Norwegian University
of Life Sciences

Master's Thesis 2022 30 ECTS
Norwegian University of Life Sciences
School of Economics and Business

Canadian Housing Prices – A Case Study Using Macroeconomic Variables and Machine Learning

Anita Sarah Linares
Applied Economics and Sustainability

Acknowledgements

Submission of this thesis marks the end of my two-year Master's Program at NMBU. The challenges I faced throughout the program allowed me to rise to the occasion and broaden my horizons. While undertaking the thesis alone was challenging, it pushed me to develop both my programming skills in Python and theory knowledge in economics.

I am grateful for the opportunity to work with my supervisor Dag Einar Sommervoll. Under his tutelage, I made significant improvements and gained valuable knowledge in the methods used in this thesis. I am also grateful to all the professors I have met throughout my program; the wisdom they parted to me motivated me to pursue the topics discussed in this thesis.

I would also like to thank Cheng Zu (Colin) Cui for always taking the time to edit my paper.

Lastly, I'd like to thank my parents for their continuous support and encouragement for me to pursue my dreams throughout my entire academic career.

Ås, 2022 May 16.

Anita Sarah Linares

Abstract

This thesis uses cointegration methods to investigate the extent to which segmentation manifests in the Canadian housing markets. By applying the Johansen, Engel-Granger, and FMOLS models, inference on market segmentation and long-run relationships of the associated macroeconomic variables are evaluated. The initial result from the analysis brings to light the issue of using the national housing price index to predict local housing prices and the necessity to study local macroeconomics variables that can influence housing prices. Furthermore, XGBoost, LASSO, and Random Forest will be utilised to predict housing prices using the aforementioned cointegrated variables. Using machine learning methods showed that the variables chosen can reasonably predict local housing prices but ultimately also displayed the limitations of the data set, highlighting the need for more data and features to reduce overfitting.

Contents

List of Figures

List of Tables

1 Introduction

Home ownership, marriage, and raising a family are essential life milestones for many Canadians. However, during the past couple of decades, home ownership has become one of the least obtainable milestones for many. One of the reasons home ownership is slipping away from Canadians is the mismatch between rising prices and stagnant wages. It leaves many Canadians wondering if home ownership will ever be within grasp. In a quick span of 6 years, the housing price to income ratio has increased by more than 40% (Statista, 2022). Understanding current trends and creating credible forecasts for the future housing market will give consumers foresight into how they may achieve home ownership. Despite Canada being a large and dynamic economy with vastly different labour markets and geographical environments, the current housing price index is a national index. Using a single index to represent the Canadian housing market may not be entirely accurate. Therefore, testing whether Canada has a segregated housing market is a critical analysis that should be undertaken. Following then is an analysis of the effectiveness of machine learning methods in out-of-sample predictions in the housing market. This study will attempt to answer two critical questions related to the Canadian housing markets:

RQ1: *Is there housing market segregation in Canada? If so, can a single housing price index be used to predict housing prices in Canada?*

RQ2: *Can local macroeconomic variables be used as proxies to predict regional housing prices accurately*?

To answer these questions, we will begin by evaluating pricing trends in the Canadian housing market over the past few decades. This is followed by a literature review of influential macroeconomic variables in the market and machine learning applications in economic theory. Then, an empirical evaluation of the Canadian housing index is also conducted to verify if a national index can be used to gauge local pricing movements. Furthermore, this study will evaluate the ability of regional variables to predict prices. As discussed by Abraham and Hendershott, 1994, variables within localized regions may account for the heterogeneity

within the real estate markets. Cointegration tests with housing prices and a vector of selected macroeconomic variables will be evaluated to establish the relationship between prices and chosen variables. If cointegration exists, it is reasonable to conclude that the variables can be proxies for local housing prices. After selecting and verifying the variables, predictive modelling on pricing movements is examined. This is done using machine learning algorithms including Random Forest, LASSO, and XGBoost. RMSE, MAE, with R^2 used as the evaluation metric for each method. Finally, an analysis of housing prices and the out-of-sample forecast will be evaluated with the best-performing machine learning methods.

1.1 Background

Many believe that housing price movements are correlated to business cycles within the market; however, the Canadian housing market has proven to be irregular. An example of this in recent history is the global financial crisis of 2008/2009. Unlike the American housing market which faced rapid growth followed by a sharp decline in prices, the Canadian market did not. Monetary policies and market conditions were comparable during the same time, which suggests that other factors influenced the outcomes in the Canadian market (MacGee, 2009).

In recent years, the idea that foreign investment is one of the key drivers for the soaring housing prices, especially in metropolises like Vancouver in Toronto, has been gaining popularity. Governments have imposed a 'foreign buyer's tax' in response to public outcry. In 2016, British Columbia's premier imposed taxes on foreign real estate investments at 15% of the value. Soon after, the Ontario government followed suit by applying the same 15% tax on foreign property investments. The immediate results of the new tax law led to a noticeable reduction in property sales year over year in the two largest Canadian cities - Vancouver and Toronto (Allan, 2019). However, in the years that followed, an evaluation showed that the reduction in sales had minimal impact on the prices (Allan, 2019). This finding disputes the notion that foreigners purchasing properties is the primary driver of housing prices in Canada. In fact, Bunce et al., 2020 note that it is not even a secondary price driver. What does show to be a primary driver of housing prices, especially in GTA, is the zoning laws associated with urban development (Bunce et al., 2020).

Unexpected housing price movements are also seen during the onset of the Covid-19 pandemic. The pandemic drastically slowed economic activity and altered consumer behaviours. Canada, like many other countries in the world, implemented lockdown policies. Many jobs were lost as a result of the lockdown, which brought to light the impending issue that many Canadians may default on their mortgages due to lost income. The Canadian government announced that mortgage deferrals would be an option to mitigate the problem. In addition to the deferrals, many Canadians applied for government income support. With the

two-front relief from the government, consumers did not suffer a housing crisis. Yet, prices continued to trend on an upward trajectory. While government policies may signal issues in the housing market, financial reports say otherwise. In a report published by the Bank of Canada (BOC), they noted a year-over-year increase of over 17% in housing prices during the pandemic (Khan et al., 2021), with the Greater Toronto and Ottawa areas experiencing the most price growth during the pandemic. Furthermore, the report mentioned that housing prices were driven by second time home buyers who took advantage of the low mortgage rates (Khan et al., 2021).

In complete contrast to the housing market during the same period, gasoline prices, another common indicator of economic strength in western Canada, dipped to 77 cents/litre in the province of Alberta. The lowest since 2016. Although various factors determine gasoline prices, this drastic price drop was mainly caused by consumer lifestyle changes brought upon by the pandemic. Despite most other industries suffering setbacks, housing prices continued to rise throughout Canada during the same time (Khan et al., 2021). The contrast between these two markets further highlights the irregularities that exist in the Canadian housing market.

Ranking second by the largest landmass globally, it is important to consider regional heterogeneity when evaluating housing prices. This study investigates two aspects that the current forecasting methods fail to address. First, to what extent is the Canadian housing market segmented? Secondly, should regional heterogeneity exist in the market, can an aggregate housing price index accurately predict housing prices throughout Canada? A time-series analysis of panel data is performed to answer these questions. This paper will evaluate market dynamics by using tests for cointegration. Initial evidence through academic literature in the Canadian market suggests that there are idiosyncratic conditions that may indicate that price movements are determined within a localized region (Allen et al., 2009). To investigate, this study will begin by conducting Augmented Dickey-Fuller and cointegration tests on non-stationary time series. Evaluating these tests will validate if there is indeed a single market.

2 Literature Review

This section begins by evaluating the literature on macroeconomic variables and their importance for housing price predictions, followed by a discussion on machine learning in economics. Reviewing the literature on macroeconomic variables in the housing market and machine learning applicability provides the foundation for applying these methods in this thesis.

2.1 Variable selection

In a paper written by Demers et al., 2005, they highlight that the price-to-rent ratio, age, and wealth are important variables for influencing the quantity demanded in housing. Price-to-rent ratios allowed consumers to decide when it would be more 'reasonable' to become owners or continue to be tenants. As for age, ownership of a home would be doubtful for ownership under 18 years of age. Therefore, it would be best to omit this age group when using an age variable. This is proven by disaggregating the data, as noted in Demers et al., 2005, which showed that the age group between 15+ or, 25-44, would be best suited for forecasting housing markets as it would show the preferences of multi-income households, women entering the market, and urbanization. As for wealth, Demers et al., 2005, deemed that this variable is an important to consider based on the permanent income hypothesis, where the consumption of this period is proportional to the expected income of a household.

Volatility within the housing market is driven in part by changes in consumer consumption, as written by Piazzesi and Schneider, 2016. The marginal propensity to consume and exogenous shocks to the housing market were some variables to evaluate. Piazzesi and Schneider, 2016 mention that some notable exogenous shocks to the market are wealth, changes in expectations, and prices relative to housing services. Their analysis notes that the changes in wealth held an overall effect within the market. The issue of simply looking at wealth is that once transaction costs are incorporated into the evaluation of a home, the importance of how wealth affects the market no longer holds.

Lastly, in a recent book published by Bunce et al., 2020, another key driver for housing prices, especially within the GTA region, has been the income-to-debt ratio. In their discussion, they remarked how the Canadian government has relied on market players to be the ones to stimulate the economy and not the government. Essentially, the government would enable a policy that allows cheap interest rates for consumer investments. This would allow individuals to borrow at a lower cost but increase the income-to-debt ratio. This approach to stimulate the economy has been a key driver in the housing markets and is one indicator that the authors discuss – that when interest rates are low, it will increase the marginal propensity to consume, which is a key driver in the housing market.

An ample amount of economic literature suggests that relevant macroeconomic variables may infer how housing prices are influenced. The macroeconomic variables suggested in the mentioned literature will be used in this thesis to evaluate cointegration and out-of-sample predictions.

2.2 Machine Learning and Econometrics

In its essence, statistical machine learning (ML) utilizes computation to train prediction models. It is widely used to derive business insights such as performance trends to optimize operations. In other instances, ML can be used to optimize public transport, sewage, and water facilities further (Visvizi et al., 2018). Other cases not widely discussed are ML and econometric practices - where ML does not necessarily have a strong focus on the inference of our variables, while in econometrics, there is (Mullainathan and Spiess, 2017; Varian, 2014). When applying econometric theory to the data set, we are evaluating our parameters and their relationship with the dependant variable.

Mullainathan and Spiess (2017), further discusses the synergy between econometrics and ML. Literature on ML suggests that variable selection is based on market intuition and trends (Mullainathan and Spiess, 2017; Varian, 2014). However, this method may cause overfitting issues due to using variables that do not have a concrete mathematical relationship to the

target data. Econometric theory bridges this gap by providing the means to regularize estimators or data-driven regularization parameters (Abadie and Kasy, 2018). Econometric theories can also be applied to ML to make inferences. As mentioned in Varian, 2014, such inferences have the potential to drive policy decisions. Even though it is easy to infer causal effects by examining the ML algorithm's strong predictive abilities, it would be better to evaluate the impact of the features in the model.

In this study, by combining both econometric principles and ML, we can conclude the causal effects our chosen variables have on housing prices (Varian, 2014). The topic is further explored with macroeconomic variables to evaluate machine learning algorithms and their limitations.

3 Market Segmentation

The overarching goal of this section is to answer if and to what extent there is housing market segregation in Canada. The first step in answering this question is to gather historical price data from the main cities in British Columbia, Alberta, Ontario, Quebec, and Newfoundland and Labrador. Firstly, testing if the series is stationary or not is done by running an ADF test. Then, a cointegration test is used to evaluate if the markets move in tandem. Running these tests provided answers to the first research question.

3.1 Data

All data is collected from Statistics Canada and MLS CREA (Multiple Listing Service Canadian Real Estate Association). The time horizon of our data set will be monthly data from January 2005 till January 2022[1]. The CREA data set consists of 205 periods or 205 months. The MLS' Housing Price Index (HPI) is calculated by taking the aggregate collection of 18 Canadian housing markets: Vancouver Island, Victoria, Greater Vancouver, Fraser Valley, Okanagan Valley, Calgary, Edmonton, Regina, Saskatoon, Guelph, Hamilton- Burlington, Oakville-Milton, Barrie and District, Greater Toronto, Niagara Region, Ottawa, Greater Montreal, and Greater Moncton. The data used to compare with HPI index will be local for Victoria, Greater Vancouver Area, Calgary, Edmonton, Ottawa, Greater Toronto Area, Quebec City, Montreal, and St. John's, Halifax. The location of these cities span across the entire country, and may be referenced in Figure 3.1

[1]Data was collected on 02.01.22

Figure 3.1: Canada's Political Divisions

Source: From Her Majesty the Queen in Right of Canada, Natural Resources Canada. (2006). [Image]. *Canada Political Division - English.*

Table 3.1.1: Summary Statistics, x100,000

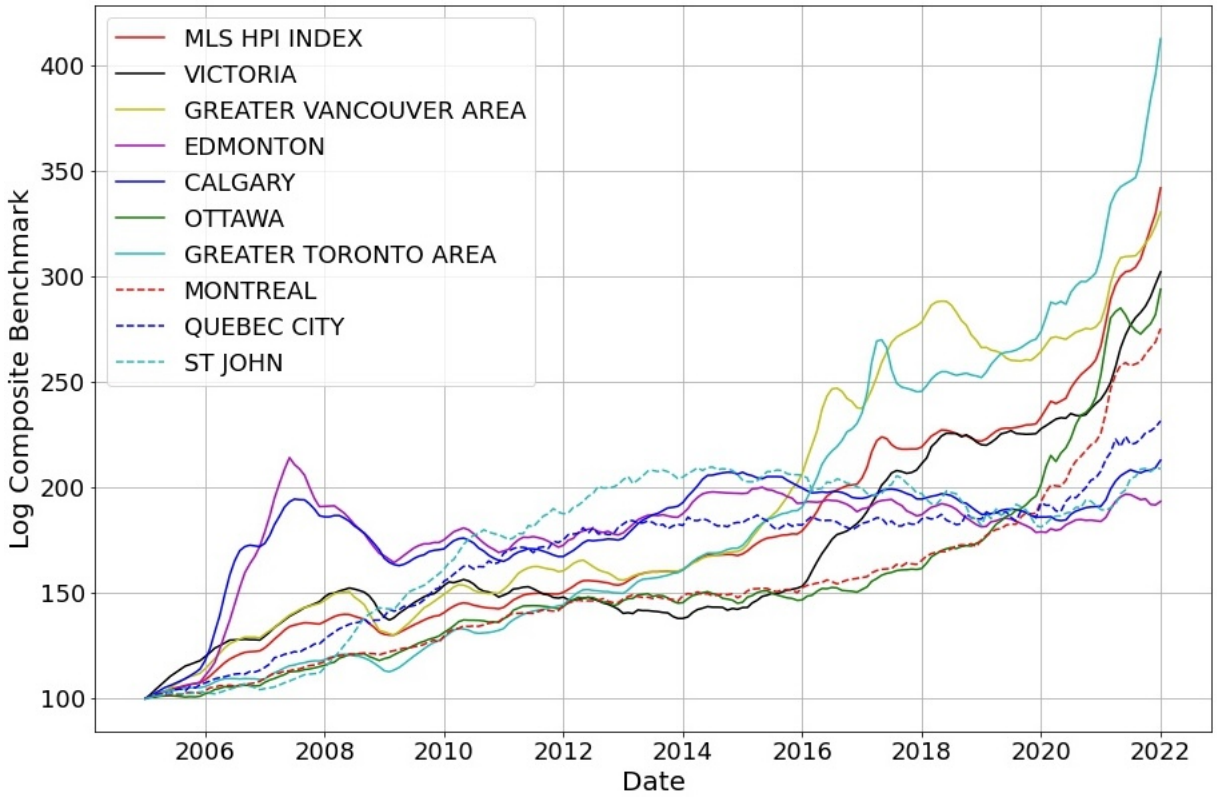
	CAD	VIC	GVA	CAL	EDM	OTT	GTA	MON	QUE	ST.J
count	205	205	205	205	205	205	205	205	205	205
mean	4.3	5.2	7.3	3.9	3.1	3.6	5.6	2.9	2.2	2.4
std	1.3	1.4	2.4	0.52	0.42	1.2	2.3	0.74	0.42	0.52
min	2.4	3.1	3.8	2.2	1.8	2.3	3.1	1.9	1.3	1.4
max	8.3	9.2	12.55	4.6	3.8	6.9	12.6	5.3	3.1	3.9

Note: Variables CAD, VIC, GVA, CAL, EDM, OTT, GTA, MON, QUE, ST.J, are defined as the followed: Canada, Victoria, Greater Vancouver Area, Calgary, Edmonton, Ottawa, Greater Toronto Area, Montreal, Quebec City, and St. John's.

A brief overview of the data shows that all housing markets have increased in housing prices

since 2005. This is shown in Figure 3.2 below. Slower growth compared to other cities is found within the Atlantic region and some western cities. Initial evaluation of the graph in Figure 3.2 alludes to potential market segmentation with room for improvement using further estimations.

Figure 3.2: MLS CREA HPI, Rebased 100



Another interesting point to consider in Figure 3.2 is the sharp increase in the HPI during 2006-2007 in the province of Alberta. Edmonton and Calgary are the two largest cities there, with their local economy being driven by oil and gas prices. The rapid expansion and contraction of the housing market are correlated with the oil commodities during the same period. As a result, the Alberta market had the most notable decrease in housing prices compared to the rest of the country. This trend also contributes to the narrative that there is possible market segmentation within Canada, as market trends do not affect the country equally.

3.2 Cointegration

Cointegration in time series analysis is an effective way to answer the question *Is there housing market segmentation in Canada?* When conducting economic time series analysis, the goal is to make a causal inference on our variables. Cointegration methods allow us to investigate the relationship between variables by determining if there is a long-run relationship between them. Should two variables be cointegrated, it is then safe to conclude that those variables do not deviate from the equilibrium in the long run.

In this study, two time series that are examined: (x_t, y_t) the national housing price index and local housing price indices. By evaluating if x_t and y_t are cointegrated or not, we can determine if the housing markets are segregated from the national price index. Should the variables not be cointegrated in the long run, it would motivate us to use other idiosyncratic variables to analyze the market.

3.2.1 Definition

Cointegration is defined as two time series of x_t and y_t, which are integrated in the first order $I(1)$, suggesting a unit root present, and there is a parameter that is a stationary process. In other words, if x_t,y_t are both $I(1)$, there exists a linear combination such as Equation 3.2.1 below,

$$z_t = m + ax_t + by_t \tag{3.2.1}$$

that is both $I(0)$ and has a zero mean, then x_t, y_t is said to be cointegrated of the order b, d where $d \geq b \geq 0$ (Asteriou and Hall, 2015). Given the series is integrated of order d and there is a linear combination of these variables, the linear combination of the variables would then be expressed as $a_1x_t + a_2y_t$, and that they are integrated in the order of $d - b$. The parameters, a_1, a_2 are defined as the cointegrating vector, and the relationship is expressed as: $x_t, y_t \sim CI(d, b)$.

3.2.2 Error Correction Model and Vector Error Correction Model (ECM and VECM)

To better understand ECM, consider a time series model based on the assumption that two series are cointegrated in the long run, where they have two time series, x_t and y_t, and they have the equation as follows:

$$y_t = \beta_1 + \beta_2 x_t + u_t \tag{3.2.2}$$

taking the residual, we will have:

$$\hat{u}_t = y_t - \hat{\beta}_1 - \hat{\beta}_2 x_2 \tag{3.2.3}$$

which suggests that y_t and x_t are cointegrated given that $\hat{u}_t \sim I(0)$. We can then represent the ECM model to be as:

$$\Delta y_t = a_0 + b_1 \Delta x_t - \pi \hat{u}_t - 1 + e_t \tag{3.2.4}$$

For Equation 3.2.4, it has solved the issues with spurious regression, given that everything is stationary. Equation 3.2.3 and x_t,y_t are also stationary because they are assumed to be $I(1)$ by the assumption of cointegration.

In other scenarios where multiple cointegrated equations are present, applying the VECM theorem would be more appropriate. The VECM model is a multivariate equation and is an extension of the ECM model (Asteriou and Hall, 2015).

To represent VECM, a multivariate equation is used instead of a bivariate equation. The variables in the equation are x_t, y_t and w_t, which are endogenous. To represent the model, we will use the equation expressed by Asteriou and Hall (2015) below,

$$z_t = a_1 z_{t-1} + a_2 z_{t-2} + ... + a_k z t_k + u_t \tag{3.2.5}$$

where z_t can be expressed in matrix notation as $z_t = [y_t, x_t, w_t]$. From Equation 3.2.5, we can deduce that the VECM can be written as the following,

$$\Delta z_t = \Gamma_1 \Delta z_{t-1} + \Gamma_2 \Delta z_{t-2} + ... + \Gamma_{k-1} \Delta z_{t-k-1} + \Pi z_{t-1} + u_t \tag{3.2.6}$$

Equation 3.2.6 now allows us to see the relationship with our variables from Π. Π represents the matrix of our variables represented by z_t. The VECM is commonly used with the Johansen test, where we use the maximum likelihood ratio to determine the rank of Π.

3.2.3 Determining order of integration

Dickey-Fuller Test (DF) examines the presence of a unit root in order to determine the order of integration. When applying a simple DF test, let us consider an AR(1) process, where we are testing to see if y_t is a stationary time series or not (if there is a unit root or not). The null hypothesis is $H_0 : \phi = 1$ and the alternative is $H_1 : \phi < 1$.

$$y_t = \phi y_{t-1} + u_t \tag{3.2.7}$$

An alternative expression of the DF test can be seen by the following,

$$\begin{aligned} y_t - y_{t-1} &= (\phi - 1)y_{t-1} + u_t \\ \Delta y_t &= (\phi - 1)y_{t-1} + u_t \\ \Delta y_t &= \beta y_{t-1} + u_t \end{aligned}$$

Where our hypothesis testing is now $H_0 : \beta = 0$ and $H_1 : \beta < 0$. Should we test y_t and find that our DF test shows that $\beta = 0$, then the time series is said to be non-stationary or a random walk.

An extension of this test, and what will be mostly used throughout this study, is an Augmented Dickey-Fuller test (ADF). The ADF test allows for additional lagged terms of the dependent variable to eliminate auto-correlation. To determine how many lagged variables should be included, Bayesian Information Criteria (BIC) and Akaike information criterion (AIC) are applied.

The three forms of this test are as the following:

$$\Delta y_t = \beta y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_t \tag{3.2.8}$$

$$\Delta y_t = \alpha_0 + \beta y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_t \tag{3.2.9}$$

$$\Delta y_t = \alpha_0 + \beta y_{t-1} + \alpha_2 t + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_t \tag{3.2.10}$$

All three equations share the same critical values except variables α_0 and $\alpha_2 t$. These variables represent the deterministic trends between each expression. Additionally, Asteriou and Hall, 2015 notes that the most common approach to start the estimation is Equation 3.2.10, which represents constant and trend. Equation 3.2.10 will be applied for all the following ADF tests unless otherwise stated.

3.2.4 Johansen's Test

One of the two cointegration methods used in this thesis to examine market segmentation in the Canadian market is Johansen's test. The advantage of using Johansen's test in comparison to Engel-Granger's test is, for example, the ability to gain estimates from two cointegrating vectors. The simplification is shown in Equation 3.2.14 below, which shows the long-term relationship in linear form.

Another advantage of using Johansen's test is the ability to determine differing speeds of the coefficients ($a_{11}a_{21}a_{31}$). Furthermore, the single and multi-equation method are considered to be the same when $a_{11} = a_{31} = 0$. When this condition is held, we can also assume that x_t and w_t are weakly exogenous (Asteriou and Hall, 2015).

Before we begin to show the long-run linear relationship, refer back to Equation 3.2.6. It should be noted that the long-run relationship between variables is shown with variable Π. The decomposition of this variable can be shown as $\Pi = \alpha\beta'$, where α represents the speed of adjustments to equilibrium coefficients and β' is the matrix of the long-run coefficients.

The variable $\beta' Z_{t-1}$ in Equation 3.2.6 is equivalent to the error correction term and contains $n - 1$ vectors in a multivariate framework (Asteriou and Hall, 2015). To mathematically represent this, consider the following case where we have two lagged terms. We can then

represent Equation 3.2.6 as

$$\begin{pmatrix} \Delta y_t \\ \Delta x_t \\ \Delta w_t \end{pmatrix} = \Gamma_1 \begin{pmatrix} \Delta y_{t-1} \\ \Delta x_{t-1} \\ \Delta w_{t-1} \end{pmatrix} + \Pi \begin{pmatrix} y_{t-1} \\ x_{t-1} \\ w_{t-1} \end{pmatrix} + e_t \tag{3.2.11}$$

can also be represented as,

$$\begin{pmatrix} \Delta y_t \\ \Delta x_t \\ \Delta w_t \end{pmatrix} = \Gamma_1 \begin{pmatrix} \Delta y_{t-1} \\ \Delta x_{t-1} \\ \Delta w_{t-1} \end{pmatrix} + \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \begin{pmatrix} \beta_{11} & \beta_{21} & \beta_{31} \\ \beta_{12} & \beta_{22} & \beta_{32} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \\ w_{t-1} \end{pmatrix} + e_t \tag{3.2.12}$$

the error correction variable on the LHS of the equation ($\Pi_1 z_{t-1}$) can be represented as,

$$\Pi z_{t-1} = ([a_{11}\beta_{11} + a_{12}\beta_{12}][a_{11}\beta_{21} + a_{12} + \beta_{22}][a_{11}\beta_{31} + a_{12}\beta_{32}]) \begin{pmatrix} y_{t-1} \\ x_{t-1} \\ w_{t-1} \end{pmatrix} \tag{3.2.13}$$

then rewriting the first row shown in Equation 3.2.13 reveals in Equation 3.2.14 the relationship between the two cointegrating vectors, where variables a_{11} and a_{12} represent the speed of adjustments.

$$\Pi_1 z_{t_1} = a_{11}(\beta_{11} y_{t-1} + \beta_{21} x_{t-1} + \beta_{31} w_{t-1}) + a_{12}(\beta_{12} y_{t-1} + \beta_{22} x_{t-1} + \beta_{32} w_{t-1}) \tag{3.2.14}$$

3.2.5 Cointegration with Engel-Granger

The second test for cointegration used in this thesis is the Engel-Granger test. Like the Johansen test, the Engel-Granger test shows the relationship between non-stationary series and long-run equilibrium. To understand the method, let us consider two time series x_t and y_t where $y_t \sim I(0)$ and $x_t \sim I(1)$. The linear combination of x_t and y_t can be represented as,

$$\theta_1 y_t + \theta_2 x_t \tag{3.2.15}$$

which is a stationary, or a $I(1)$ series. Equation 3.2.15 may also represent the same series if x_t and y_t are $I(1)$. The first step to estimating the parameters is conducting a test of integration, usually done with an ADF test. Should the results show that the variables are $I(1)$, a cointegration analysis may then be conducted. The next step would be to evaluate the long-run relationship between the two series. This is shown below where we test the residual of the following equation.

$$y_t = \beta_1 + \beta_2 x_t + e_t \tag{3.2.16}$$

Should the residuals from Equation 3.2.16 show that there is no cointegration, it is then concluded that the results are spurious. If this is not the case, we would have super-consistent estimators (Asteriou and Hall, 2015), meaning the parameter β is converging faster than if it was stationary.

The last consideration in the Engel-Granger process is evaluating the integration of residuals from the long-run equation, $\hat{e_t}$. By testing for the residuals with the ADF test and finding that the series is $I(0)$, we will reject the null hypothesis that x_t and y_t are not cointegrated. This then allow the application of the ECM and evaluate the long-run and short-run relationships.

3.3 Results

3.3.1 Augmented Dickey-Fuller (ADF) results

Before examining the time series, it is necessary to determine if the data is stationary or integrated $I(0)$. When performing the ADF test, BIC will be used to determine the number of lags used.

Python's `adfuller` function is applied to generate results, shown in Table 3.3.1. In summary, the results from the ADF test show that we cannot reject the null hypothesis at the 5% significance level in all cases. Given that the results show that both the national housing price index and the local housing price index are $I(1)$, we may use cointegration methods to

investigate further whether if the local macroeconomic variables are cointegrated with the local housing price index.

Table 3.3.1: ADF Test: local house indices

Unit root testing at logarithmic levels			
Variables	*Test Stats*	*P-value*	*k*
Victoria	0.801	0.992	3
Vancouver	-0.418	0.907	3
Edmonton	-2.200	0.220	3
Calgary	-2.310	0.152	3
Toronto	2.223	0.999	2
Ottawa	2.134	0.999	1
Montreal	2.232	0.999	1
Quebec City	-2.379	0.148	1
St John's	-2.339	0.160	3

Model used for this test is $\Delta y_t = \alpha_0 + \beta y_{t-1} + \alpha_2 t + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_t$.

Chosen lag for k was determined using BIC

3.3.2 Cointegration results

We will apply Johansen's test to our time series $I(1)$ to analyze cointegration. This will uncover any long-run cointegration between the local and the national housing price index. It should be noted that when it comes to hypothesis testing, there are two tests that we can reference. One of the methods is based on a likelihood ratio test for the trace matrix, where the null hypothesis is the number of cointegrating vectors less than or equal to r. The second is the maximum eigenvalue method, where the null hypothesis is that $\Pi = r$ against the scalar value $r + 1$. Both methods test for cointegration where the null hypothesis for the trace statistic contains at least one cointegrating relationship, and the eigenvalue alternative is r +1 vectors.

The results of the cointegration test are shown in Table 3.3.2 below. Johansen's test[2] show that the results are consistent with a market that is the opposite of a well-integrated market with the national housing price, with only three cointegrating vectors as indicated from the λ max statistics and two with the trace statistics. Should the national housing price be a good representation of the housing market, we would find that the results would show the opposite with a single I(1) variable.

Table 3.3.2: Cointegration Ranking for CREA data

H_0	H_1	λmax	Crt. Values	Trace	Crt. Value
$r = 0$	$r \geq 1$	111.144747	74.7434	462.380765	273.3838
$r \leq 1$	$r \geq 2$	86.202351	68.503	351.236018	328.2226
$r \leq 2$	$r \geq 3$	82.703416	62.1741	265.033667	287.1891
$r \leq 3$	$r \geq 4$	55.021025	55.8171	182.330250	250.0778
$r \leq 4$	$r \geq 5$	38.500805	49.4095	127.309225	169.9829
$r \leq 5$	$r \geq 6$	33.839452	42.8612	88.808420	97.7748
$r \leq 6$	$r \geq 7$	23.984586	36.193	54.968968	62.5202
$r \leq 7$	$r \geq 8$	19.350758	29.2631	30.984382	41.0815
$r \leq 8$	$r \geq 9$	11.631902	21.7465	11.633625	23.1485
$r \leq 9$	$r \geq 10$	11.631902	6.6349	0.001723	6.6349

Note: Evaluated at the 1% level.

Table 3.3.2 suggests that the national housing price index may be an ineffective measurement for the Canadian housing market, and using it to measure individual cities may be misleading. Furthermore, it also shows a limited long-run relationship between the national housing prices and the market index. These results are to be expected as the Canadian economy is diverse, and different variables may influence local markets. The landscape of Canada is comprised of heterogeneous provincial and municipal regions.

Given the results shown from Table 3.3.2, further evaluation into housing market hetero-

[2]To conduct the Johansen test, the Python package `coint_johansen` from the `statsmodels.tsa.vector_ar.vecm` library was used.

geneity should be done. Previous studies suggest (Abraham and Hendershott, 1994; Allen et al., 2009; Baffoe-Bonnie, 1998) that macroeconomic variables may explain fluctuations in the housing prices should the markets be deemed heterogeneous. Therefore, it is essential to choose appropriate variables for the market to derive appropriate inferences and predictions. Overall, the results from this section suggest that the Canadian housing market is not cointegrated. The next section aims to investigate local macroeconomic variables and causal inference.

4 Macroeconomic Variables

Using the Engle-Granger test to determine local macroeconomic variables will allow us to evaluate the long-run relationship between a vector of potential variables and the housing price index.

The purpose of investigating potential variables that may influence the shock in housing prices is to understand why some markets react differently and how they affect time series properties. Therefore, it is important to validate potential proxies for the housing price index to evaluate which variables impact housing prices and conclude inference. Additionally, should the variables show a long-run equilibrium, these variables may be used in ML models.

The methodology of choosing local variables was first suggested by Capozza and Helsley, 1990. They discussed urbanization from an agricultural landscape and the proxy variables which may have influenced these changes. Specifically, Capozza and Helsley, 1990 discussed the value of land, cost of development, rent increases, and household income. Additional work on this topic is also found in Abraham and Hendershott, 1994, where he used variables such as building permits, real income, interest, etc., as explanatory macroeconomic variables for the housing market. These methods are contrary to the traditional hedonic model that is frequently used for housing market analysis.

We will use the variables outlined in Abraham and Hendershott, 1994 and additional vari-

ables mentioned by Statistics Canada as key housing market indicators. These variables are: construction cost, CPI (Consumer Price Index) for shelter, construction of new prices, income, population, and the five-year mortgage rate.

4.1 Model

4.1.1 Fully-Modified OLS Model

FMOLS model is an extension of the work done by Phillips and Hansen (1990). It is created to eliminate endogeneity effects by using non-parametric methods. In other words, it eliminates the serial correlation between the cointegration equation and the explanatory variables (stochastic regressors). Applying the FMOLS model, we will further investigate cointegration and the long-run relationship between variables by evaluating the coefficients. When evaluating the parameter estimates from the FMOLS model, we can assume that the estimated parameters β are asymptotically unbiased and are asymptotically efficient (Phillips and Hansen, 1990). This will allow us to conclude the long-run relationship and further inference on macroeconomic variables in the cities studied.

4.1.2 Macroeconomic Variables

Given that local proxy variables such as building permits and union wages are limited in the available data set's time horizon, we will instead use the suggested variables from Statistic Canada and others mentioned in other literature (Abraham and Hendershott, 1994; Allen et al., 2009). We can estimate the proxy variables with the local housing price index as follows:

$$CREA_i = \beta_0 + \beta_1 ShelterIndex_t^i + \beta_2 ConstructionStart_t^i + \beta_3 Population_t^i + \beta_4 MortgageRate^i t + \beta_4 Wage_t^i + u_t^i \quad (4.1.1)$$

where the time horizon will be consistent, starting from Jan 2005 - to Jan 2022 with monthly observations. An explanation from Statistics Canada on each variable is as follows:

Construction Costs

The construction start variable[3], encompasses housing starts, housing under construction, and housing completions. Construction is defined by moment when the foundation is poured for the home or the equivalence for dwellings without a basement. Observing this variable in our model allows us to consider the activity and supply within the market.

Shelter Index

The shelter index[4], is one of the sub-indexes which make up part of the CPI. It takes the weighted average of other related shelter prices such as rented/owned accommodation and water/electricity/gas. The shelter index takes the local aggregate of the related items from the index. The inclusion of this variable provides further insights into whether individuals may be able to afford housing based on consumption costs.

Wage

The wage variable[5] is a weekly observation from the National Occupational Classification (NOC) for those over the age of 15. The table is further partitioned into wages from different professions such as management, business and finance, natural and applied sciences etc., and the final data used take into consideration all the professions. Allen et al. (2009) also cite that the inclusion of a wage variable directly relates to the costs of a house or improvements. If the wage of construction workers goes up, the cost of housing construction will increase and thus so will housing prices.

Population

The population variable[6] is the observations of those who are 15 years and older and able to

[3]Variable is from Table 34-10-015-01, from Statistics Canada, 2022e

[4]From Table 18-10-0004-01, Statistics Canada, 2022c

[5]From Table 14-10-0287-01, Statistics Canada, 2022b

[6]From Table 14-10-0287-01, Statistics Canada, 2022a

work. It is important to note that the limitation of this variable is that the individual age group could not be disaggregated. Despite this, Statistics Canada has noted that population is considered to be one of the key indicators for housing prices within Canada.

5 - Year Mortgage Rate

Last variable considered in this thesis is the 5-year mortgage rate[7]. Allen et al. (2009) use the five-year mortgage rate as a proxy for economic activity, as it shows the cost of home ownership; therefore, it is included in our analysis.

4.2 Results

Firstly, we used Engel-Granger method to determine if the selected proxy variables are cointegrated with the housing price index in the long run by taking the residuals of Equation 4.1.1 and tested for cointegration. If there is evidence that the variables are cointegrated, the FMOLS model is then applied. Using the FMOLS model allows us to make inferences to the selected parameters and evaluate their long-run relationship.

The variables shown in Equation 4.1.1 are evaluated in log-form except for the mortgage rate. Taking the residuals $\hat{u}_i$ of Equation 4.1.1 we test is they are $I(1)$. The ADF test is used next[8] to evaluate integration. The results are shown in Table 4.2.1.

These results show that cointegration exists in all cities throughout Canada, which motivates us to use the FMOLS model. Applying the FMOLS allows us to investigate the causal inference between the variables and the housing price index[9]. The output is shown in Table 4.2.2.

The results from Table 4.2.2 reveal that the parameters estimated are statistically different in each city. The coefficient signs are the same for all cities, as shown with the variables

[7]From Table 34-10-0145-01 from Statistics Canada, 2022d

[8]Applied in Python using the `adfuller` function.

[9]This is done using the `arch` and its `cointegration.FullyModifiedOLS` function.

Table 4.2.1: ADF Test of the Macroeconomic Variables and Local HPI

City	p-value
Victoria **	0.0163351
Vancouver[1]	0.091231
Edmonton *	0.016026
Calgary **	0.009144
Toronto[2]***	0.002886
Ottawa *	0.078448
Montreal***	0.002351
Quebec City ***	0.000150
St. John **	0.020592

[1] Greater Vancouver Area

[2] Greater Toronto Area

Significant at 1% [***], 5% [**], and 10% [*]

in construction and the shelter index. Construction has been concluded in literature to be a reliable and statistically significant variable used to predict housing prices (Abraham and Hendershott, 1994; Jud and Winkler, 2002). The results from Table 4.2.2 reveal that cities that report statistically significant change in construction start suggest that a one percent increase corresponds to a 0.15%to a .32% increase in housing prices. This variable is insignificant in the two cities in British Columbia, suggesting that other local idiosyncratic variables may be driving prices.

The population variable is statistically significant except for Montreal and Calgary. Edmonton's coefficient for the population variable is negative, which casts some doubt on the validity of our estimation for the city. This result is counter intuitive, but the reason may be how the data is collected using trend-cycle. Trend-cycle, as defined by Statistics Canada, is seasonally adjusted time series which has been smoothed (Fortier et al., 2019). The trend of a time series gives long-run information in seasonally adjusted data, and the cycle adjusts

Table 4.2.2: Equation 4.1.1: FMOLS Model Results

City	Shelter	Construction Start	Population	Mortgage	Wage
Victoria	3.6***	0.02	1.7**	0.2	-0.6
	(0.00)	(0.62)	(0.02)	(0.13)	(0.31)
Vancouver[1]	0.8	0.1	4.3***	0.2	-0.4
	(0.34)	(0.20)	(0.00)	(0.22)	(0.60)
Edmonton	2.9***	0.1***	-2.7**	-0.1	0.2
	(0.00)	(0.01)	(0.03)	(0.61)	(0.75)
Calgary	2.4***	0.2***	-1.5	-0.03	0.05**
	(0.00)	(0.00)	(0.17)	(0.79)	(0.95)
Toronto[2]	2.6***	0.3***	2.5**	0.0	-0.5
	(0.00)	(0.00)	(0.01)	(0.99)	(0.11)
Ottawa	1.7*	0.2***	1.2	-0.01	2.8***
	(0.07)	(0.00)	(0.51)	(0.95)	(0.00)
Montreal	2.7***	0.3***	0.8	-0.2***	0.2
	(0.00)	(0.00)	(0.27)	(0.00)	(0.16)
Quebec City	3.6***	0.05*	4.9***	-0.3***	-2.4
	(0.00)	(0.07)	(0.00)	(0.00)	(0.33)
St. John's	0.2	0.2***	14.7 ***	0.1*	0.1
	(0.41)	(0.00)	(0.00)	(0.06)	(0.41)

[1] Greater Vancouver Area

[2] Greater Toronto Area

Significant at 1% [***], 5% [**], and 10% [*]

(smooths) out seasonal data during times of expansion and contraction. The 'cycle' component of the adjusted data is defined as being smoothed around long-run trends where there is fluctuations during expansion and contraction periods.

When considering wage, half the cities report a negative relationship, and the majority of the cities show the variable as statistically insignificant. Labour mobility may explain the

difference between positive and negative coefficients within the country. Wage and labour markets are often correlated with one another and influence housing prices, as suggested by Head and Lloyd-Ellis, 2012. When wage differs between cities, the willingness to move depends on vacancy rates and housing prices. Although this paper will not go into depth about labour mobility, renters likely have higher mobility than homeowners (Head and Lloyd-Ellis, 2012).

Mortgage rates, similar to wage, exhibit different coefficients between cities, and they show that this variable is statistically insignificant in some metropolises. Economic intuition suggests that if the cost of borrowing were to decrease, the demand for housing should increase, and too will the prices. This means that the decreased cost of borrowing does not directly affect housing prices, but it is the by-product of the increased housing demand. In a paper written by Adelino et al., 2012, they pointed out that credit markets do not directly respond to housing demand, but they suggested a directional effect such that a decrease in the cost of borrowing leads to an increase in housing prices.

Findings in this section further suggested that the housing markets throughout Canada are heterogeneous. Our analysis to see if there is a long-run equilibrium with the national index against the local index suggests no cointegration - which prompts us to explore other explanatory variables. Evidence from the chosen idiosyncratic variables shows that there are also mixed results from city to city. However, overall the variables show that they cointegrated with housing prices in the long run. The next part of this paper explores how effective the chosen variables are for predicting prices.

5 Machine Learning Application

Machine learning is becoming ever more the norm for understanding data trends. Machine learning is an excellent choice in some cases because it excels at predicting dependent variables on our independent variable. The x variables used in machine learning algorithms are sometimes called *features* or *predictors*. When working with machine learning algorithms, it

is important to note that some of the coefficients on the variables may be different when we are evaluating the results. This is due largely in part that the goal of machine learning is to get the best out-of-sample predictions (Mullainathan and Spiess, 2017; Varian, 2014).

One of the issues of machine learning is overfitting our data. There are various prediction functions we may choose from, and it is essential when choosing which function to apply to our data set, we consider the following: firstly, we choose the best loss-minimization function, and secondly, we optimize our function (McInerney, 2017; Mullainathan and Spiess, 2017). Therefore it is essential to tune the algorithm using methods, such as cross-validation.

When predicting housing prices, applying machine learning methods and theory is not uncommon. There are two types of machine learning methods, supervised and unsupervised learning. Unsupervised learning is discovering patterns through clusters without a training set, supervised learning has a labeled data set and aims to predict out-of-sample (Mullainathan and Spiess, 2017). This thesis aims to apply supervised learning to the model as expressed in Section 4. We can use machine learning with its predicting capabilities while using economic theory to provide causal inference to our results.

5.1 Random Forest

A random forest can create accurate out-of-sample predictions by employing ensembles of decision trees. As outlined by Varian (2014), the problem that can arise when applying a random forest model to the data is the method can be "a bit of a black box." In other words, it is difficult to derive insights as to how our features interact with other variables within the data set. What random forest can provide us, according to Varian (2014), is how well our variable does as a predictor by providing further accuracy when included in the algorithm.

5.2 LASSO

LASSO (Least Absolute Shrinkage and Selection Operator) is a penalized regression model that allows the model to have a selected number of nonzero values. It is also considered a reliable predictor when tested in practice (Mullainathan and Spiess, 2017; Varian, 2014). The goal of a LASSO function is to minimize the following equation:

$$\sum_{i=1}^{n}(y - \sum x_{i,j}\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{5.2.1}$$

where β may be minimized to zero, the results are easier to interpret. Our λ variable is the tuning parameter in the equation, which adjusts to the amount of shrinkage. Therefore, if we have a large λ, we suggest that the estimates are biased and will be eliminated from our estimates. In the opposite remark, if we have a small or decreasing λ, the variance increases.

5.3 XGBoost

XGBoost, or extreme gradient boost, is a regularized gradient boosting algorithm that strengthens weaker predictions to produce more accurate out-of-sample predictions. One of the key features of XGBoost is that the algorithm is scalable, meaning that it is faster when processing compared to other novel boosting algorithms.

Boosting is a common method applied in machine learning algorithms to predict housing prices. This idea was first coined by Kearns, (1988). The core idea is to use weaker algorithms which may then be turned into more robust predictions rather than guessing at random. When using a boosting algorithm, the main objective is to continuously add on from the negative gradient function of a loss function. Where the loss function is defined by how well our coefficients are in conjunction with our primary data set. The additive nature of boosting is adding a higher weight to data outliers, and assigning a lower weight to data points that are easier to classify. This process repeats to improve the prediction model.

6 Implementation

This section will first discuss the primary library used to execute ML algorithms, sklearn. Following which will be a discussion of turning, training, and testing. Lastly, we will be discussing the metrics used to evaluate each model and how well they predict local housing prices using macroeconomic variables.

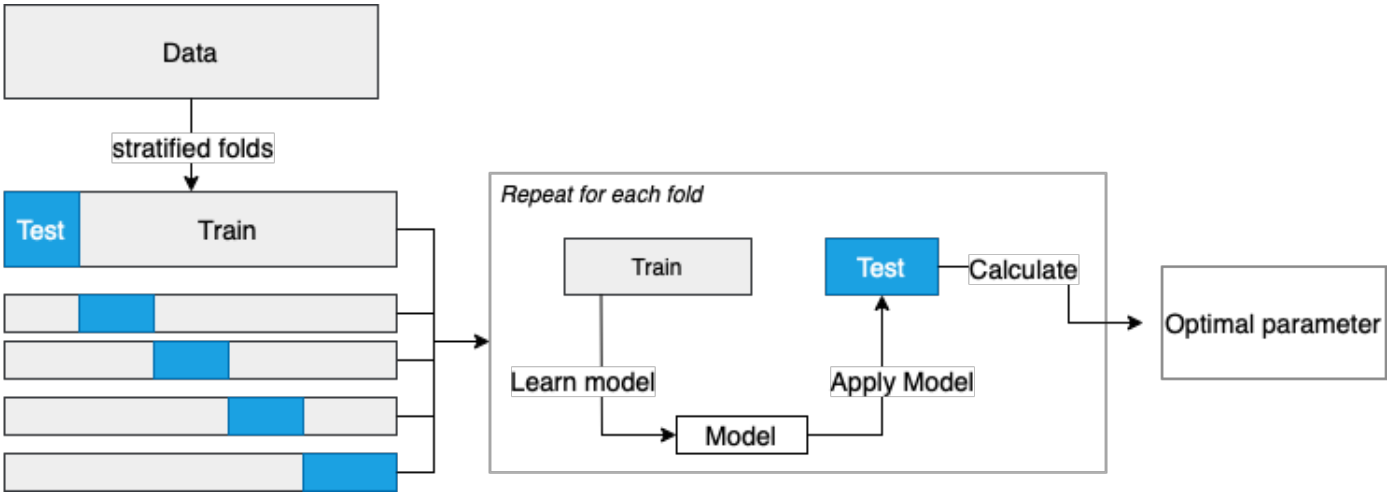
6.1 Sklearn

The main library used for ML application in this section is sklearn (or scikit-learn) and it is one of Anaconda's built-in libraries. The library has various functions such as regression and clustering. The benefits of using sklearn is that it combines SciPy and NumPy, and focuses on imperative programming (Pedregosa et al., 2011). Sklean is an efficient library for data analysis, and users benefit from its simple design and API.

6.2 Cross-validation

Tuning is a crucial step before before evaluating the performance of our ML models. To tune, we will use cross-validation (CV), where the goal is to minimize in-sample error and overfitting. Cross-validation is a statistical method that takes the data set and divide it into testing and training segments. The data used in the testing and training set are crossed over multiple times such that the data may be validated against each other. The most popular method of cross-validation is the k-fold validation. K-fold CV is defined as taking the data and equally dividing it into k parts. Where k is the number of subsets or folds. We will then fit our model k times while taking $k-1$ of the data for training and the remaining for testing. For each fold, we are rotating or stratifying the data set. Taking the average prediction of each fold will give us the optimal parameter for our model. Tuning our model as outlined in this section provided us the theoretical optimal parameters for each model.

Figure 6.1: *k-fold* Cross-Validation Schematic



Tuning for final parameter for the model

To apply these methods in Python, we will use the functions found in the `sklearn` library. Function `train_test_split` is used to split the data 20%, 80%, where 80% will be used for training and the remaining for testing. `GridSearchCV` is used to tune for the final model.

For each of the chosen ML models (XGBoost, Random Forests, and Lasso), there are unique ways to tune their hyperparameters for the final model before using the testing. For a tree-based model, we may tune the model by reducing the tree's maximum depth, which will limit the model from overfitting. For boosting model, we may control overfitting by adjusting gamma values, L1 and L2 regularization. As for lasso, tuning the hyperparameter for alpha will allow better generalization.

6.3 Evaluation of Models

Understanding how well each model performs is an integral part of this paper. Testing to see how well macroeconomic variables work in predicting local housing prices may provide us with further causal inference. The best performance indicators are Root Mean Square

Error (RMSE) and Mean Absolute Error (MAE). These metrics will be reported for all cities. Although there are ongoing debate about which indicator is better, this paper will not explore the topic in great depth and simply report both.

6.3.1 RMSE

As shown in the equation below, RMSE is the standard deviation of the residuals, and the value it returns explains how closely the predictions are from the actual values. How RMSE differs from MAE is that it penalizes variance and gives more significant weight to larger absolute values rather than smaller absolute values (Chai and Draxler, 2014). RMSE is also known for being sensitive to outliers, and outliers in the data set must be taken into careful consideration (Chai and Draxler, 2014).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=i}^{n}(y_i - \hat{y}_i)^2} \tag{6.3.1}$$

RMSE is commonly reported in ML evaluations, and one of its advantages over using MAE is that RMSE does not use absolute value, which is not preferred in some calculations (Chai and Draxler, 2014). Overall, when it comes to understanding how well a model perform, comprehending the data set and how RMSE is applied will reveal the performance of our models and how accurately they predict the pricing.

6.3.2 MAE

MAE is another standard metric used in determining how well a model fits the data, where the same weight to all errors is returned. MAE is calculated using the summation of the absolute values of the errors divided by n, where x_i is the actual value, y_i is the prediction value, and n is the number of observations.

$$MAE = \frac{1}{n}\sum_{i=i}^{n}|y_i - x_i| \tag{6.3.2}$$

In literature, one of the reasons to choose MAE over RMSE is because there is a lower sample variance in comparison (Brassington, 2017). Furthermore, one of the benefits of using MAE is that it is a linear model and, as such, provides little ambiguity (Brassington, 2017; Chai and Draxler, 2014; Willmott and Matsuura, 2005). However, this is debated in the literature.

6.3.3 R^2

R^2 is a commonly used statistical measure that explains the amount of variance between the dependent and independent variables in a given model. R^2 is represented from 0 to 1, where 0 means our model does not explain the variability around the mean, while 1 means the opposite. RMSE, MAE, and R^2 will be used in conjunction to determine the goodness of fit for our ML application.

7 Results

We begin by describing the results of each model from each city, starting with Lasso, then Random Forest, and lastly XGBoost. Discussion of the overall results based on what is reported from the previously discussed metrics will follow. Additionally, evaluation of the hyperparameters will also be discussed to highlight their significance and how they impact the prediction of the model.

7.1 Data

The features used in the machine learning models are the shelter index, population, construction starts, mortgage rate, and wage. There are 205 observations for each feature in

each city, see Appendix A. Shown in Table 7.1.1 below is an example data set used for Victoria.

Table 7.1.1: Summary Statistics: Victoria

	Victoria	Shelter	Population	Const.	Mort.Rate	Wage
count	205	205	205	205	205	205
mean	516504	116	3854	34	4	25
std	135581	6	301	8	0.95	3
min	304700	105	3370	13	3	20
25%	432300	113	3595	27	4	23
50%	456700	114	3832	36	4	26
75%	632200	120	4109	40	5	28
max	920400	137	4396	54	7	33

Note: Victoria represents the local housing price index in Canadian Dollars. *Population*, has been adjusted and may be represented by x1,000. *Const.* (x 1,000) represents construction starts, and *Mort.Rate* is the 5 year mortgage rate as a percentage. *Wage* is represented in Canadian Dollars.

7.2 Hyperparameters

The hyperparameters used from each are shown in Table 7.2.1, 7.2.2 and 7.2.3. Cross-validation techniques will be used to test for the optimal hyperparameters for each model, and the outcome is shown in the following tables.

In Table 7.2.1 below lists the hyperparameters for XGBoost for each city are shown. The parameters tuned are `learning_rate`, `max_depth`, and `min_child_weight`. To adjust `learning_rate`, the numeric value is between [0,1]. This hyperparameter determines how the model discovers patterns in the data. Adjusting `learning_rate` will prevent overfitting by shrinking the weights of the features. `max_depth` controls the depth of the trees, which also changes the degree overfitting by altering the complexity of the trees.

Lastly, `min_child_weight` adjusts for the minimum hessian that is required for a child. This variable also controls the depth of the tree by stopping the partitioning of a tree when the sum of the instance weight is less than min child weight. With an increase in the `min_child_weight` variable, the model would become more conservative.

Table 7.2.1: XGBoost: Hyperparameters

City	learning_rate	max_depth	min_child_weight
Victoria	0.1	3	2
Vancouver	0.15	3	2
Edmonton	0.15	3	4
Calgary	0.1	3	4
Toronto	0.1	3	2
Ottawa	0.1	4	2
Montreal	0.1	3	2
Quebec City	0.1	3	2
St John's	0.15	4	2

Table 7.2.2 below is the tuning parameters for the lasso function. It is tuned by adjusting the parameter `alpha`, which can be also represetned as λ. Since Lasso is a penalized regression, adjusting `alpha` is equivalent to adjusting the L1 penalty, which controls the number of features in a model. The closer the parameter `alpha` approaches 1, the fewer features are included in the model.

Table 7.2.2: Lasso: Hyperparameters

City	alpha
Victoria	0.00001
Vancouver	0.0001
Edmonton	0.0005
Calgary	0.0001
Toronto	0.0001
Ottawa	0.001
Montreal	0.0001
Quebec City	0.001
St John's	0.0001

Lastly, Table 7.2.3 below displays the hyperparameters set for the random forest model. The `max_depth` variable is similar to XGBoosts', where the argument dictates the depth of a tree. `max_features` takes into account specific features when there is a split, while `min_samples_split` controls the minimum amount of samples needed before the internal node splits. Lastly, the `min_samples_leaf` adjusts the minimum number of samples required for a node.

Table 7.2.3: Random Forest: Hyperparameter

City	max_depth	max_features	min_samples_leaf	min_samples_split
Victoria	3	2	0.1	4
Vancouver	3	0.5	0.1	4
Edmonton	3	2	0.1	3
Calgary	3	0.5	0.1	4
Toronto	3	0.5	0.1	4
Ottawa	4	0.25	0.1	3
Montreal	4	0.5	0.1	2
Quebec City	4	0.5	0.1	3
St John's	3	2	0.1	4

7.3 Overall Results

The results from the testing data are shown below in Table 7.3.1. XGBoost is the best predictor for our data due to having the lowest RMSE. This is partly due to the ensemble of the trees which in theory gives better results. When using ensemble methods, specifically an ensemble of tree, the goal is to take individually built trees and improve generalizability.

Although XGBoost has been known to outperform other ML methods over the years partly due to its scalability (Chen and Guestrin, 2016), this paper will not go into depth about its competitive advantage due to scope limitation. Furthermore, recall that its predictions are based on gradient boosting, where the predictive model is an ensemble of weak predictions, which are decision trees.

Table 7.3.1: Testing Set, Monthly Observations, 18-08-01 to 22-01-01

XGBoost			
City	**RMSE**	**MAE**	R^2
Victoria	0.0346	0.029584	0.86263
Vancouver	0.0373	0.025955	0.915601
Edmonton	0.0324	0.025955	0.721992
Calgary	0.010	0.006747	0.896348
Toronto	0.0370	0.032329	0.934454
Ottawa	0.0308	0.117473	0.876292
Montreal	0.0259	0.022531	0.883726
Quebec City	0.0266	0.024240	0.891436
St John	0.0160	0.012088	0.886417

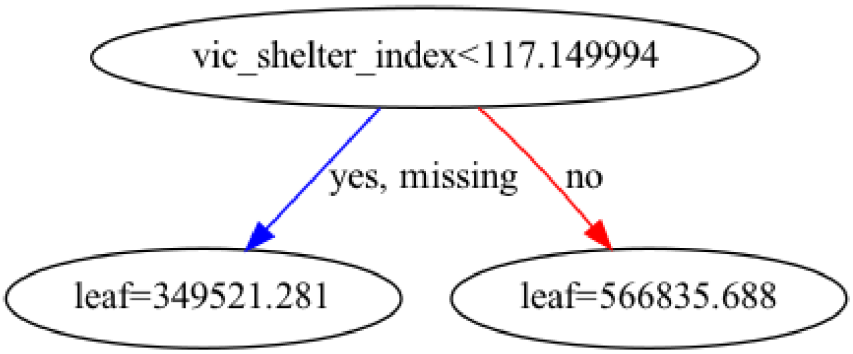
Lasso			
City	RMSE	MAE	R^2
Victoria	0.0583	0.045312	0.9449
Vancouver	0.0781	0.068042	0.9414
Edmonton	0.0834	0.067778	0.7807
Calgary	0.0663	0.052312	0.8584
Toronto	0.0489	0.037947	0.9830
Ottawa	0.0743	0.056856	0.9146
Montreal	0.0267	0.020983	0.9867
Quebec City	0.0370	0.028849	0.9669
St John	0.1065	0.088711	0.7984

Random Forest			
City	RMSE	MAE	R^2
Victoria	0.0683	0.049699	0.961513
Vancouver	0.0737	0.059236	0.973555
Edmonton	0.1194	0.058325	0.742037
Calgary	0.1062	0.054691	0.797894
Toronto	0.0693	0.049506	0.982724
Ottawa	0.0855	0.054193	0.941599
Montreal	0.0722	0.046684	0.902745
Quebec City	0.0658	0.045012	0.895452
St John	0.0576	0.041794	0.973483

Although these results are impressive at first glance, it should be noted that there is a relatively high R^2 amongst all the results, with some approaching 1. These results tells us that the models being used are overfitting the data. Although cross-validation techniques were applied to avoid this, it is clear there are other factors. Additionally, it should be noted that the R^2 values from the testing set show that the values differ in each city, with Edmonton reporting an R^2 of 72% and Toronto of 93% (as shown from XGBoost). Lasso and random forest are both similar in this respect.

Further investigation of tree's structure shows that XGBoost, while being the best predictor, gives us a tree that is overly simple. Further examples are shown in Appendix D. The Figure 7.1 below shows that our model is limited given that it is a stump with a root and two leaves. With the results from our R^2 values and the limited complexity of the trees, we should treat our results with skepticism. Another note of concern is the time dimension component in the data set due to possible serial correlation between each data point in the set, and therefore is not truly independent of one another.

Figure 7.1: Victoria - XGBoost Tree



7.3.1 XGBoost and further adjustments

The result from the previous section invites the opportunity for further tuning. Despite XGBoost performing the best, and many different tuning adjustments such as pruning, regularization, sampling, and early stopping are made, over fitment was unavoidable. Pruning a tree can be done by adjusting `gamma`, `min_child_weight`, and `max_depth`. Tuning `gamma` adjusts for regularization of the model. An increase to `gamma` dials down the complexity of the trees by minimizing the loss reduction from a split and makes the model more conservative. Adjusting these hyperparameters will reduce the size of the decision tree and remove splits during the building process. The Table 7.3.2 shows the new hyperparameters.

Table 7.3.2: XGBoost: Adjusted Hyperparameters

City	gamma	learning_rate	max_depth	min_child_weight
Victoria	3	0.1	3	4
Vancouver	3	0.15	3	4
Edmonton	2	0.15	3	6
Calgary	2	0.1	3	6
Toronto	3	0.1	3	4
Ottawa	3	0.1	2	5
Montreal	3	0.1	3	4
Quebec City	3	0.1	3	4
St John's	3	0.15	2	4

The new results from the additional tuning are shown in Table 7.3.3. Adjusting those hyperparameters showed that we have not corrected for overfitting. R^2 is notably smaller for some observations, while some cities continue to have high R^2. Given that the sample size is relatively small, tuning the model even more may not benefit the overall predictions. Instead, adding more variables to our model can be the next logical step. Furthermore, serial correlation issue may continue to be a problem if left untreated. Therefore, in addition to adding more variables, correcting this may help the out-of-sample predictions.

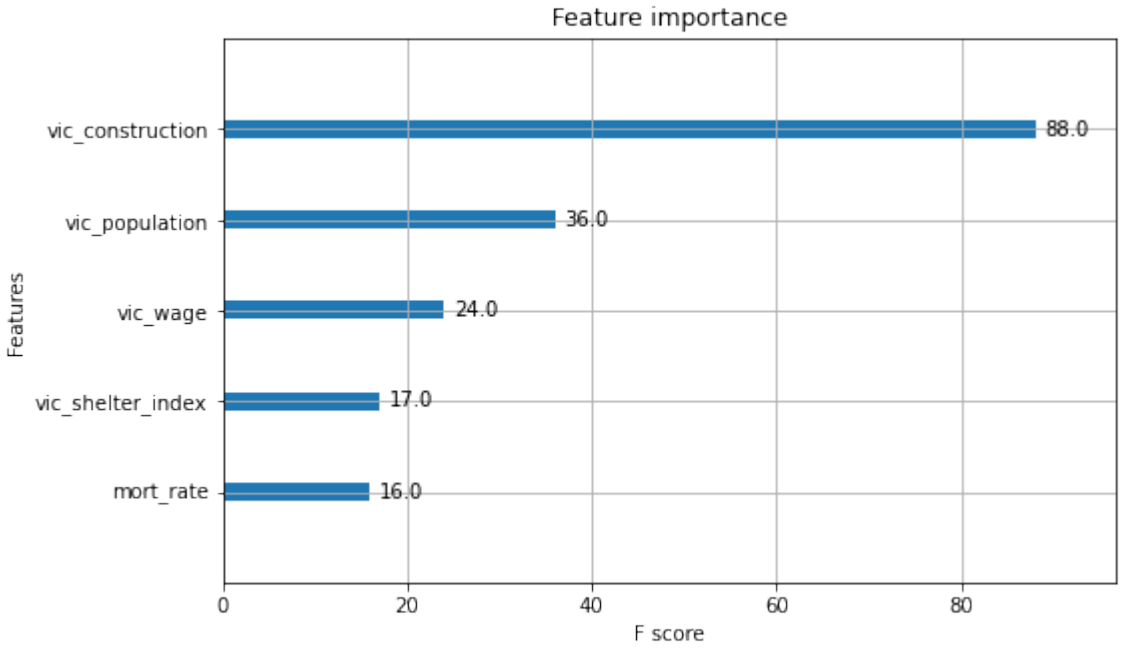
Table 7.3.3: XGBoost: Adjusted XGBoost Out-of-Sample Prediction

City	**RMSE**	**MAE**	R^2
Victoria	30165.6030	21342.304878	0.971302
Vancouver	33488.5193	26570.189787	0.990225
Edmonton	28534.2290	15106.891768	0.599322
Calgary	30471.7925	18144.892530	0.734031
Toronto	42794.6442	27463.105183	0.980453
Ottawa	45993.5932	25325.078506	0.884395
Montreal	20065.3784	14180.267912	0.954495
Quebec City	10691.4829	8093.995427	0.963374
St John	6165.2103	4766.832317	0.992176

7.3.2 Variable Importance

The XGBoost model is further evaluated with the feature importance of each city. Shown below is Figure 7.2 is an example of the variable importance for the city of Victoria. Further evaluation of other cities can be found in Appendix C.

Figure 7.2: Victoria - Variable Importance



Evaluating the features[10] for Victoria, the variable that frequently appears in our tree is `construction`, and the second feature of importance is `vic_population`. This finding is almost consistent with the FMOLS model, where `construction` has a positive coefficient. Besides Victoria, all cities report `construction` to have the largest weight, while the second feature differs from city to city.

On the other hand, to evaluate the importance of our feature and their influence on the model, we can refer to the plotted tree graph of XGBoost. This is shown in Appendix C and has been highlighted above in Figure 7.1. The variables which are the most influential for the model's calculation vary between `population` and `shelter_index`. The graphical representation of XGBoost shows that we have a simple decision tree. Although causal inference cannot be made directly from ML models, we can note that the `population` and `shelter` are important variables which influence the model the most. These results allude back to the FMOLS model, where `population` and `shelter` have relatively consistent coefficients for cities tested in our data. Statistics Canada suggests that the `shelter` is one

[10] `get_score()` function is used, and default argument for `importance_type` is applied.

of the most important variables when determining housing prices for the national housing index, as itself encompasses various factors that can determine the price of a home.

7.4 Limitations

The machine learning model above has provided limited understanding of causal inference. An improvement to this analysis can include the construction of a feature selection model. Creating a correlation matrix where there is a clear visualization of the relationship between the variables can help find patterns within the data.

The most prominent limitation of this analysis is the small sample size and the prevalence of overfitting caused by the small time dimension. Should the model have more features and data points, more accurate insights and detailed relationships can then formed by using a correlation matrix.

In conclusion, this study highlights the disadvantages of having a small data set in ML modelling, especially with values that are not truly independent of each other due to the limited time dimension. Nonetheless, contrary to the ML results in this paper, an abundant amount of literature supports ML and economic predictions, and an expanded study should be done to further analyze the feasibility of using ML modelling to predict housing prices.

8 Concluding Remarks

The results of this paper have explored non-exhaustive estimations for the Canadian housing market. Section 1 and 2 introduced current issues arising within the housing market and how we could potentially answer them with economic theory and machine learning methods.

Sections 3 and 4 explored and answered whether there was market segregation in Canada and if we could use a housing price index for the nation. Applying time series analysis and cointegration methods showed that Canadian markets are segregated, and the use of national

housing price index can misrepresent the regional markets. These results prompted further exploration of other explanatory variables.

Those variables for local markets proved to be viable predictors for housing prices, albeit with inconsistency throughout the country. The chosen variables were also tested using cointegration methods and showed that they were cointegrated in the long run. The variables were then applied to the FMOLS model, which further proved that the variables had different inferences per city, suggesting the presence of a heterogeneous housing market. The local variables used in the FMOLS model were then used to try and make out-of-sample predictions in the local housing markets.

Results from the machine learning should be taken with skepticism, as there is evidence from our model to overfit the data in the out-of-sample estimation. Further investigation of the model shows that the features used in the model are limited, and in some cases, the tree is represented as a stump with a root and two leaves. Results are shown in Section 7, and introduce the theory used in this section was explored in Sections 5 and 6.

This paper has shown that the Canadian housing market is segmented. Investigating the macroeconomic variables used in the FMOLS model proved to be asymptotically unbiased estimators, and can be used to make inferences in local markets. Lastly, the application of machine learning should in theory work in predicting prices, but the results in this study is not conclusive due to the small data size. This study, however, did prove that that macroeconomic variables can viable features in house price predicting ML algorithms.

8.1 Future Research

Further work that can improve the topic explored in this thesis include investigating if local housing indexes are an appropriate estimation for the market they represent. Given that housing is a heterogeneous commodity, supply and demand within a localized market may also drive markets to be segmented due to their location and price (Goodman and Thibodeau, 1998). Therefore, an analysis of microeconomic variables such as consumer

housing preferences may be coupled to more accurately construct housing prices models especially if a metropolis contains sub-markets.

By conducting a thorough investigation of cities throughout Canada, we can discover additional explanatory variables to act as features for our ML models. Selected features for the model can then vetted using a correlation matrix. With additional features and data, we may conclude further insights and inference to local housing markets.

References

Abadie, A., & Kasy, M. (2018). Choosing among regularized estimators in empirical economics. *Review of Economics and Statistics, forthcoming*.

Abraham, J. M., & Hendershott, P. H. (1994). Bubbles in metropolitan housing markets.

Adelino, M., Schoar, A., & Severino, F. (2012). *Credit supply and house prices: Evidence from mortgage market segmentation* (tech. rep.). National Bureau of Economic Research.

Allan, R. (2019). *An inquiry into the toronto and vancouver housing markets and assessing the impact of the change in mortgage rules and foreign buyer's tax policy* (Doctoral dissertation). Queen's University.

Allen, J., Amano, R., Byrne, D. P., & Gregory, A. W. (2009). Canadian city housing prices and urban market segmentation. *Canadian Journal of Economics/Revue canadienne d'économique*, *42*(3), 1132–1149.

Asteriou, D., & Hall, S. G. (2015). *Applied econometrics*. Macmillan International Higher Education.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Baffoe-Bonnie, J. (1998). The dynamic impact of macroeconomic aggregates on housing prices and stock of houses: A national and regional analysis. *The Journal of Real Estate Finance and Economics*, *17*(2), 179–197.

Belloni, A., Chernozhukov, V., Fernández-Val, I., & Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, *85*(1), 233–298.

Bontempi, G., Ben Taieb, S., & Borgne, Y.-A. L. (2012). Machine learning strategies for time series forecasting. *European business intelligence summer school*, 62–77.

Brassington, G. (2017). Mean absolute error and root mean square error: Which is the better metric for assessing model performance? *EGU General Assembly Conference Abstracts*, 3574.

Bunce, S., Livingstone, N., March, L., Moore, S., & Walks, A. (2020). *Critical dialogues of urban governance, development and activism: London and toronto*. University College London.

Capozza, D. R., & Helsley, R. W. (1990). The stochastic city. *Journal of urban Economics*, *28*(2), 187–203.

Carney, M. (2011). *Housing in canada*. Retrieved November 10, 2021, from https://www.bankofcanada.ca/2011/06/housing-in-canada

Chai, T., & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, *7*(3), 1247–1250.

Charpentier, A., Flachaire, E., & Ly, A. (2018). Econometrics and machine learning. *Economie et Statistique*, *505*(1), 147–169.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Coulombe, P. G., Leroux, M., Stevanovic, D., & Surprenant, S. (2020). How is machine learning useful for macroeconomic forecasting? *arXiv preprint arXiv:2008.12477*.

Dahms, K., & Ducharme, A. (2021). *Housing affordability deteriorates for a third consecutive quarter in q3 202*. Retrieved October 18, 2021, from https://www.nbc.ca/content/dam/bnc/en/rates-and-analysis/economic-analysis/housing-affordability.pdf

Demers, F. et al. (2005). *Modelling and forecasting housing investment: The case of canada* (tech. rep.). Bank of Canada.

Fortier, S., Matthews, S., & Gellatly, G. (2019). Trend-cycle estimates – frequently asked questions. Retrieved May 5, 2022, from https://www.statcan.gc.ca/en/dai/btd/tce-faq

Goodman, A. C., & Thibodeau, T. G. (1998). Housing market segmentation. *Journal of housing economics*, *7*(2), 121–143.

Götz, T. B., & Knetsch, T. A. (2019). Google data in bridge equation models for german gdp. *International Journal of Forecasting*, *35*(1), 45–66.

Granger, C. W., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of econometrics*, *2*(2), 111–120.

Head, A., & Lloyd-Ellis, H. (2012). Housing liquidity, mobility, and the labour market. *Review of Economic Studies*, *79*(4), 1559–1589.

Jud, G. D., & Winkler, D. T. (2002). The dynamics of metropolitan housing prices. *The journal of real estate research*, *23*(1/2), 29–46.

Kearns, M. (1988). Thoughts on hypothesis boosting. *Unpublished manuscript*, *45*, 105.

Khan, M., Bilyk, O., & Ackman, M. (2021). *Update on housing market imbalances and household indebtedness*. Retrieved November 10, 2021, from https://www.bankofcanada.ca/2021/04/staff-analytical-note-2021-4/

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237–293.

MacGee, J. (2009). Why didn't canada's housing market go bust?, sl: Federal reserve bank of cleveland.

McInerney, J. (2017). An empirical bayes approach to optimizing machine learning algorithms. *NIPS*, 2712–2721.

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology*, *45*, 27–45.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Oren, K., Tony, S., & Michael, D. (2021). *Affordable housing will get increasingly harder to find*. Retrieved November 10, 2021, from https://resources.oxfordeconomics.com/hubfs/Content%5C%20Hub%5C%20RBs/open20210518012500.pdf

Paruchuri, H. (2021). Conceptualization of machine learning in economic forecasting. *Asian Business Review*, *11*(2), 51–58.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Phillips, P. C., & Hansen, B. E. (1990). Statistical inference in instrumental variables regression with i (1) processes. *The Review of Economic Studies*, *57*(1), 99–125.

Piazzesi, M., & Schneider, M. (2016). Housing and macroeconomics. *Handbook of macroeconomics*, *2*, 1547–1640.

Renigier-Biłozor, M., & Wiśniewski, R. (2012). The impact of macroeconomic factors on residential property prices indices in europe. *Aestimum*, 149–166.

Schembri, L. L. (2014). Housing finance in canada: Looking back to move forward. *National Institute Economic Review*, *230*, R45–R57.

Singh, A. (2021). *Canada housing market: Slower price growth*. Retrieved October 18, 2021, from https://www.moodysanalytics.com/-/media/article/2021/10-canada-housing-market-outlook.pdf

Statista. (2022). House price to income ratio in canada from 3rd quarter 2015 to 4th quarter 2021. Retrieved April 12, 2022, from https://www.statista.com/statistics/591782/house-price-to-income-ratio-canada/

Statistics Canada. (2022a). *Table 14-10-0287-01. Labour force characteristics, monthly seasonally adjusted and trend-cycle, last 5 months* [Data table]. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410028701

Statistics Canada. (2022b). *Table 14-10-0306-01. Employee wages by occupation, monthly, unadjusted for seasonality* [Data table]. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410030601

Statistics Canada. (2022c). *Table 18-10-0004-01. Consumer Price Index, monthly, not seasonally adjusted* [Data table]. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810000401

Statistics Canada. (2022d). *Table 34-10-0145-01 Canada Mortgage and Housing Corporation, conventional mortgage lending rate, 5-year term* [Data table]. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410014501

Statistics Canada. (2022e). *Table 34-10-0159-01. Canada Mortgage and Housing Corporation, housing starts, all areas, Canada and provinces, 6-month moving average* [Data table]. https://doi.org/10.25318/3410015901-eng

Tay, F. E., & Shen, L. (2002). Economic and financial prediction using rough sets model. *European Journal of Operational Research*, *141*(3), 641–659.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

Visvizi, A., Lytras, M. D., Damiani, E., & Mathkour, H. (2018). Policy making for smart cities: Innovation and social inclusive economic growth for sustainability. *Journal of Science and Technology Policy Management*.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, *30*(1), 79–82.

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. *Journal of the royal statistical society*, *89*(1), 1–63.

Appendices

A Macroeconomic Variables Summary Statistics

All dollar values from each city variable and wage are represented in Canadian Dollars. *Const (x1,000)* or construction, represents construction starts. *Mort. Rate* is the 5-year mortgage rate and is represented as a percentage. The population variable has also been adjusted and may by x1,000.

The time horizon for all the data sets in Appendix A is from 01/01/05 - 01/01/22. All observations are monthly. The data collected is from Statistics Canada.

Table A.0.1: Summary Statistics: Vancouver

	Vancouver	Shelter	Population	Const.	Mort.Rate	Wage
count	205	205	205	205	205	205
mean	729692	116	3854	34	5	26
std	237954	7	302	8	0.95	3
min	379900	105	3370	123	3	20
25%	556200	113	3596	27	4	24
50%	6174000	114	3832	36	4	26
75%	989700 120	4109	40	5	28	
max	1255200	138	4396	54	7	34

Table A.0.2: Summary Statistics: Edmonton

	Edmonton	Shelter	Population	Const.	Mort.Rate	Wage
count	205	205	205	205	205	205
mean	313403	155	31230	33	5	29
std	41733	17	285	9	0.95	4
min	175600	110	2569	14	3	20
25%	307600	148	2911	26	4	26
50%	323800	157	3164	31	4	29
75%	338200	166	3362	39	5	32
max	376000	186	3574	53	7	36

Table A.0.3: Summary Statistics: Calgary

	Calgary	Shelter	Population	Const.	Mort.Rate	Wage
count	205	205	205	205	205	205
mean	389620	155	3130	33	5	29
std	51966	17	284	8	0.95	4
min	215500	109	2568	14	3	20
25%	368300.	147	2911	26	4	25
50%	402300	157	3164	31	4	29
75%	423900	166	3362	38	5	32
max	458800	186	3573	53	7	36

Table A.0.4: Summary Statistics: Toronto

	Toronto	Shelter	Population	Const.	Mort.Rate	Wage
count	205	205	205	205	205	205
mean	563243	130	11133	72	5	27
std	225529	14	695	11	.95	3
min	305500	107	9982	45	3	21
25%	367200	119	10542	64	4	24
50%	479700	127	11094	73	4	26
75%	767700	141	11668	79	5	28
max	1259900	162	12436	107	7	34

Table A.0.5: Summary Statistics: Ottawa

	Ottawa	Shelter	Population	Const.	Mort.Rate	Wage
count	205	205	205	205	205	205
mean	359757	130	11133	72	5	27
std	102708	14	695	11	1	3
min	234700	107	9982	45	3	21
25%	286100	119	10542	64	4	24
50%	345800	127	11094	73	4	26
75%	377900	141	11668	79	5	28
max	689700	162	12436	107	7	34

Table A.0.6: Summary Statistics: Montreal

	Montreal	Shelter	Population	Const.	Mort.Rate	Wage
count	205	205	205	205	205	205
mean	291934	126	6666	47	5	23
std	73588	9	278	8	1	3
min	192600	108	6129	32	3	18
25%	237500	120	6446	41	4	21
50%	284500	126	6711	48	4	22
75%	315000	131	6866	50	5	25
max	530100	146	7124	81	7	30

Table A.0.7: Summary Statistics: Quebec City

	Quebec City	Shelter	Population	Const.	Mort.Rate	Wage
count	205	205	205	205	205	205
mean	220352	126	6666	47	5	23
std	42062	9	278	8	1	3
min	132100	108	6129	32	3	18
25%	187800	120	6446	41	4	21
50%	239600	126	6711	48	4	22
75%	244400	131	6866	50	5	25
max	305800	146	7124	81	7	30

Table A.0.8: Summary Statistics: St. John's

	St. John's	Shelter	Population	Const.	Mort.Rate	Wage
count	205	205	205	205	205	205
mean	241095	143	440	11	5	24
std	52237	17	8	2	1	4
min	139500	110	424	6	3	16
25%	202200	129	431	9	4	20
50%	264100	147	444	11	4	25
75%	281700	157	446	12	5	27
max	292500	172	448	15	7	31

B XGBoost: Out-of-Sample Graphs

The out-of-sample predictions are shown in Appendix B. Each prediction below shows the test set against the predicted values. The time horizon for the observations is from 2018-09-01 - 2022-01-01. Observations are monthly.

Figure B.1: Victoria - Out of Sample

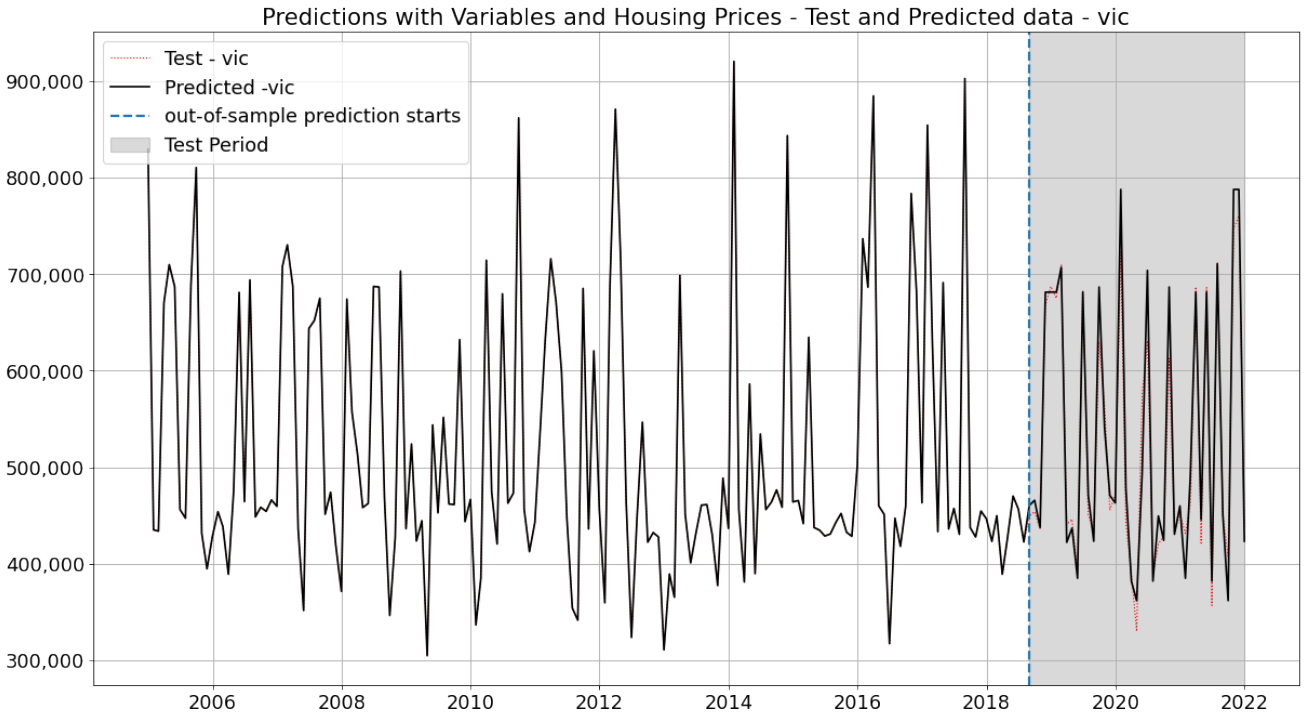


Figure B.2: Greater Vancouver Area - Out of Sample

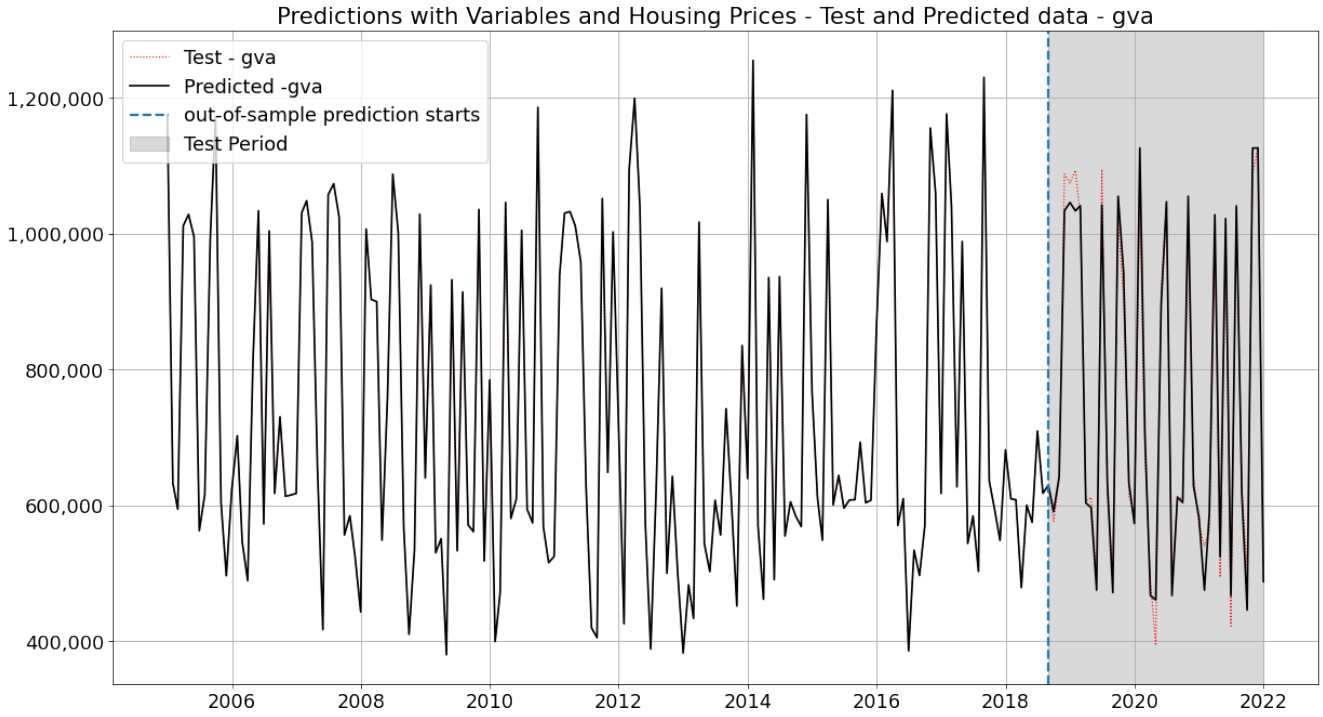


Figure B.3: Edmonton - Out of Sample

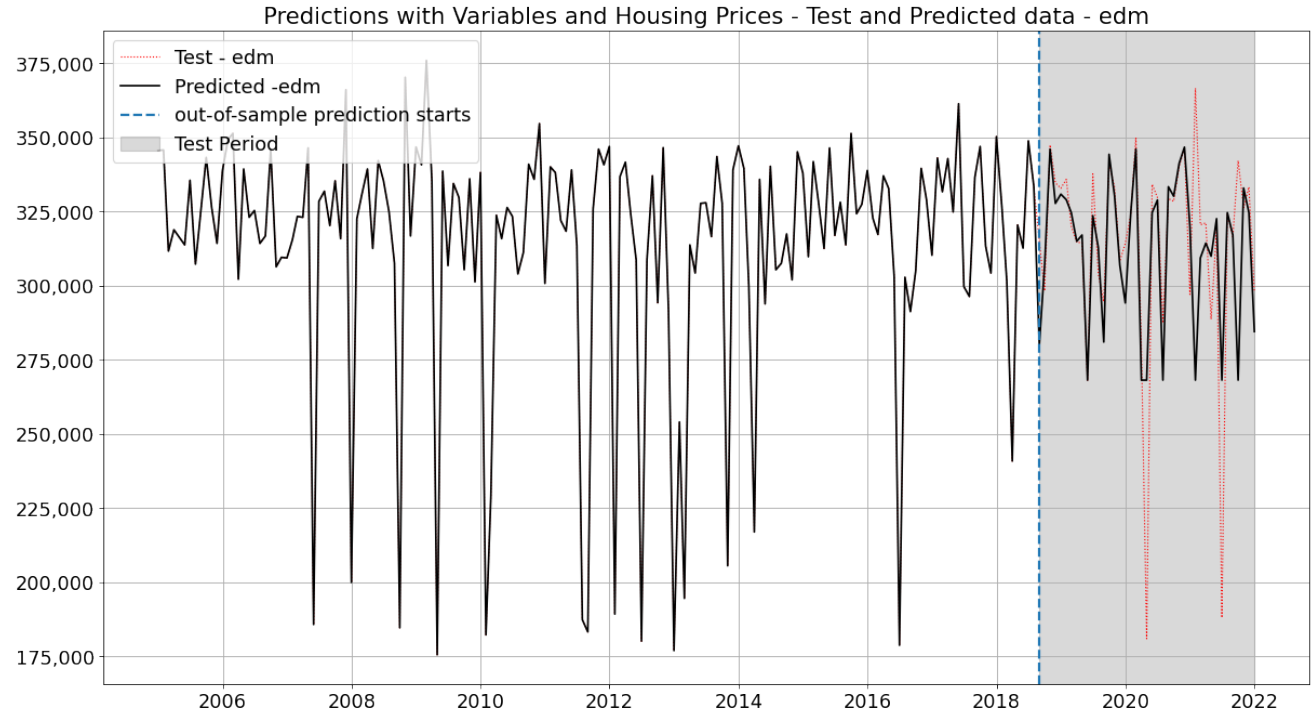
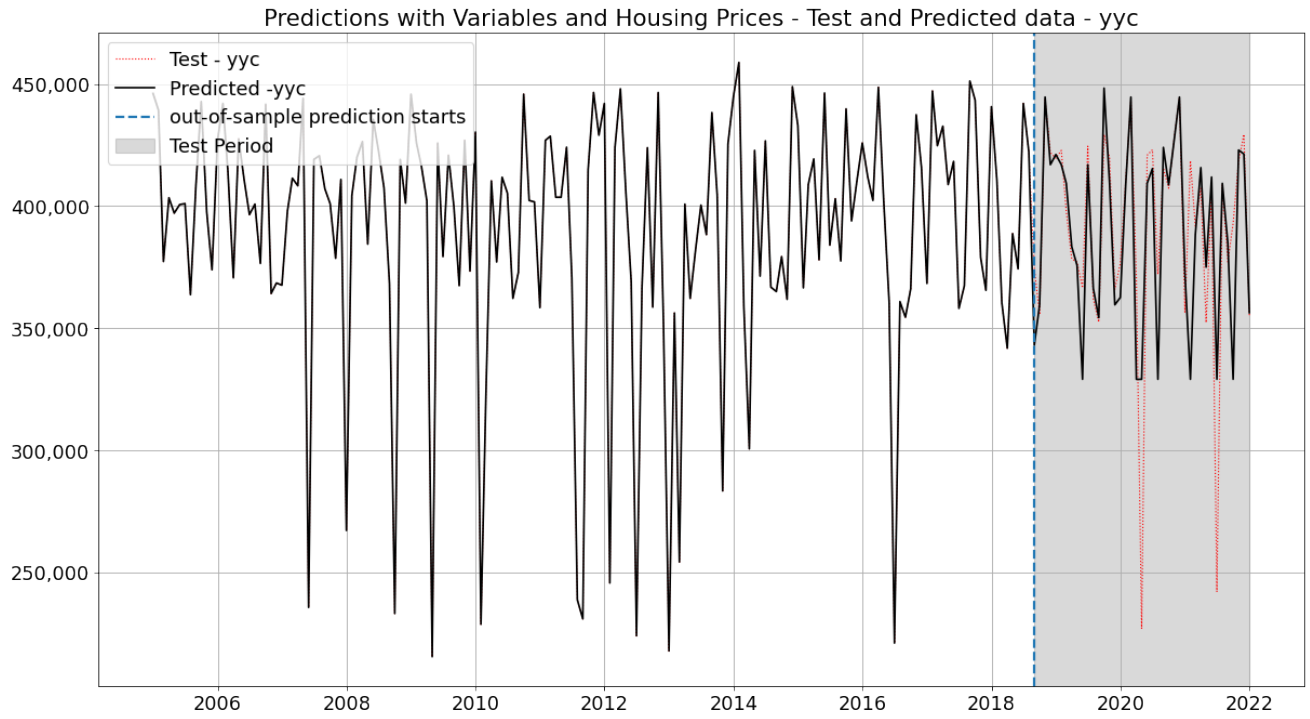


Figure B.4: Calgary - Out of Sample



* Note: YYC is Calgary respectively

Figure B.5: Greater Toronto Area - Out of Sample

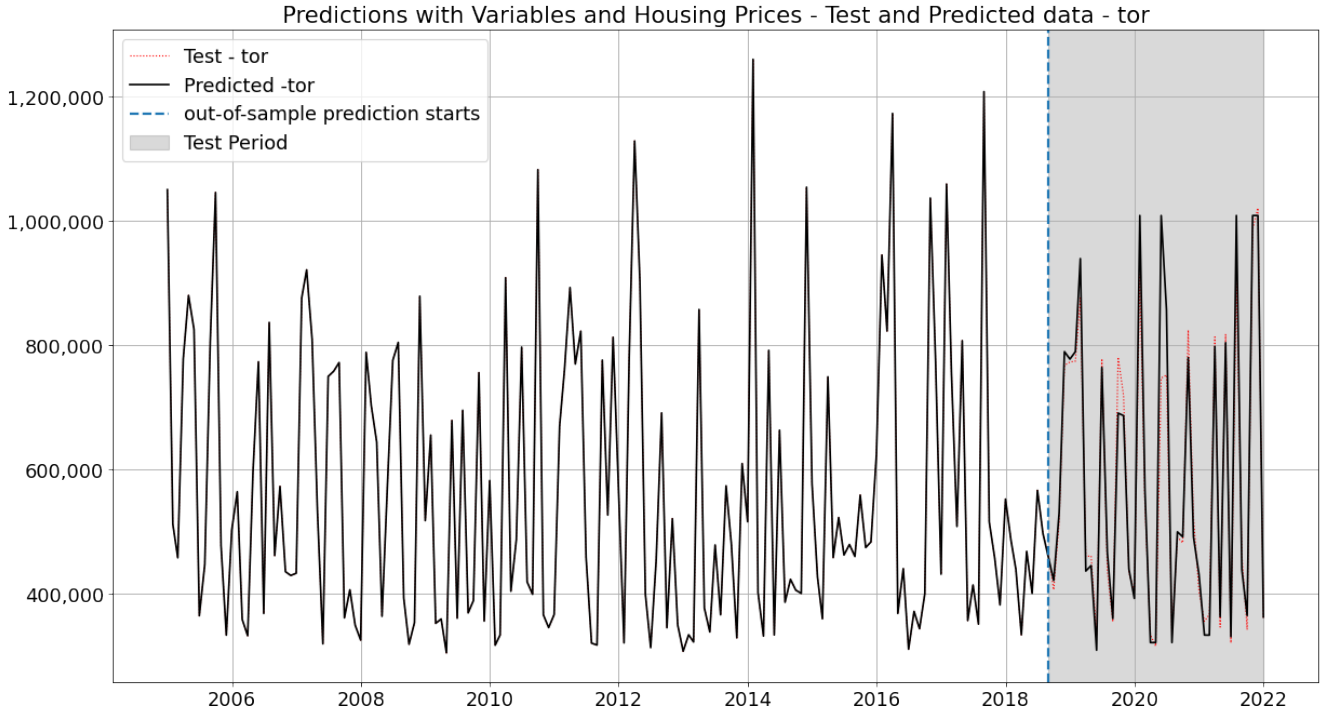


Figure B.6: Ottawa - Out of Sample

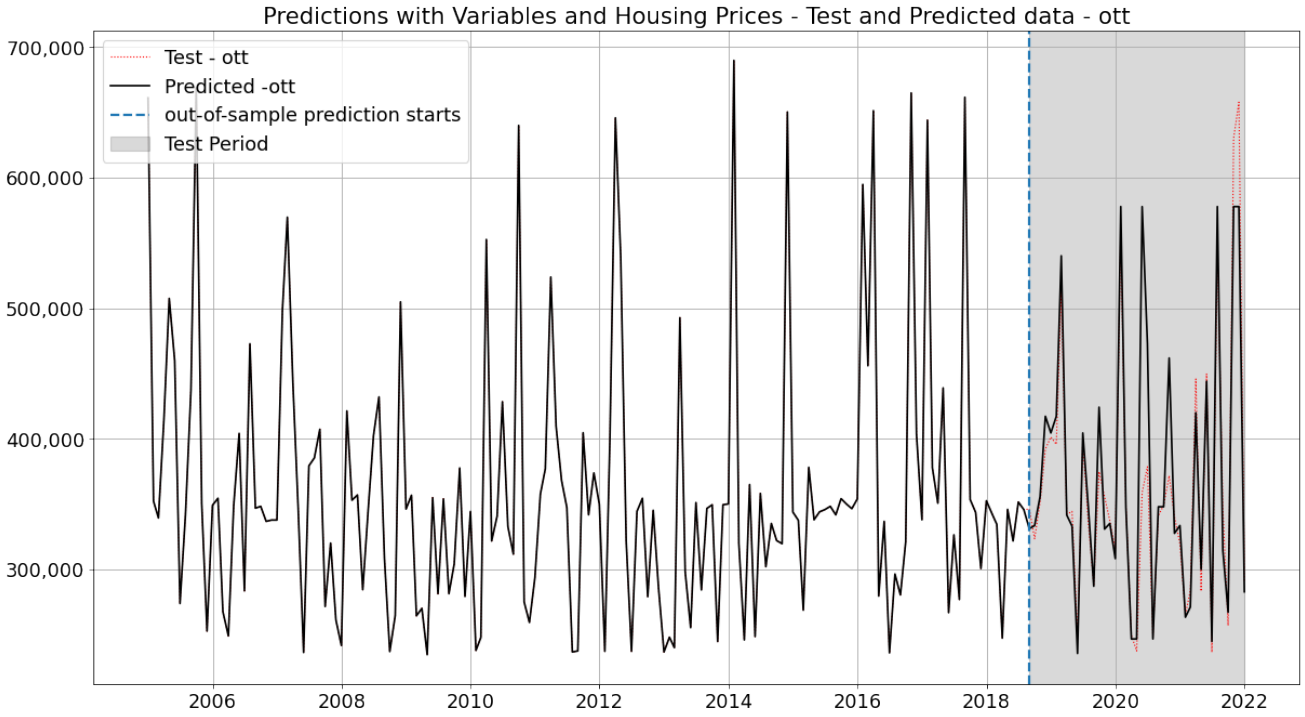
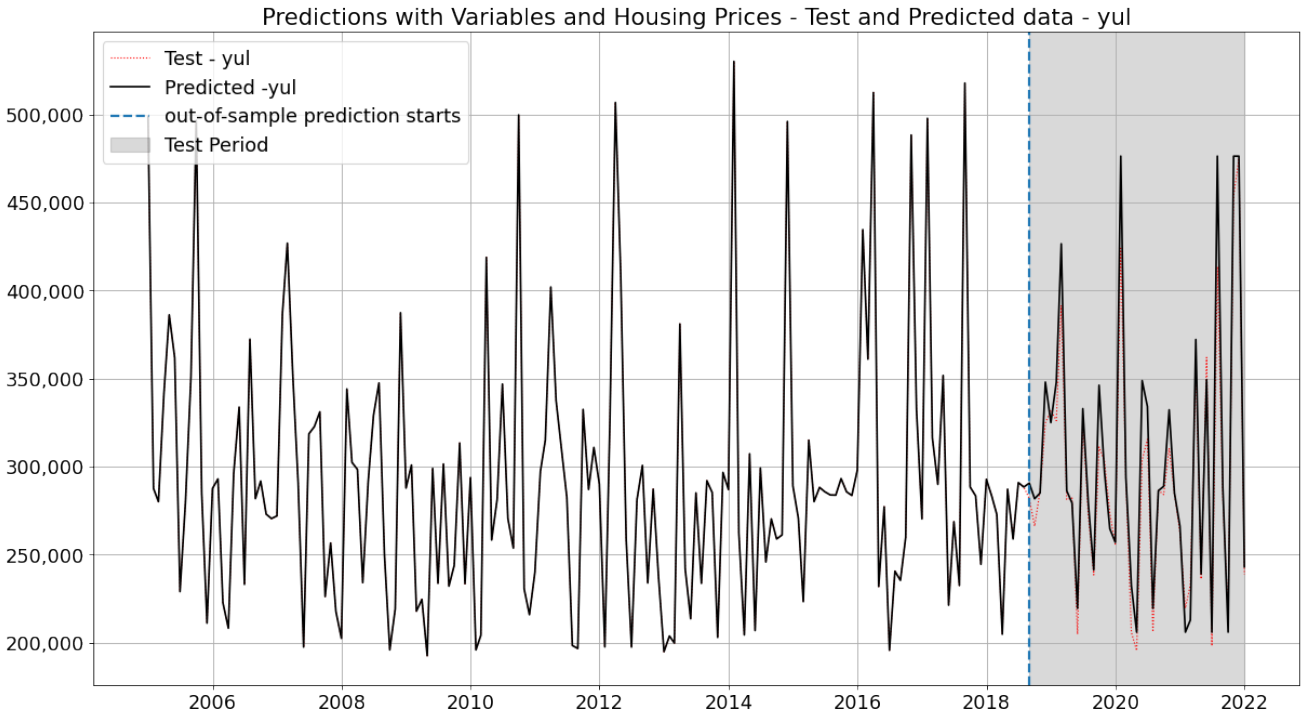


Figure B.7: Montreal - Out of Sample



* Note: YUL is Montreal respectively

Figure B.8: Quebec City - Out of Sample

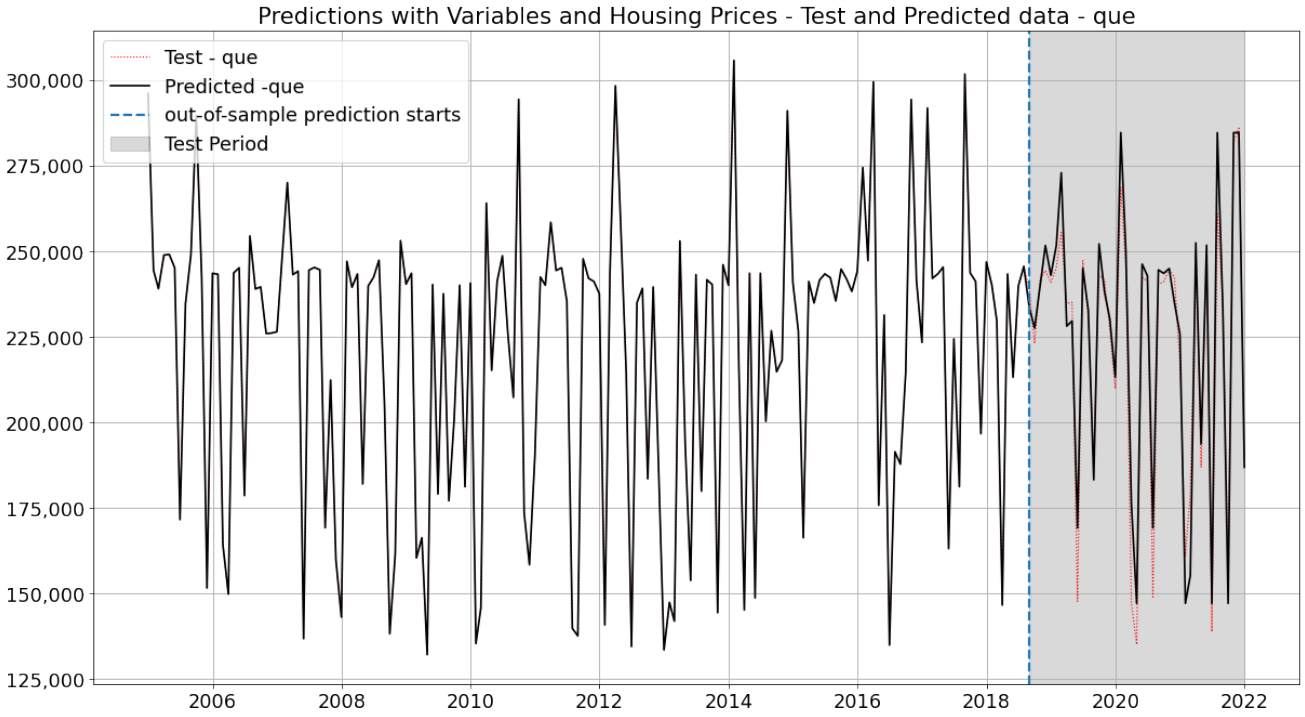
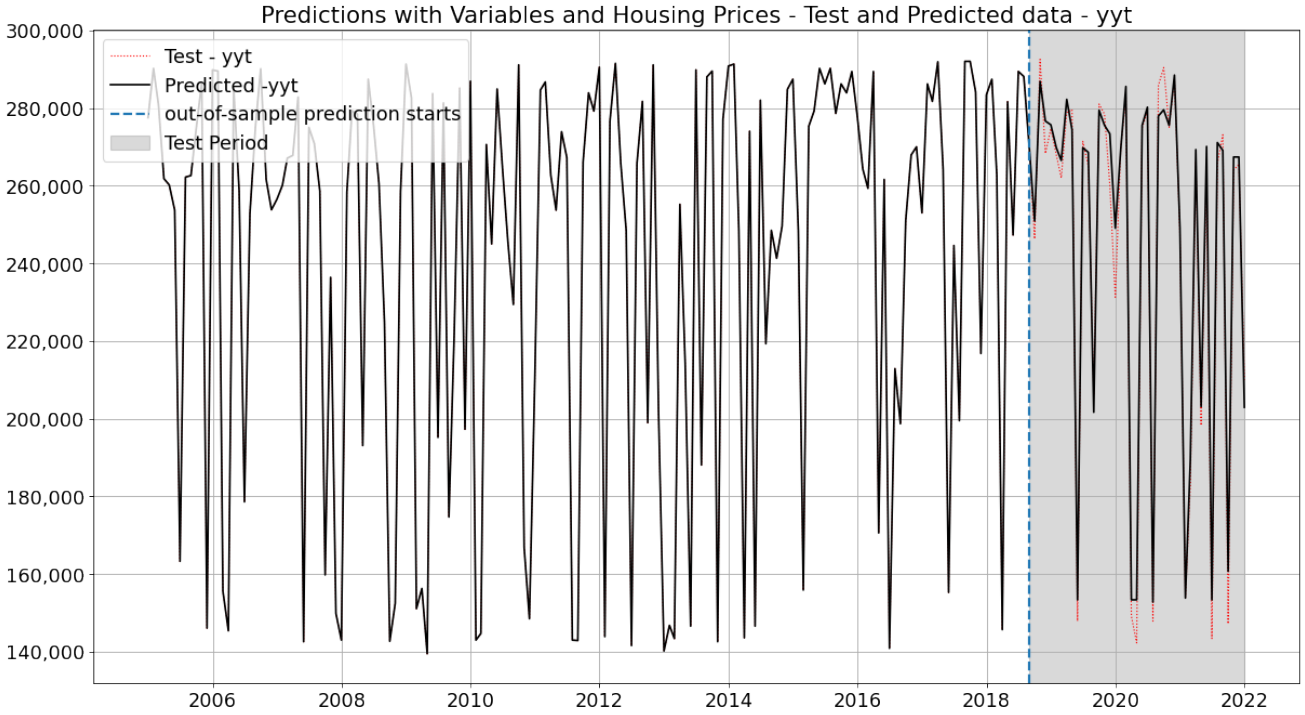


Figure B.9: St. John's - Out of Sample



* Note: YYT is St. John's respectively

C XGBoost: Variable Importance

Feature importance for all cities considered is shown in Appendix C Below. As previously mentioned in the section above, the default setting in Python was used. The values to calculate variable importance are shown in Appendix A.

Figure C.1: Victoria - Variable Importance

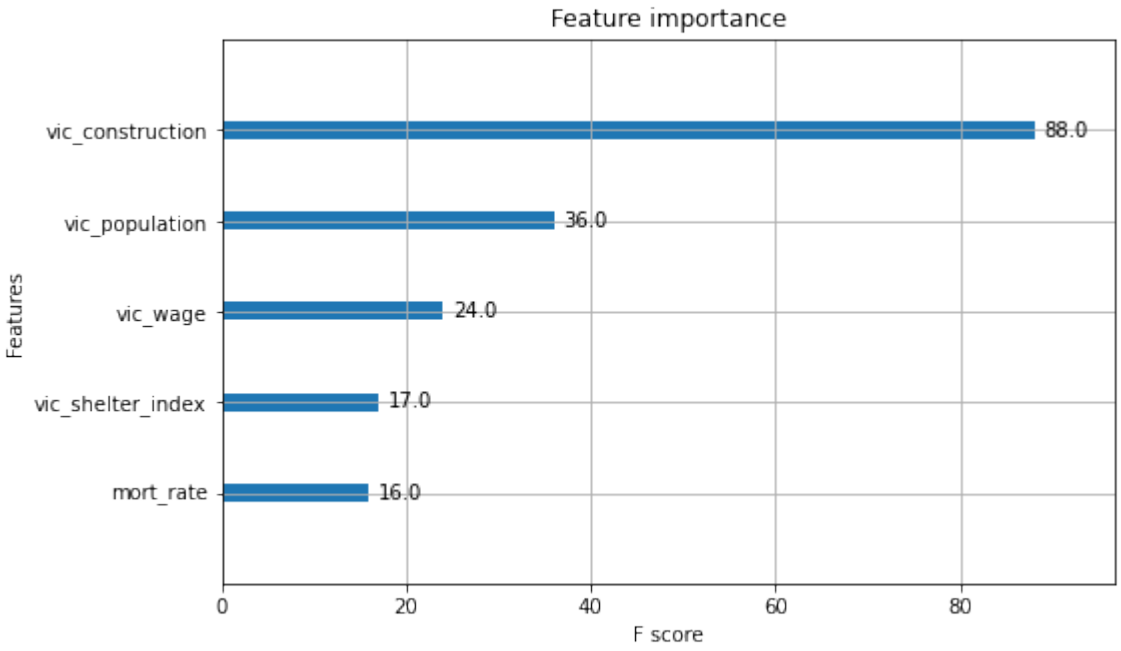


Figure C.2: Greater Vancouver Area - Variable Importance

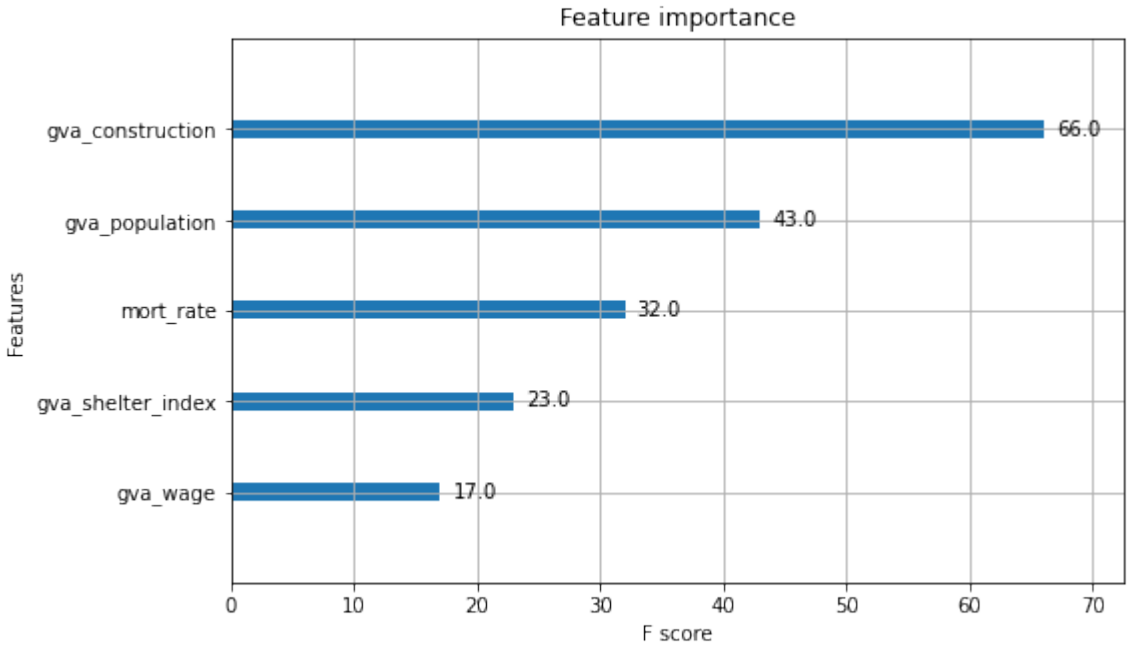


Figure C.3: Edmonton - Variable Importance

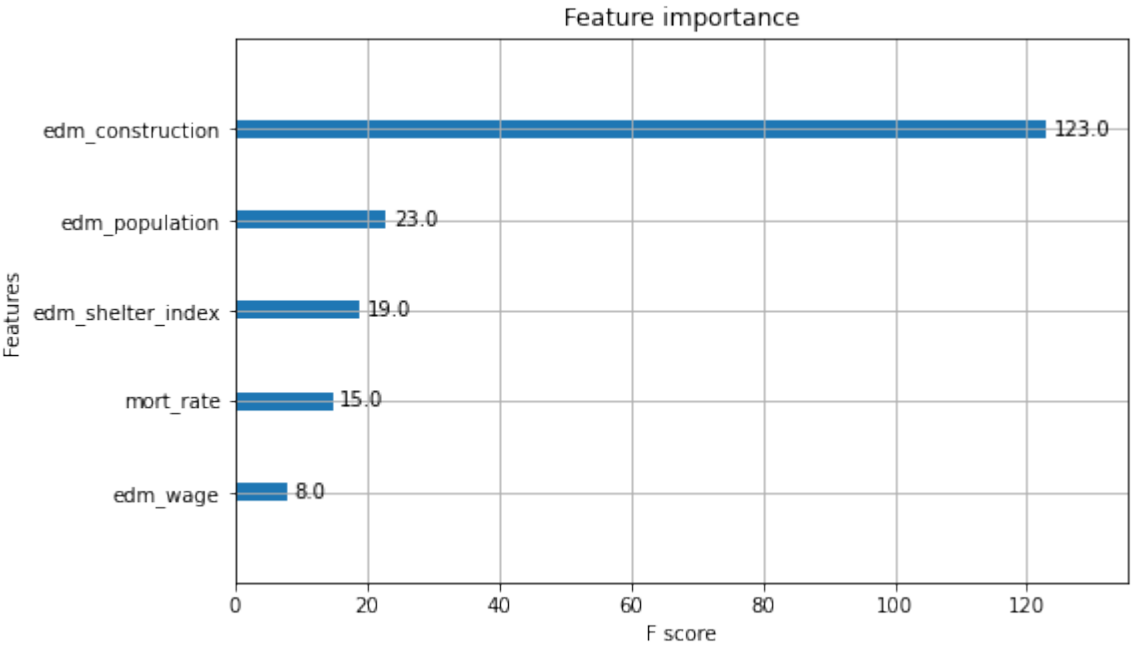
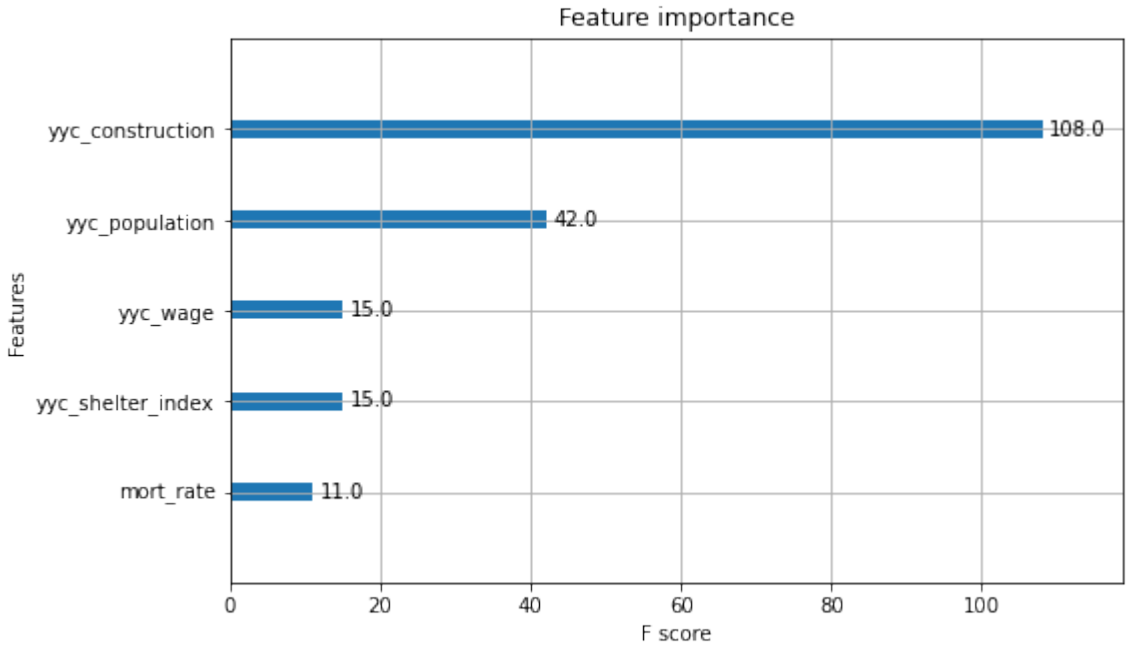


Figure C.4: Calgary - Variable Importance



*Note: YYC is Calgary respectively

Figure C.5: Greater Toronto Area - Variable Importance

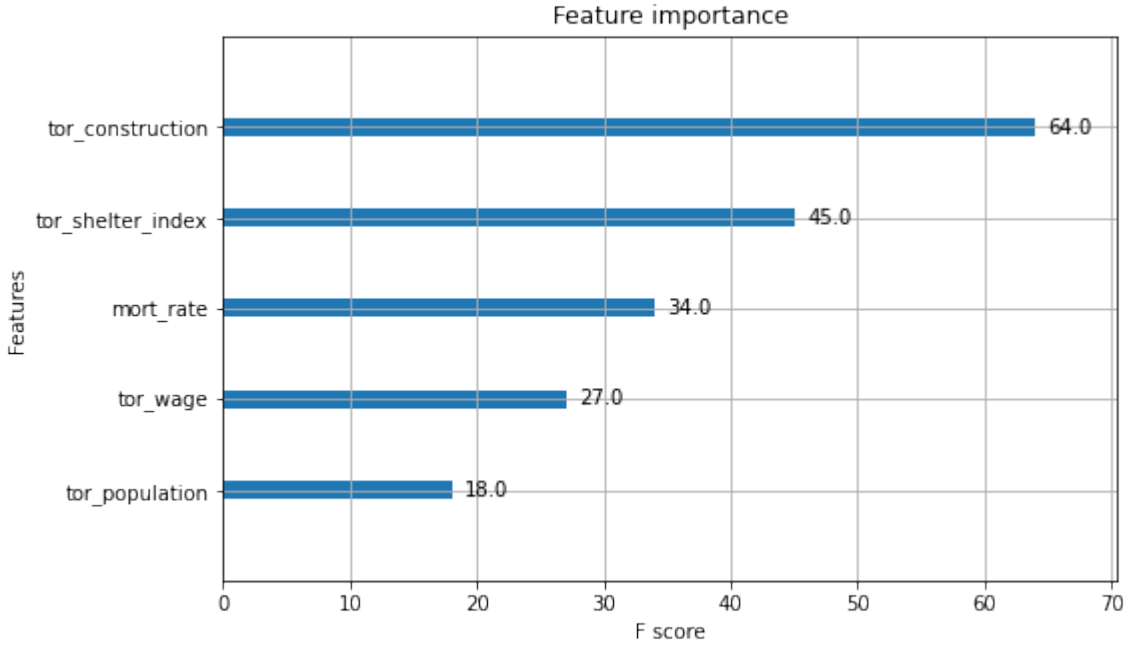


Figure C.6: Ottawa - Variable Importance

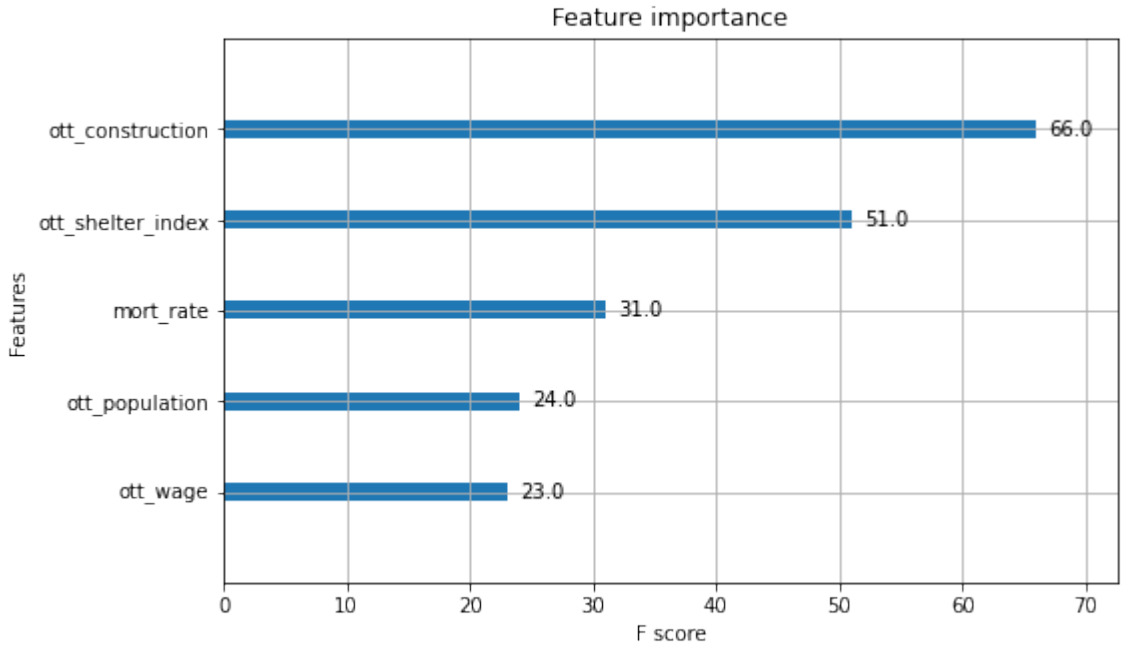


Figure C.7: Montreal - Variable Importance

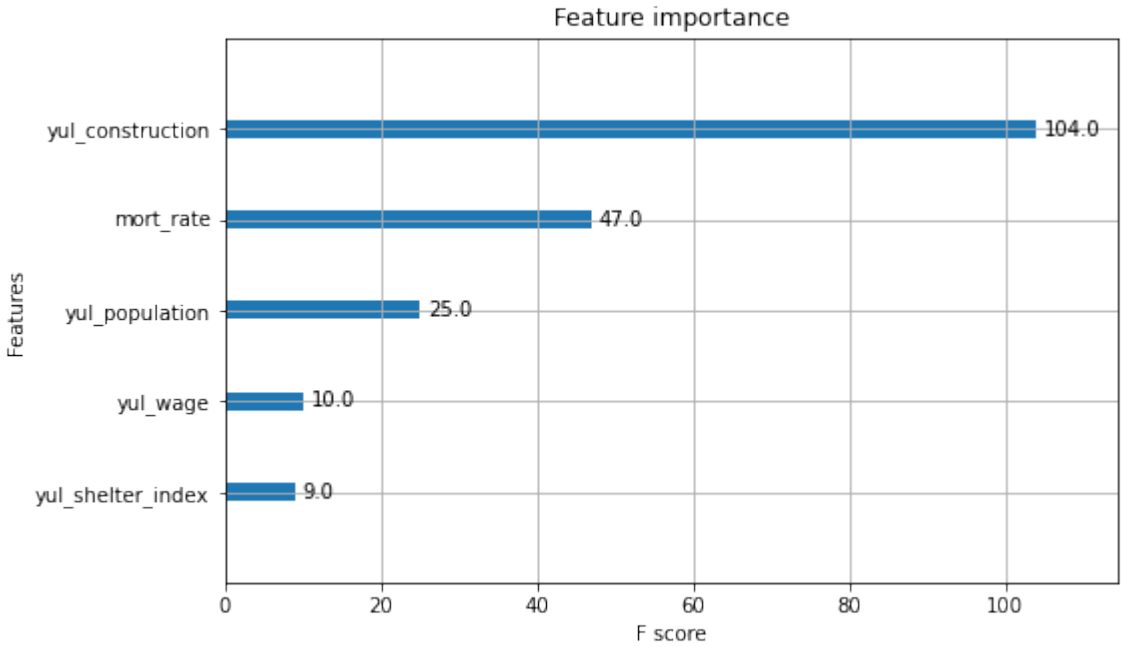


Figure C.8: Quebec City - Variable Importance

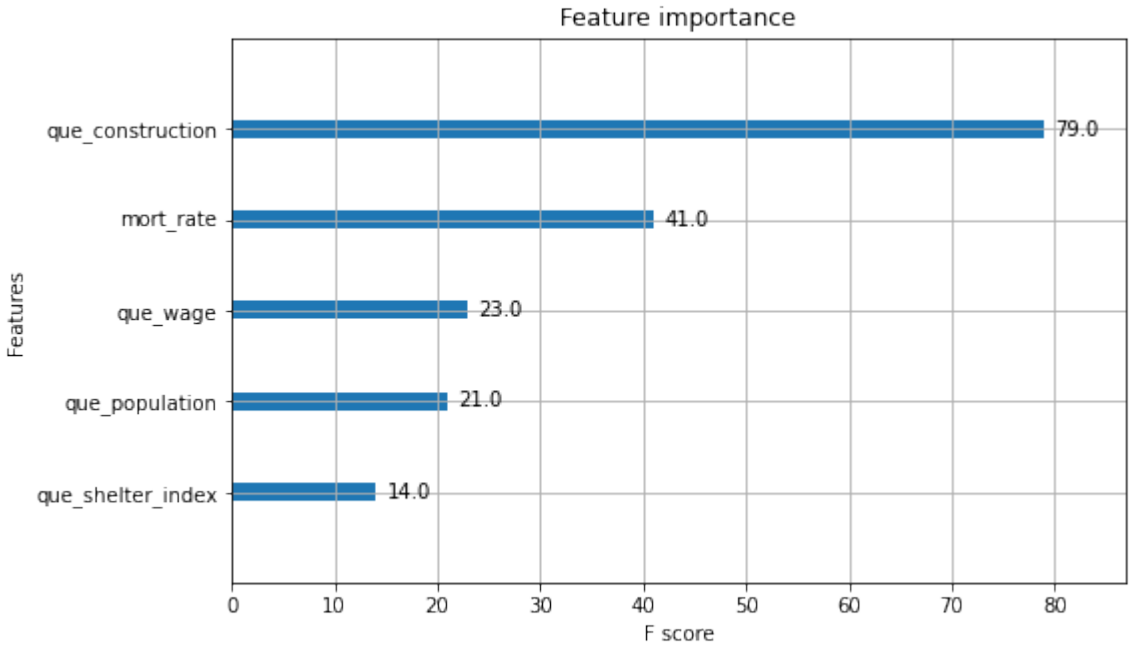
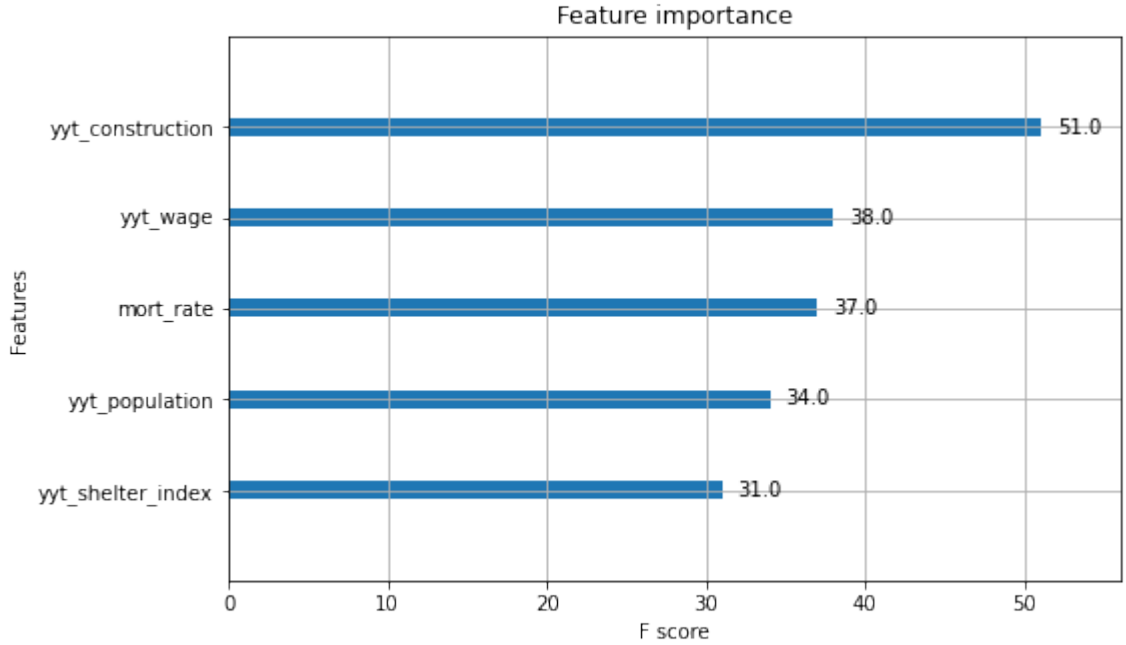


Figure C.9: St. John's - Variable Importance



*Note: YYT is St. John's respectively

D XGBoost Decision Trees

In Appendix D, below are the graphed tree from XGBoost. This is done in Python with the function plot_tree(). The data used to graph the trees are shown in Appendix A, which are the values from the macroeconomic variables used for the estimations.

Figure D.1: Victoria - XGBoost Tree

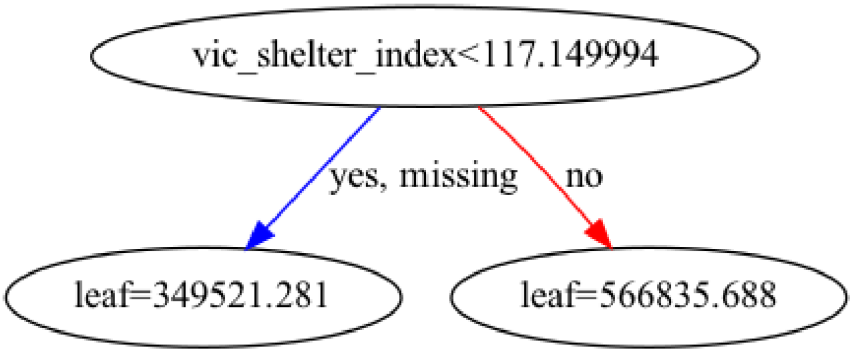


Figure D.2: Greater Vancouver Area - XGBoost Tree

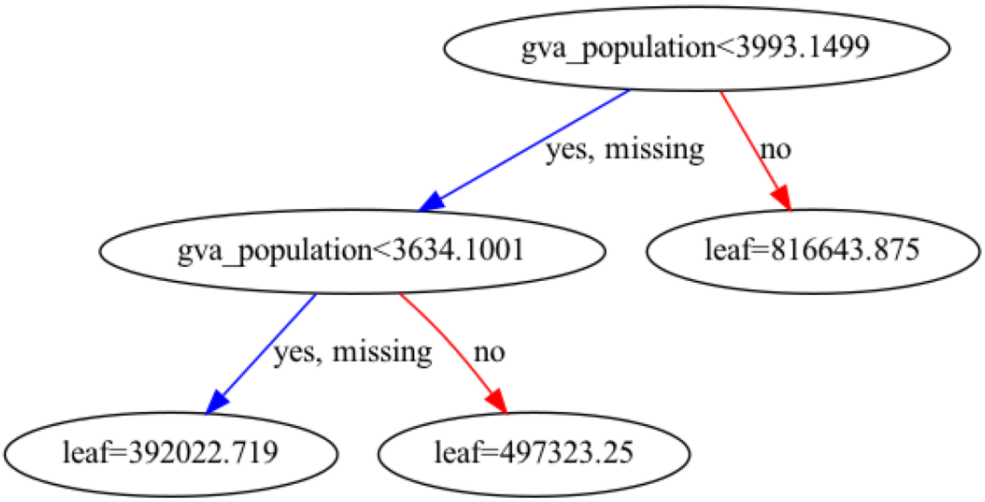


Figure D.3: Edmonton - XGBoost Tree

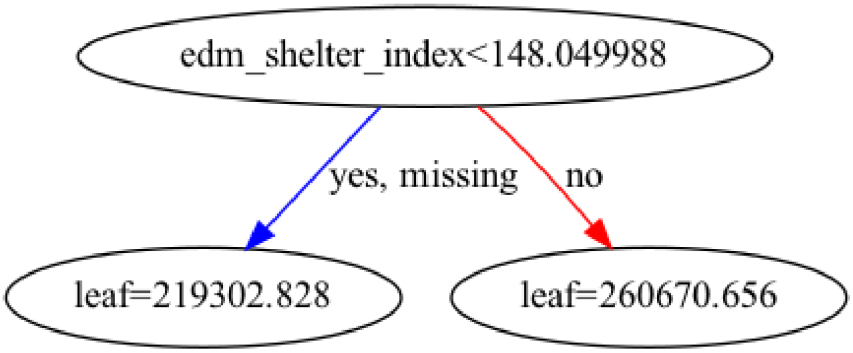


Figure D.4: Calgary - XGBoost Tree

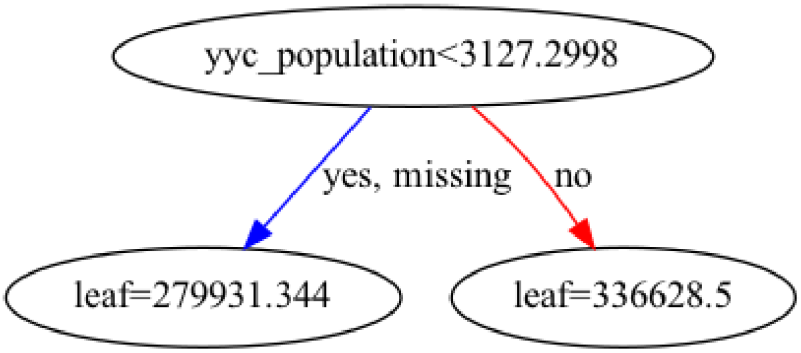


Figure D.5: Greater Toronto Area - XGBoost Tree

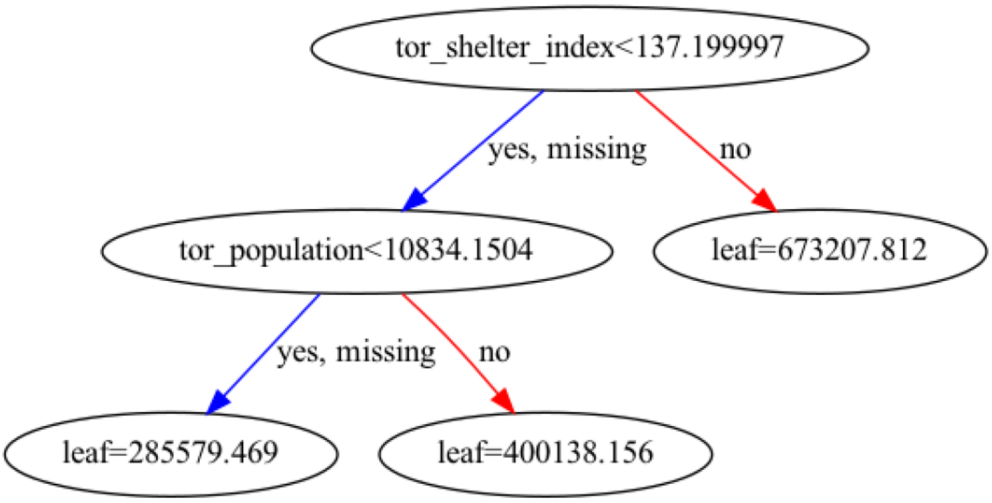


Figure D.6: Ottawa - XGBoost Tree

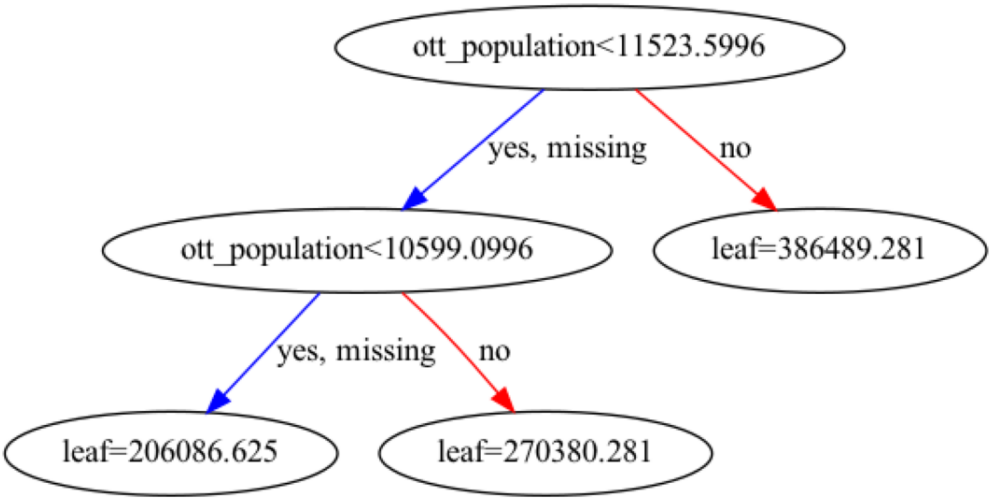


Figure D.7: Montreal - XGBoost Tree

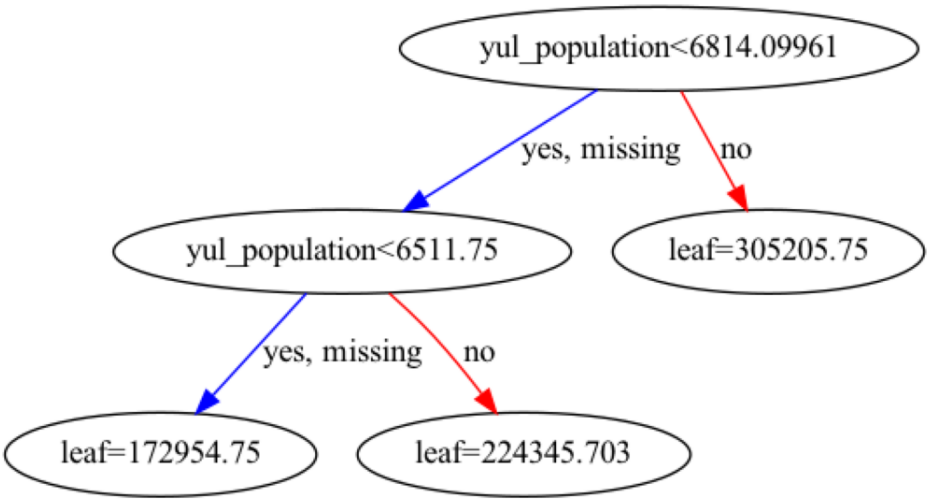


Figure D.8: Quebec City - XGBoost Tree

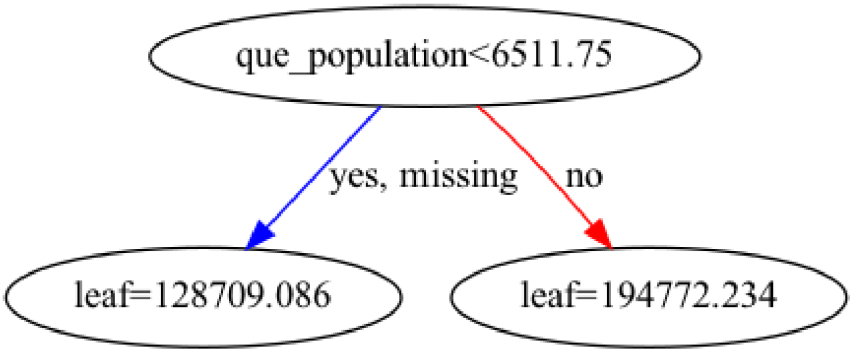
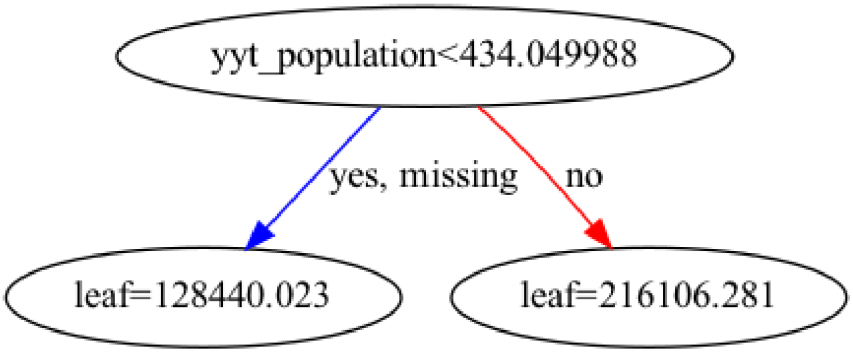


Figure D.9: St. John's - XGBoost Tree



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway