applied data preprocessing techniques, including encoding and scaling, to ensure consistent model training and a fair comparison, and did feature selection to reduce dimensionality and improve interpretability. We evaluated different models, including Linear Regression, Ridge Regression, Decision Tree, Random Forest, and Gradient Boosting, and developed a stacked meta regressor using the best parameters gained in the hyperparameter tuning process. Our results showed that the meta regressor generated the best performance with high R^2 and comparatively low RMSE. We found that data segmentation based on market conditions and including informative location features improves predictability.

Introduction

The Canadian housing market is a complicated and important topic for predictive modeling because of its significant geographical heterogeneity and price volatility. The affordability and value of housing have changed significantly between cities and provinces in recent years due to a combination of factors such as population growth, rising interest rates, and post-pandemic migratory patterns. Therefore, developers, investors, and policymakers who want to comprehend the dynamics of the local market as well as buyers and sellers can benefit from accurate and scalable methods for projecting home prices.

While previous studies have explored the use of machine learning for housing price prediction in Canada, many have relied on macroeconomic time series data or examined narrowly defined subsets of the market. In this project, we introduce a cross-sectional machine learning method to estimate individual house prices based on their identifiable features at the time of listing. Instead of anticipating future prices over a period, our emphasis

is on understanding how specific property-level attributes—like location, size, number of rooms, and amenities—affect prices in the present market. We utilize a dataset collected from Remax Canada and available on Kaggle (February 2025), which comprises thousands of listings from throughout the nation.

To enhance prediction accuracy and model robustness, we:

- Perform data preprocessing, including handling missing values, encoding categorical variables, and standardizing numerical features;
- Segment the dataset into four subsets– over 2.5 million CAD, high population/urbanized area with price less than 2.5 million CAD, rich resource/rural area with price less than 2.5 million CAD, small population/coastal area price less than 2.5 million CAD– to assess performance across different market conditions.
- Apply feature selection(except for Luxury Market) to reduce dimensionality and runtime from over 3,000 one-hot encoded features to 550 informative ones;
- Compare the performance of multiple regression models, including linear regression, random forests, and gradient boosting;
- Tune hyperparameters and develop a stacked meta-classifier to ensemble model predictions;

Background

Brännlund et al. (2023) at the Bank of Canada used monthly economic data from 1981 to 2019 to assess how well machine learning models could predict short-term changes in Canadian housing prices and existing home sales. They evaluated multilayer perceptrons (MLPs) and support vector regression (SVR), and compared their results with linear benchmarks. The authors stated that the improvement was restricted when using traditional macroeconomic time series inputs alone, even though the machine learning models occasionally showed somewhat higher predictive power. To properly use machine learning in real estate prediction, they suggested adding richer, more unique data (such as internet listings, unstructured, or high-frequency data) (Brännlund et al., 2023).

Another related study was conducted by Linares (2022), which examined fragmented housing markets across Canadian provinces in order to approach the issue from a regional viewpoint. Linares showed that ML models could accurately forecast regional home values using a collection of cointegrated macroeconomic data, including models such as XGBoost, LASSO

regression, and random forests. The study did, however, point out that overfitting was a problem because of the data's constraints and that adding more detailed information (such as specific property attributes) would probably enhance the model's performance.

In contrast to these previous studies, our project adopts a cross-sectional approach using recent, property-level data. The dataset we used includes detailed real estate listings from across Canada, with each row representing a specific house along with its physical and locational features (e.g., square footage, number of bedrooms, heating type, city, and province). Instead of simulating market-wide temporal patterns, our objective is to forecast sale prices based just on the characteristics of a certain property at a particular moment in time.

Methods

Since our target variable, housing prices, is a numerical value, we consider the following machine learning models:

- **Linear Regression**: Baseline model assuming a linear relationship. It minimizes mean square error(MSE) and can serve as a benchmark, especially when our target variable is numerical.
- **Ridge Regression**: Linear regression with L2 regularization, which helps to prevent overfitting when the feature space has high dimension.
- **Random Forest Regressor**: An ensemble method using bagged decision trees. Each tree is trained on a bootstrapped subset of the data, and final predictions are averaged so that it’s robust against overfitting.
- **Gradient Boosting Regressor**: An ensemble method that builds trees sequentially, where each new tree focuses on correcting the residuals (errors) of the previous ones. It is effective at capturing complex patterns and offers flexibility for hyperparameter tuning.
- **Stacked Meta-Regressor**: We combined the above models using a stacking strategy. The predictions from the base models (Linear, Ridge, Random Forest, and Gradient Boosting) are input to a final Linear Regression meta-model, which learns to optimally combine them. This approach can often outperform any single base model by leveraging their complementary strengths.

Experiment/Results

Data description

The Canada housing data we used came from Kaggle (Remax Canada, Feb 2025), which contains more than 44000 observations and 22 features including location, size, property type, number of bedrooms and bathrooms, and additional features like heating, flooring, garage availability, etc..

Data preprocessing

Our initial preprocessing includes:

- Dropped columns with more than 60% missing values.
- Imputed missing values
    - **Categorical**: mode imputation.
    - **Numerical**: mean imputation
- One-hot encoded all categorical features (e.g., city, province, property type)
- Standardized numerical features (zero mean, unit variance).

After examining our data, we discovered our target variable, housing price, is very right-skewed, and the log transformation didn't do much when we ran the models. We wanted to use all our data but prevent the right skewness from influencing the overall prediction, so we decided to split our data using 2.5M as the threshold.

More importantly, we needed to find a way to correctly incorporate the City and Province feature. If we treat each city and province as one feature, we would have over 3000 features after one-hot coding. If we separate each province into its own subset, the dataset will be too small to train a good model. Therefore, we decided to split the properties with prices under 2.5M further into three subsets based on their economic and housing market conditions. For the high dimensionality resulting from the City feature for the three under 2.5M subsets, we feature-selected the overall top 550 features to keep important City information. This step significantly improved training time and helped maintain model interpretability without degrading predictive performance. By the end, we have four subsets in total:

1. **High Population Area:** ON, QC, BC
2. **Resource-Rich Area:** AB, SK, MB

3. **Smaller Population Area:** all other provinces
4. **Luxury Market:** All properties with prices over 2.5 million CAD.

Each subset was used for separate model training and evaluation to assess performance across different housing markets.

Model comparison

	High Population		Resource Rich		Smaller Population		Luxury Market	
	RMSE	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**
Linear	291791	0.71	169296	0.73	226160	0.45	2428183	0.18
Ridge	292375	0.70	169024	0.73	221199	0.48	2418672	0.19
Random Forest	252523	0.78	138270	0.82	207835	0.54	1894134	0.50
Gradient Boosting	287610	0.71	173444	0.72	209562	0.53	1843109	0.53
Decision Tree	329819	0.62	172699	0.72	259365	0.28	2426052	0.18

Linear Regression and Ridge Regression had good performance in High Population and Resource Rich subsets (R^2 ~0.70–0.73). This suggests that in these two datasets, some features may change proportionally to price, making simpler models relatively effective in regions with more predictable housing trends. However, their performance dropped in the Smaller Population and especially in the Luxury Market (R^2 ~ 0.18–0.48). This is likely due to the nonlinear relationships or complex interactions. For example, in the Luxury Market, price might have an exponential or other non-linear relationship with our features.

The Decision Tree model consistently showed weaker performance across all subsets. This indicates that the model failed to capture much of the variability in housing prices. Given the potential of overfitting and the underperformance of a single decision tree model, we decided to not continue with this model and shifted focus to ensemble methods. Both Random Forest and Gradient Boosting showed the best overall performance. Random Forest achieved the best overall R^2 in the Resource Rich subset ($R^2 = 0.82$) and strong results in High Population ($R^2 = 0.78$). Gradient Boosting had the strongest performance in those three subsets under 2.5 million CAD. Both Random Forest and Gradient Boosting had good performance in the

Luxury Market subset with Gradient Boosting slightly outperforming Random Forest(R^2 = 0.53 vs. 0.50).

Across four subsets, model performance varied and reflected our subsampling choices. The differences in data size and price variability directly affected model performance. The High Population and the Resource Rich subsets were relatively large and stable, which enables models to achieve higher prediction accuracy. In contrast, the Smaller Population and Luxury Market subsets were smaller and had higher variance, resulting in a less than ideal performance.

Hyperparameter tuning

To further improve model performance, we conducted hyperparameter tuning using grid search with cross-validation. This allowed us to further select the model and identify the best parameter combinations for each regional subset.

Subset: High Population

Model	RMSE	R^2	Best Hyperparameters
Linear	291791	0.71	/
Ridge	292110	0.71	alpha': 0.7, 'solver': 'svd'
Random Forest	252729	0.78	'max_depth': None, 'n_estimators': 500
Gradient Boosting	248043	0.79	'learning_rate': 0.1, 'max_depth': 5, 'min_samples_split': 5, 'n_estimators': 500

For the High Population subset, Ridge Regression and Random Forest did not benefit much from hyperparameter tuning. However, Gradient Boosting improved from R^2 = 0.71 to 0.79 after tuning. Thus, among all models, Gradient Boosting with tuned parameters achieved the best performance for this subset.

Subset: Resource Rich

Model	RMSE	R^2	Best Hyperparameters
Linear	169295	0.73	/
Ridge	168781	0.73	'alpha': 0.5, 'solver': 'svd'
Random Forest	139136	0.82	'max_depth': None, 'n_estimators': 500

Gradient Boosting	154187	0.78	'learning_rate': 0.1, 'max_depth': 5, 'min_samples_split': 2, 'n_estimators': 500

Similar to the High Population subset, hyperparameter tuning had little effect on the performance of Ridge Regression and Random Forest for the Resource Rich dataset. Gradient Boosting improved slightly after tuning, but still underperforms Random Forest. Random Forest with default-tuned settings and after tuning shows a similar $R^2 = 0.82$, but before tuning, it has a lower RMSE. Thus, Random Forest with default-tuned settings remains the most effective model for this subset.

Subset: Smaller Population

Model	RMSE	R^2	Best Hyperparameters
Linear	226160	0.45	/
Ridge	221458	0.48	'alpha': 0.8, 'solver': 'auto'
Random Forest	211552	0.52	max_depth': 20, 'n_estimators': 500
Gradient Boosting	204676	0.55	'learning_rate': 0.1, 'max_depth': 5, 'min_samples_split': 5, 'n_estimators': 500

For the Smaller Population subset, Ridge Regression maintained similar performance after tuning. Random Forest showed a slight decline in performance (R^2 dropped from 0.54 to 0.52), possibly due to overfitting from a deeper tree configuration. Also, the Smaller Population subset has relatively fewer samples, so small changes in model configuration may yield unstable results. Gradient Boosting benefited from hyperparameter tuning and achieved a better performance. Thus, Gradient Boosting with tuned parameters achieved the best performance for this subset.

Subset: Luxury Market

Model	RMSE	R^2	Best Hyperparameters
Linear	2428183	0.18	/
Ridge	2383085	0.21	'alpha': 1.0, 'solver': 'auto'
Random Forest	1893989	0.50	'max_depth': 10, 'n_estimators': 100
Gradient Boosting	1843109	0.53	'learning_rate': 0.1, 'max_depth': 3, 'min_samples_split': 2, 'n_estimators': 100

In the Luxury Market subset, only Ridge Regression benefited from the hyperparameter tuning, while Random Forest and Gradient Boosting had similar performance compared to the default setting. This is probably due to the high variability in this subset, making it difficult for fine-tuned parameters to generalize better than already robust default configurations. However, among all models, Gradient Boosting still achieved the best overall performance for this subset.

Meta Regressor

We created a meta regressor by stacking Linear regression, Ridge regression, Random Forest regressor, Gradient Boosting regressor with a linear regression meta classifier.

Subset	RMSE	R^2
High Population/Urbanized	239926	0.801
Rich Resources/Rural	134998	0.830
Small Population/Coastal	203963	0.556
Luxury Market	1834244	0.53

Meta Regressor led to improved or matching performance across most subsets. For the High Population and Rich Resources subsets, the meta regressor improved model performance (R^2 improved from 0.79 to 0.801 and from 0.82 to 0.83). This shows that blending models helped capture both linear and nonlinear patterns in these two subsets. For the Smaller Population subset, the meta regressor only improved the performance marginally, which is likely due to fewer samples and higher noise in this subset. For the Luxury Market subset, performance remained similar to the best Gradient Boosting model ($R^2 = 0.53$), suggesting that stacking did not yield further gains, possibly due to extreme variability and limited sample size.

Discussion

Overall Performance & Segmentation Impact

Our results showed that overall, machine learning models are good for housing price predictions across Canada when trained on structured, property-level data. Across most subsets, our models had strong performance in terms of RMSE and R^2. Although our RMSE values are in the order of 10^6, they are consistent with the scale reported in previous studies

and are comparatively lower in magnitude, even before using the meta regressor. Furthermore, the R^2 in the High Population and Resource Rich subset is particularly strong, reaching 0.801 and 0.830, respectively. We attribute this improved performance to the data segmentation based on geographic and market characteristics. By dividing the data into four subsets, we were able to train our model on more structured data so that there was less noise and enhanced predictability.

Model Comparison

In this study, we explored and compared various models to understand how different approaches perform under diverse market conditions. Ridge Regression and Linear Regression performed well in larger and more stable subsets like High Population and Resource Rich, but struggled in subsets with limited data dn high variance, such as the Luxury Market and Smaller Population. Ensemble tree-based models, Random Forest and Gradient Boosting, consistently outperformed across most subsets than linear models. Their ability to capture nonlinear relationships and interactions allowed them to have better generalizability.

Value of Subset-Specific Modeling

One of the key lessons from this study is the value of subset-specific modeling. By our subsampling choice, we were able to isolate different price dynamics and better fit our models. Furthermore, the Stacked Meta-Regressor further improved performance by combining the strengths of multiple models, showing that the ensemble learning model captures key patterns while reducing the weaknesses of each individual model. This reinforces the value of subsampling modeling and suggests that tailoring strategies to data characteristics is essential for building accurate and generalizable price prediction in real estate.

Future Work

Although we had strong performance in several subsets, results varied across regions. The performance in R^2 for the Small Population ($R^2 = 0.556$) was modest due to the relatively small data size. Further work could train the model using a larger dataset. Additionally, the performance for the Luxury Market was the weakest overall (RMSE = 1834244, $R^2 = 0.53$). This is likely due to both the limited availability of high-end housing data and the volatility in

luxury housing prices. Since little work has been done on this part of the market, future work could explore other predictive features and more robust modeling strategies, such as incorporating external economic indicators or handling outlier-driven distributions.

Contibution

Both: data preprocessing, feature selection, model training, and evaluation for codes

Fay: hyperparameter tuning for codes; Github repository, and the second half of the report

Lily: meta regressor for codes; first half of the report

Code

Github repository: https://github.com/fayyyyyxy/Canada-Housing-Prices-Prediction.git

Reference

- **Brännlund, J., Lao, H., MacIsaac, M., & Yang, J.** (2023). *Predicting Changes in Canadian Housing Markets with Machine Learning.* Bank of Canada Staff Discussion Paper 2023-21. Available at: https://www.bankofcanada.ca/2023/09/staff-discussion-paper-2023-21/
- **Linares, M.** (2022). *Predicting Canadian House Prices Using Machine Learning and Macroeconomic Data.*Master's thesis, University of Waterloo. Available at: https://uwspace.uwaterloo.ca/handle/10012/18809
- **Yuliia Bulana.** (2025). *Canada Housing* [Data set]. Kaggle. Retrieved from: https://www.kaggle.com/datasets/yuliiabulana/canada-housing
- **Koch, A., Peremyslova, M., & Lemanowicz, L.** (2018). *Zestimate Bazinga: Predicting the Selling Price for Condos in Downtown Vancouver.* Stanford CS229 Project Report. Available at: https://cs229.stanford.edu/proj2018/report/86.pdf