Smith SCHOOL OF BUSINESS Queens University | Master of Management in Artificial Intelligence

MMAI 823

AI in Finance

Professor Matt Thompson

Predicting the Single-Family Home Price Index for Greater Toronto

Final Project

Team Alfred

June 30th, 2021

Student Name	Student Number
Kelsey Pericak	10006993
Mohammad Raza	10134906
Nezil Gopinadh	20247162
Shayaan Mehdi	20253255
Shilpa Vishwanath	20253142
Sky Tang	20251377
Theebak Sothilingam	20241245
Yvan Large	20246196

Table of Contents

1.0 Executive Summary

Team Alfred built a Bayesian Regression that predicts the home price index for single-family homes in greater Toronto. Using hypotheses about the housing market, we began by collecting relevant data that was available online for free. Then, we explored the data by evaluating correlations, generating growth rates, and conducting bivariate analyses. The data was cleaned by detecting and either removing or adjusting outliers, and it was modeled using three linear regression methods with no feature selection. To evolve the prediction's performance, feature selection occurred and brought the dataset down to 9 features from 35. Following hyper-tuning, the Bayesian model's R^2 performance (our best model) increased by 0.24 to 0.984. Our final model accurately predicted the change in single-family home prices by leveraging data about the bank rate, construction starts and completions, unemployment, temperature, the month, and self-lagged data by one month and one year. Though initially developed via curiosity by the team, our model could be used by Toronto residents, policy makers, banks, developers, and real estate agents (to name a few) for purchasing, selling, and other money generating purposes.

2.0 Business Problem

Team Alfred found an opportunity to create a financial model that predicts future real-estate prices. Many members of our team have recently purchased, sold, or considered real estate investments. Curiosity led us to seek the reasoning behind real estate price fluctuations in Greater Toronto, where most of us live. Such insights could guide our future purchasing decisions. Beyond the advice from real-estate agents, there was not a commonly known source for which we could rely on to receive future investment advice. Our goal was to use free, public, and historical data to predict the change in home prices for single-family homes one month in advance.

The single-family home price index was selected as our target variable. This index was created by the Canadian Real Estate Association (CREA). The CREA allows anyone to export its data after accepting their conditions to not use said information for commercial purposes[1]. It is calculated using a complex formula that considers several variables such as the number of rooms in a home, the price of a home and the square footage of a property. The growth in price for a

[1] *MLS® HPI data – Terms of Use*. CREA. https://www.crea.ca/hpi-tools-terms-of-use/.

typical single-family home in greater Toronto is captured with the HPI. Single-family homes are defined by the CREA as one or multiple storey homes with a kitchen and living room on the main floor and if two-storeys or more then bedrooms on the higher floors. This grouping encompasses a wide array of residential types: "back split, bi-level, bungalow, hillside bungalow, hillside split, 2-storey split and 3 level split" if one floor, and "4 level split, 5 level split, one-and-a-half storey, two-storey, two-and-a-half storey, and three-storey" if multiple floors[2]. The index has a base value of 100 which was set in January 2005. One unit of the index can be explained as one percentage from the base index price[3].

3.0 Approach

The below CRISP-DM steps (excluding deployment) were taken to accomplish this project. Notice the data preparation step. Following dataset creation and exploratory analysis, we were careful to split our dataset using a time sensitive approach so that no data leakage would occur. Test data was unseen during the training stage, and it occurred following October 2018. Training data was capped to September 2018 for modeling. Features were lagged as a secondary precaution. More about the features in section 4.

- Data
    - Dataset creation
    - Exploratory analysis and understanding
    - Data preparation
        - Cleaning
        - Time sensitive split
            - Train (85%) from January 2006 to September 2018
            - Test (15%) split from October 2018
- Modeling
    - Base model creation and evaluation
    - Feature selection and hyper parameter tuning
    - Prediction

[2] *Property Type Definitions*. CREA. https://www.crea.ca/wp-content/uploads/2016/02/benchmark_home_definitions_for_tableau_en.pdf .

[3] *MLS® Home Price Index Methodology*. CREA. https://www.crea.ca/wp-content/uploads/2016/07/HPI_Methodology.pdf.

4.0 Dataset Creation

4.1 Methodology

To build our dataset, we identified potential drivers of the HPI index. For each hypothesis, we conducted a web search to find relevant data. We then manually exported the data if it was freely available. Finally, each datapoint was aggregated on a monthly basis since we were predicting a monthly home price index. Multiple data sources were leveraged, with Statistics Canada as the main, trusted dataset provider. Whenever possible, we exported data for Toronto. If Toronto data was unavailable then we would try to find data about Ontario, and if Ontario data was unavailable then we would check if national data could be utilized. In total, the team was able to create 35 features with lags of both one month and one year against the target month.

4.2 Data

The following data were added to our baseline dataset. See **Appendix A** for a response to each hypothesis mentioned in this section.

4.2.1 Home Price Index (HPI)

The single-family home price indices from the CREA of other large cities including Calgary, Vancouver and Montreal were added to our feature set in hopes of finding a lagged relationship between the housing prices throughout Canada. We hypothesized that the growth in Toronto's housing price index might be correlated with the growth of Vancouver's housing price index. For other cities, we were mainly inquisitive.

4.2.2 Consumer Price Index (CPI)

Cost of living and inflation are captured by the consumer price indices created by Statistics Canada. We chose to include two price indices in our dataset. The CPI for shelter was the first. The CPI for shelter encompasses the cost of rent by taking into consideration renovations and amenities. Home ownership costs that are considered by the index include interest, mortgages, maintenance costs, and other accommodation expenses[4]. We hypothesized that this CPI from Statistics Canada would have a high correlation with the HPI from the CREA given that both

[4] Government of Canada, S. C. Tracking the cost of shelter. Government of Canada, Statistics Canada. https://www.statcan.gc.ca/eng/blog/cs/shelter-cost.

utilize prices for products and services within the same industry. The second index we included was the CPI for all goods and services which represents inflation in Canada as a whole.

4.2.3 Seasonality

To account for seasonality, we added a column that shows the month number. Month number was calculated with a simple python function. We hypothesized that seasonality might exist in the data after visualizing the line graph of HPI evolution over time.

4.2.4 Weather

Extreme conditions like snow may have an impact on people's willingness to move downtown. Winter traffic and limited outdoor activities (ex. skiing and snow shoeing) are available downtown from November to February. We also figured that people would not want to move their belongings in the winter due to poor road conditions caused by snow and ice. Imagine trekking through 20 centimeters of snow to place your heavy couch into your moving truck. These circumstances could decrease demand and potentially prices. To gather weather data, we calculated monthly average snowfall in centimeters from daily statistics and we did the same for temperature. The website called Toronto.weatherstats.ca was used to export this data[5].

4.2.5 Gross Domestic Product (GDP)

As stated in Canada's National Observer website within an article titled "Canada's housing market is on fire — and headed for disaster", housing makes up roughly 10% of Canada's GDP[6]. The GDP exemplifies Canada's economic output and captures our overall standard of living. We extracted this information from Statistics Canada by using basic prices for the real estate industry on a monthly basis. Given the inability to choose non- adjusted data to maintain variance, we chose to use the trading-day adjusted data with 2012 constant prices[7]. We hypothesized a positive correlation between GDP and the target.

4.2.6 Retail Sales

Retail sales provide a consumer spending perspective that could aide in understanding consumer behavior around preparing houses to sell. We extracted this information from Statistics

[5] *Number of Days of Snow - Monthly data for Toronto*. Amateur Weather Statistics for Toronto, Ontario. https://toronto.weatherstats.ca/charts/count_snow-monthly.html.

[6] Fawcett, M. (2021, March 15). Canada's housing market is on fire - and headed for disaster. Canada's National Observer. https://www.nationalobserver.com/2021/03/15/opinion/canadas-housing-market-fire-and-headed-disaster.

[7] Government of Canada, Statistics Canada. (2021, June 1). Gross domestic product (GDP) at basic prices, by industry, monthly. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3610043401.

Canada for Toronto specifically[8]. Retail sales cover a wide variety of merchandise such as motor vehicles, furniture and furnishings, groceries, health, jewelry, and miscellaneous. No concrete hypotheses were made; we were mainly interested in how spending occurs holistically.

4.2.7 Construction Starts and Completions

We hypothesized that the construction commencements and completions of Toronto dwellings could be good indicators of real estate prices given that newer homes typically cost more, and that more construction starts could be a sign of population growth or housing demand. This data was made available by Statistics Canada[9].

4.2.8 Population and Income

Toronto's yearly population data was retrieved from a dataset maintained by the United Nations. The data was then broken down into months by linearly extrapolating the yearly growth rate[10] [11]. Population was hypothesized to have a relationship with construction. Median annual income for both couple families and lone parent families were also added as features to assess how Toronto's average wealth related to real estate prices[12]. Were citizens becoming wealthier and driving the prices of real estate higher? Though data was limited, we considered the emergence of technology companies in Toronto and the potential increase of jobs with higher salaries. Recent data was missing for income from our source, so we imputed the last three months of 2020 with SimpleImputer by scikit-learn.

4.2.9 Bank Rate

The Bank of Canada bank (lending) rate impacts the interest on mortgages that consumers can receive from banks. An increase in the lending rate will typically increase the monthly payments[13] that a homeowner is required to make. We hypothesized that a lowering bank rate could

[8]Government of Canada, Statistics Canada. (2021, June 23). Retail trade sales by industry. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2010000802&pickMembers%5B0%5D=1.9&cubeTimeFrame.startMonth=11&cubeTimeFrame.startYear=2020&cubeTimeFrame.endMonth=03&cubeTimeFrame.endYear=2021&referencePeriods=20201101%2C20210301.

[9] Government of Canada, Statistics Canada. (2021, June 16). Canada Mortgage and Housing Corporation, housing starts, under construction and completions in centres 10,000 and over, Canada, provinces, selected census metropolitan areas. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410014301.

[10] United Nations. World Population Prospects - Population Division. United Nations. https://population.un.org/wpp/.

[11] Toronto, Canada Metro Area Population 1950-2021. MacroTrends. https://www.macrotrends.net/cities/20402/toronto/population.

[12] Government of Canada, Statistics Canada. (2020, September 2). Distribution of total income by census family type and age of older partner, parent or individual. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1110001201.

[13] *What's behind your mortgage rate*. Bank of Canada. https://www.bankofcanada.ca/2020/05/whats-behind-your-mortgage-rate/.

increase demand for mortgages, and vice versa. This data was sourced from Statistics Canada[14]; however, it was also available on the Bank of Canada website.

4.2.9 Unemployment Rate

We hypothesized that unemployment could have a positive or negative impact on the housing market by either increasing the wealth gap or decreasing spending capacity and therefore prices. This data was sourced from Statistics Canada and filtered for Toronto with population ranging from 25 to 54 years old[15].

4.2.10 University Enrollment

The university enrollment data for Ontario residents was retrieved from Statistics Canada[16]. Increased university enrollment may correlate with increased earnings potential as well as an increased understanding of investment in real estate, which may drive up prices[17]. We looked at enrolment for all degrees.

4.2.11 Median Age

The Median age of Ontario residents was retrieved from Statistics Canada, too[18]. We pondered if increasing population ages would impact the volume of homes purchased as an aging population may result in more elderly selling homes.

---

[14]Government of Canada, Statistics Canada. (2021, May 28). Financial market statistics, last Wednesday unless otherwise stated, Bank of Canada. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1010012201&cubeTimeFrame.startMonth=01&cubeTimeFrame.startYear=2004&cubeTimeFrame.endMonth=04&cubeTimeFrame.endYear=2021&referencePeriods=20040101%2C20210401.

[15] Government of Canada, Statistics Canada. (2021, June 4). Labour force characteristics by immigrant status, three-month moving average, unadjusted for seasonality. https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?Pid=1410008201.

[16] Government of Canada, Statistics Canada. (2020, November 25). Postsecondary enrolments, by registration status, institution type, status of student in Canada and gender. https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?Pid=3710001801#timeframe.

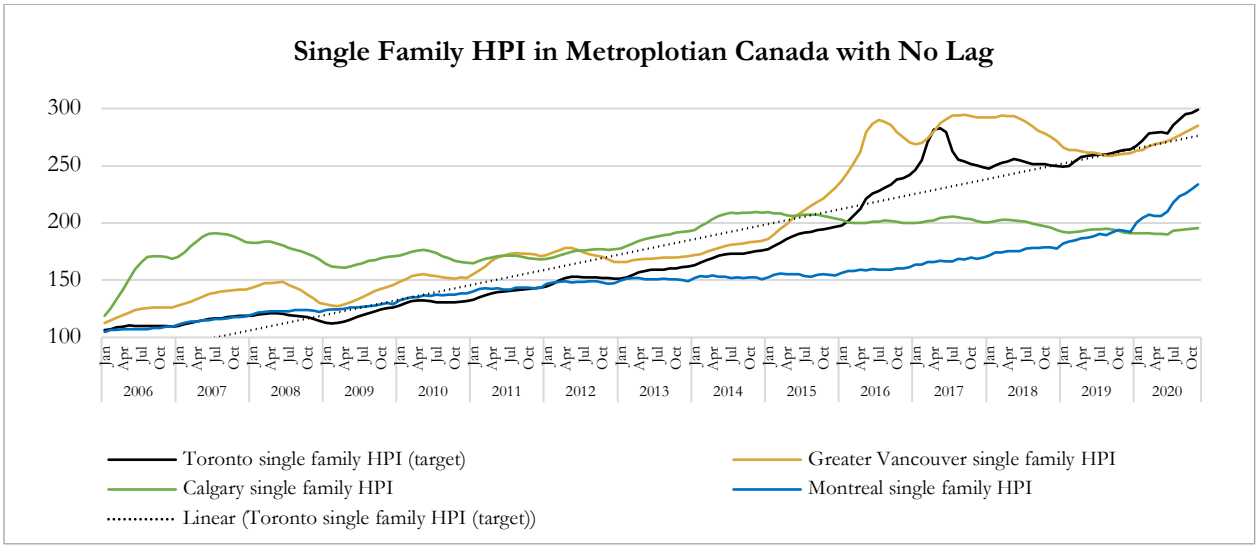
[17] Government of Canada, Statistics Canada. (2020, November 25). Postsecondary enrolments, by field of study, registration status, program type, credential type and gender. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?Pid=3710001101&pickmembers%5B0%5D=1.7&pickmembers%5B1%5D=2.2&pickmembers%5B2%5D=6.2&cubetimeframe.startyear=2003%2B%2F%2B2004&cubetimeframe.endyear=2018%2B%2F%2B2019&referenceperiods=20030101%2C20180101.

[18]Government of Canada, Statistics Canada. (2020, September 29). Population estimates on July 1st, by age and sex. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?Pid=1710000501&pickmembers%5B0%5D=1.7&pickmembers%5B1%5D=2.1&cubetimeframe.startyear=2005&cubetimeframe.endyear=2020&referenceperiods=20050101%2C20200101.

5.0 Exploratory Analysis and Data Preparation

Exploratory data analysis was performed on the HPI and features in our dataset. A correlation chart was created for the entire dataset. We observed that 19 features (all lagged by one month or one year) had a correlation above 0.9 with the target. The top two most interesting correlations were the metropolitan Ontario CPI for shelter and the retail sales for Toronto. These positive correlations taught us that spending is a comprehensive activity that can be observed across rental, ownership and all other shopping purchases. The top negative correlation was the 1 year lagged bank rate at -0.48, and (however low) unemployment rate was also negatively correlated. See **Appendix C** for a visual representation of the bank rate decreasing as HPI increases in 2020. It is also important to note that a pandemic occurred in 2020, and that we do not assume a causational relationship between bank rate and HPI.

Figure 1: Historical time series of HPI calculated by CREA



As seen in Figure 1, the single-family HPI for greater Toronto increased by 1.5X from 2015 to 2020. Moreover, when zoomed in, we noticed troughs or stability during winter months such as November and December (consistent with our hypotheses). The housing market in Vancouver exhibited a similar upward trend in HPI to that of greater Toronto. Montreal was steady until the later parts of 2017 as it began the upward trend and started to close the gap nearing Vancouver and Toronto. Contrast to Calgary, the housing price index did not correlate as well with the other cities and remained steady throughout. In addition, two significant peaks were noticed

for Vancouver and Toronto. Those peaks are often referred to as “housing bubbles”, and they occurred because of foreign tax regulations which impacted foreign purchases of homes[19]. Based on this, we decided to drop the Calgary housing price index and the bubble time range.

Months during the housing bubble were also noticed in the normality testing. Figure 2 presents a scatter plot of predictions versus residuals on training data as well as a histogram for those residuals. This visualization was made to ensure our data would be suitable for linear modeling. Notice the change is graphs once the outliers detected in our residual plot were removed.

Figure 2: Normality testing

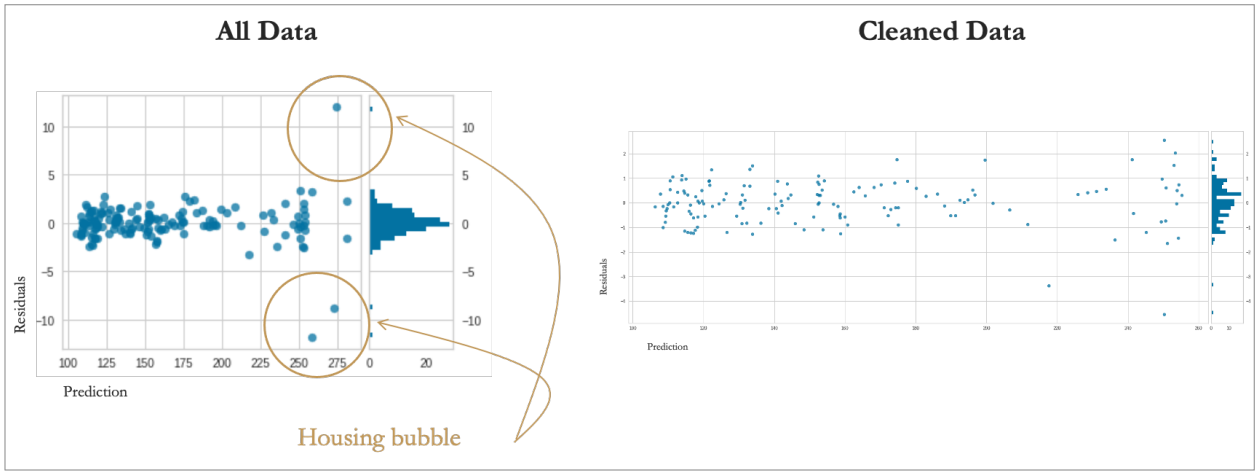
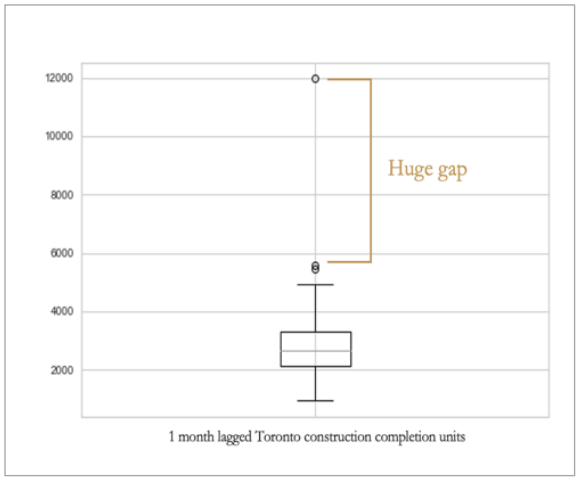


Figure 3: Boxplot



A boxplot analysis was performed to detect other outliers within the feature set. We reviewed the boxplot graphs for every feature and noticed significant outliers for Toronto construction completion units. To not skew any normalized data with this high maximum value, we capped the construction units to six thousand. All other features were deemed reasonable enough for modeling. See **Appendix D** for examples.

After exploring and pre-processing, the data was ready for modeled.

[19] Pasalis, J. Special Report: Lessons Learned From Toronto's 2017 Real Estate Bubble. Move Smartly. https://www.movesmartly.com/lessons-learned-from-toronto-2017-real-estate-bubble

6.0 Modeling

6.1 Evaluation

All code and hyper-tuning results can be found at this public GitHub repository. We started by using all aforementioned features to create base linear models for the regression. The key metrics for our use case were R^2 on test data and the RMSE on test data. R^2 measures the variance of a dependent variable explained by the independent variables in the model[20], and the root mean square error (RMSE) is the standard deviation of the residuals. Both metrics were introduced to us in the AI in Marketing course as suitable performance measurements for regression modeling.

Table 1: Base model performance

Metric	ARD	Linear Regression	Bayesian
R^2 on training data	0.99	0.99	0.99
R^2 on test data	0.97	0.76	0.74
RMSE on test data	2.76	8.0	8.43

Our best base model was the ARD regression followed by the linear regression and then the Bayesian regression. ARD (Automatic Relevance Determination) regression fits the data with a Bayesian Ridge regression. By shifting the coefficient weights slightly towards zero, it stabilizes them compared to the OLS (ordinary least squares) estimator[21]. Linear regression and Bayesian regression were already introduced to us in the Machine Learning and AI Technology course. Linear regression is a model in which an output variable can be determined using a linear combination of inputs variables. While Bayesian regression predicts by making use of regression weight possibilities together with their weighted posterior probabilities.

Due to these performances (refer to Table 1 above), we first tuned the ARD regression but learned that only two features were selected (the rest had coefficients of 0). Those features were one-month self-lagged data and the temperature lagged by one month. This was too simplistic to

[20] Fernando, J. (2021, May 19). R-Squared. https://www.investopedia.com/terms/r/r-squared.asp#:~:text=R%2Dsquared%20(R2),variables%20in%20a%20regression%20model.&text=It%20may%20also%20be%20known%20as%20the%20coefficient%20of%20determination.

[21] Automatic Relevance Determination Regression (ARD) — scikit-learn 0.24.2 documentation. Scikit-Learn. https://scikit-learn.org/stable/auto_examples/linear_model/plot_ard.html

describe the overall picture, so we moved onto the next models. The linear regression and Bayesian regression had the same tuned performance however the Bayesian regression enabled us to tune more hyperparameters. We deemed the model superior given its ability to remain accurate if the data were to drift in the future. Linear regression with sklearn only has the fit_intercept to tune. See below for a summary:

Table 2: Tuning details

ARD	Linear Regression	Bayesian
- Eliminated - Only 2 features with non-zero coefficients - 0.980 R^2 - Skeptic of performance during uncommon economic shifts	- Eliminated - Only one hyperparameter - Gave coefficients for all features selected - High performance - 0.984 R^2 on test	- Selected - Many hyperparameters - Gave coefficients for all features selected - High performance - 0.984 R^2 on test

6.2 More About the Bayesian Model's Evolution

The baseline Bayesian Ridge Regression model was trained from Jan 2006 with all 35 features and no hyperparameter tuning. We then performed feature selection that narrowed down 35 features to 9 features. Feature selection was conducted by looking at the individual RMSE for each feature with the target variable, and by removing highly correlated variables above 0.4 or below -0.6. Pearson correlations of 0.4 and -0.6 were selected by looking at the full correlation chart and identifying reasonable cut-offs. This eliminated any multicollinearity and ensured nicely distributed features with enough variance to have an impact on the predictions. See **Appendices E** and **F**.

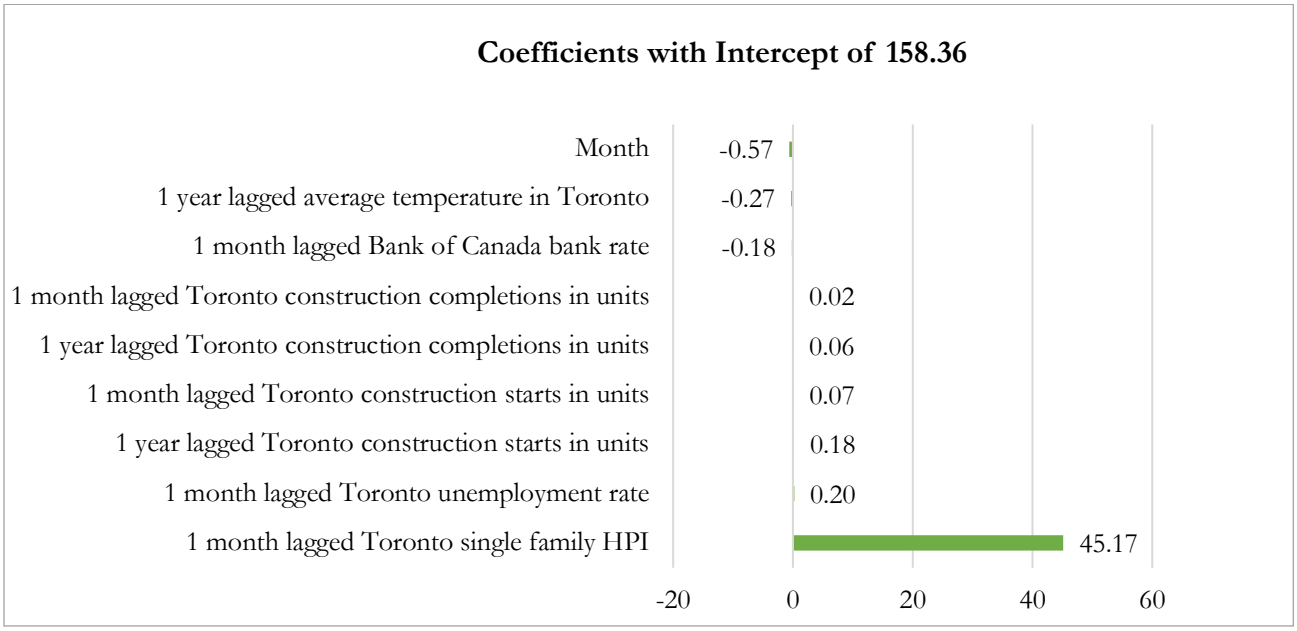
We then performed hyperparameter tuning on those 9 features and obtained the best values for intercept as True, alpha 1 (which is the shape parameter of gamma distribution over alpha parameter) as 10 and alpha 2 (which is the inverse scale parameter) as 0.001[22]. We forced the number of iterations to 1,000 so that the model would learn the optimal parameters and achieve good performance. We also ensured that the data was normalized by forcing the normalize parameter to True. With this configuration, we arrived at an R^2 of 0.984 on test and a Root Mean Square Error of 2.08, which we think is very reasonable.

[22] *sklearn.linear_model.BayesianRidge*. scikit. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html.

6.3 Parsimonious Model

The model is parsimonious because it performs well with limited features, and it has explainable coefficients. We noticed that the R^2 and RMSE values remained consistent irrespective of whether we standardized the features or not. Upon standardization (for explainable AI with the coefficients), the parsimonious equation showed the highest negative coefficients for the prediction month at –0.57, the 1 year lagged average temperature in Toronto at –0.27, and 1 month lagged Bank of Canada rate at –0.18. We saw the highest positive coefficients for 1 month lagged Toronto single-family HPI at 45.17, and an intercept of 158.36. The other positive coefficients had much lower values ranging from 0.02 for the 1 month lagged Toronto construction completions and 0.20 for the 1 month lagged Toronto unemployment rate. An interesting observation was that 4 of the 6 positive correlations were related to construction.

Figure 4: Parsimonious Equation (rounded and standardized)

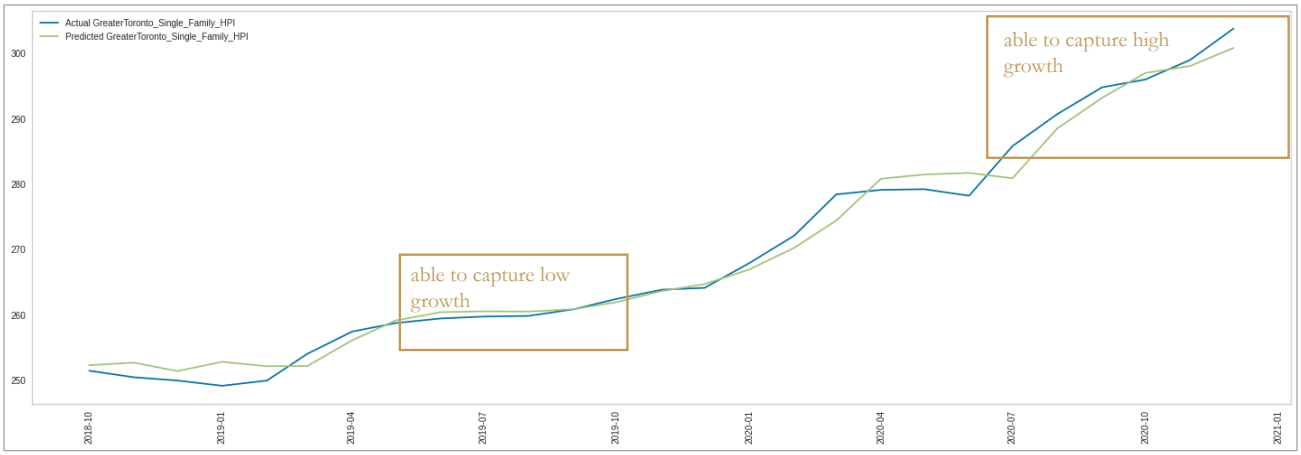


6.4 Model Output on Test Data

We tested our Bayesian model on unseen data from October 2018 onwards and we found that it was good at capturing peaks and valleys such as low growth in the second half of 2019 and even high growth at the end of 2020 during the pandemic. Despite the pandemic's uncommon impact on the economy, our model has proven it has strong predictive power. It was robust and

able to adapt to the various economic circumstances. However, we noticed that the model slightly under-predicted the increase in house prices during the early stages of the pandemic.

Figure 5: Prediction vs Actual results



6.5 Model Prediction

We reported the coefficients of our model on normalized (hyper-parameter) but not standardized data to predict the HPI for May 2021 and arrived at an index value of 350.57, an increase of 2.28 percentage points since April 2020 (347.2). Once the HPI for May has been released we will be able compare the prediction. We gain competitive advantage by predicting this information before it becomes available to the public.

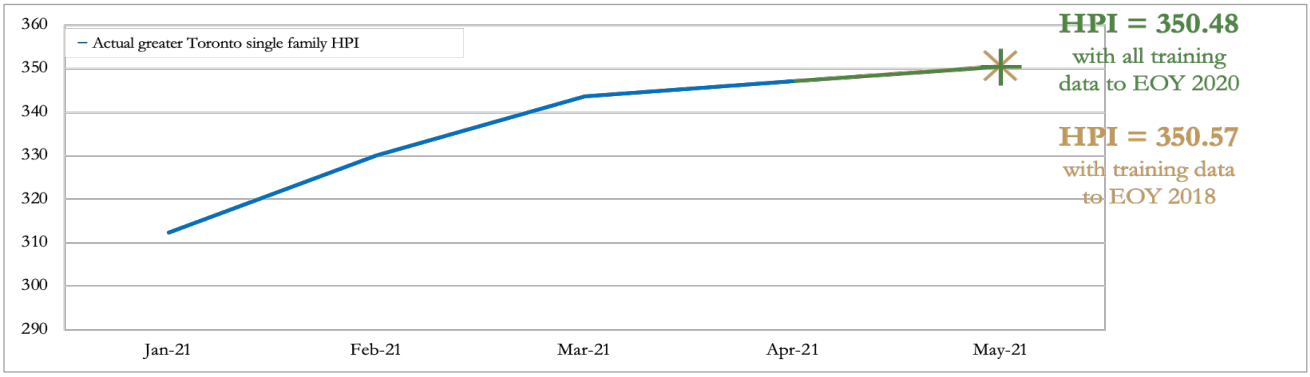
Figure 6: Prediction for May (unreleased) using model on normalized but not standardized data

Feature	Coefficient	Value	Product
Intercept	-1.12	1	-1.12
1 month lagged Toronto single-family HPI	1.01	347.20	350.32
Month	-0.16	5	-0.82
1 month lagged bank of Canada bank rate	-0.13	0.50	-0.07
1 month lagged average temperature in Toronto	-0.03	7.89	-0.22
1 year lagged construction completion units in Toronto	6.71e-05	2,942	0.20
1 month lagged construction starts units in Toronto	7.31e-05	2,802	0.21
1 year lagged construction starts units in Toronto	0.00019	2,558	0.49
1 month lagged construction completion units in Toronto	2.55e-05	2,822	0.08
1 month lagged unemployment in Toronto	0.18	8.50	1.51
			350.57

6.6 Data Drift

To assess data drift, we trained our Bayesian model with all training data from January 2006 until December 2020 and compared the prediction for May 2021 that we got with the model trained from January 2006 until December 2018. We found a miniscule difference in the prediction for May of 350.48, showing there was little data drift. This means that the model would not need to be retrained frequently; however, it should still be validated and monitored to ensure that it continues to capture unexpected changes is real estate prices such as housing bubbles (if ever deployed). As you can see in the graph below, the HPI is very similar across both models.

Figure 7: 2021 Single-family HPI with 2 predictions showing little to no data drift

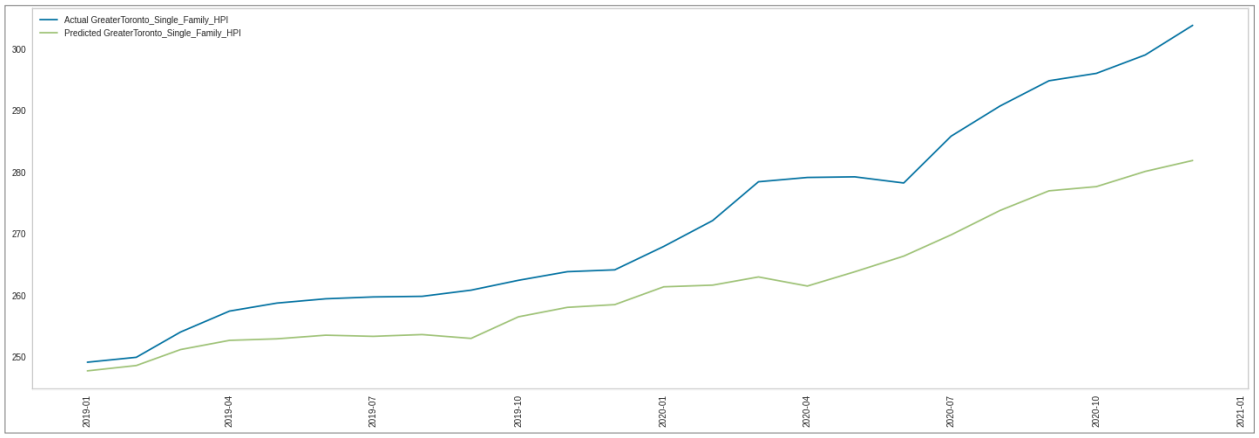


6.7 Model with No Self-Lag and Standardized Data

Finally, we conducted an exploratory experiment to see how our baseline model behaved without self-lag. We wondered if we could leverage data external to the CREA to predict their index. Given the highly correlated features to our target, we felt confident about the possibility.

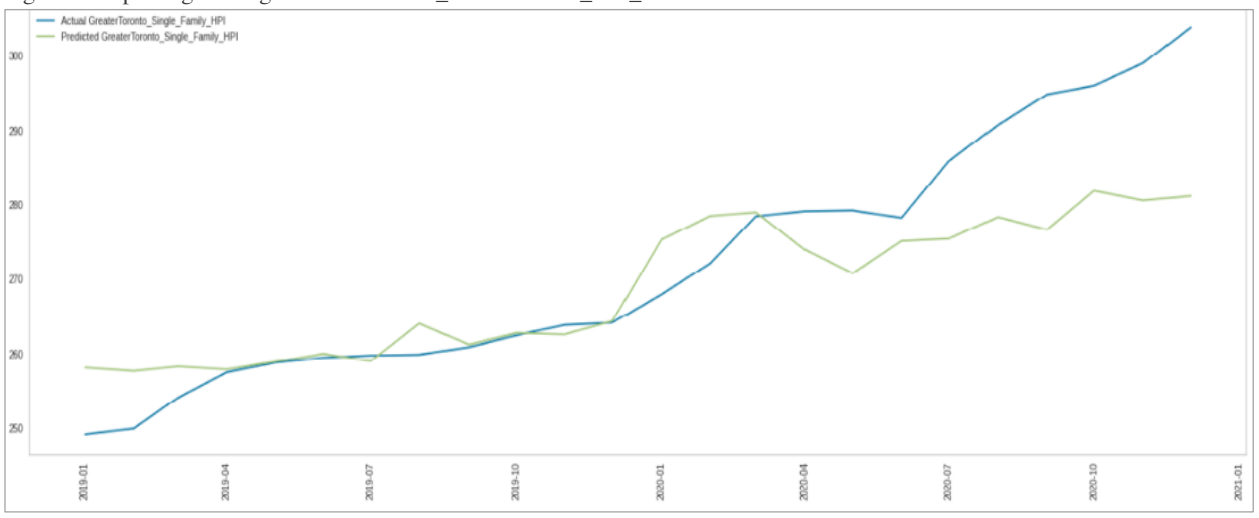
We kept all features but removed the self-lag feature with the highest coefficient, the 1 month lagged Toronto single-family HPI, and the 1-year self-lag too. We dropped 2016 and 2017 data that contained the housing bubbles in Canada as a whole, and we still arrived at an RMSE of 12.02. See figure 8 on the next page for a visual representation of the predictions on test data.

Figure 8: Actual vs Prediction results with all but self-lag feature



We then retrained our best model (9 features) by replacing the self-lag feature with the highly correlated 1-year lagged Ontario CPI shelter, dropped 2016 and 2017 data again, and arrived at an RMSE of 9.19 on test and R^2 value of 0.66 on test. See **Appendix G** for the coefficients. Also see figure 9 below. This methodology worked well until the pandemic, where it could have benefited with some self-lag information to capture and re-adjust for the out-of-normal changes in housing prices.

Figure 9: Replacing self-lag feature with 1Yr_MetroOntario_CPI_Shelter



7.0 Conclusion

7.1 Recommendations

The model we built can be used by individuals, private firms and government agencies alike. Let us discuss some of its recommended usages, implications, and improvements.

The HPI value predicted by our model can be used by consumers like home buyers or investors to plan the best time for refinancing, buying, or selling a home. Mortgage and Real Estate Agents can optimize their marketing spend and pre-plan personnel for peak seasons using the model's predictions. Banks can anticipate high mortgage demand as the prime rate in Canada lowers. Real estate developers can use our model to adjust prices in proportion to changes in HPI. Additionally, home repair and renovation services firms can either keep stable or adjust their prices in proportion to the changes in HPI. They can also plan for demand in materials accordingly. Government agencies can gain insights to proactively evaluate property taxes for the future. Policy makers can also act proactively based on various socio-economic factors like unemployment rate.

7.2 Model Implications

We explored the model's limitations and areas of improvement. The first limitation is that there is a possibility of data drift over time due to unpredictable circumstances like the housing bubbles we observed. The model performance needs to be monitored and the model should be retrained if it is predicting poorly in the future. The second limitation is that if the base year changes for either the HPI or CPI, our model will also require retraining. The third implication is that our model uses the prior month's data as lagged features, so it is reliant on each source releasing data in a timely manner. However, forecasting could be done to fill in the equation if necessary. For this reason, our model can only predict one month in advance, too.

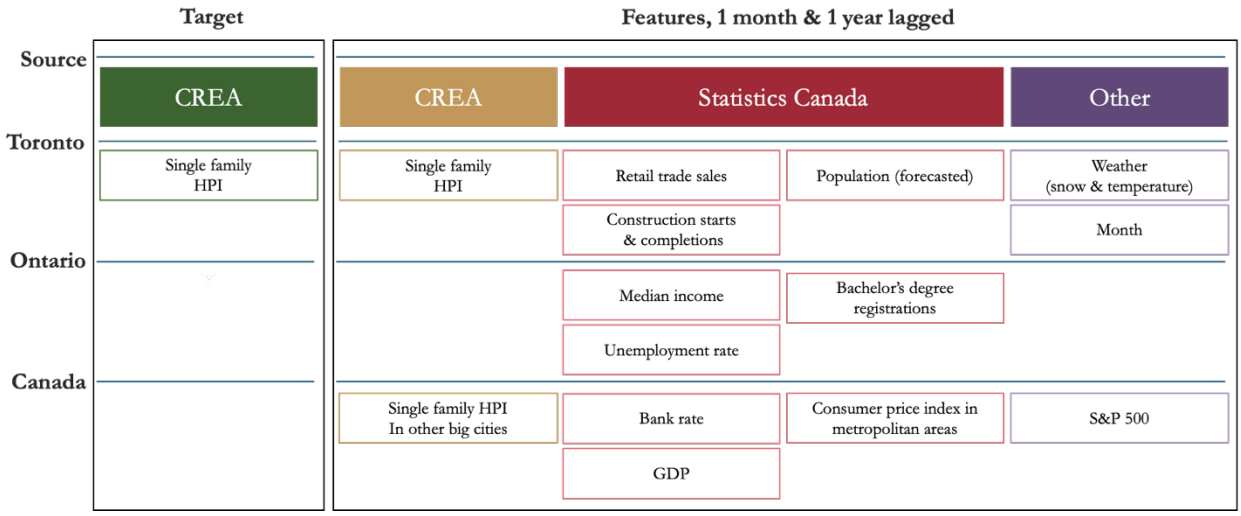
7.3 Model Improvements

In conclusion, we have brainstormed a few ways that the model could be improved. First, we would like to explore immigration data from the CIC to help understand how the arrival of new immigrants drives up housing demand. Second, we believe scraping social media data for sentiment analysis around housing trends, or policy changes that influence housing bubbles, could be another rich area to generate insights and create features. Finally, we are interested in testing a similar model and approach on other cities and home types for Canada-wide HPI predictions.

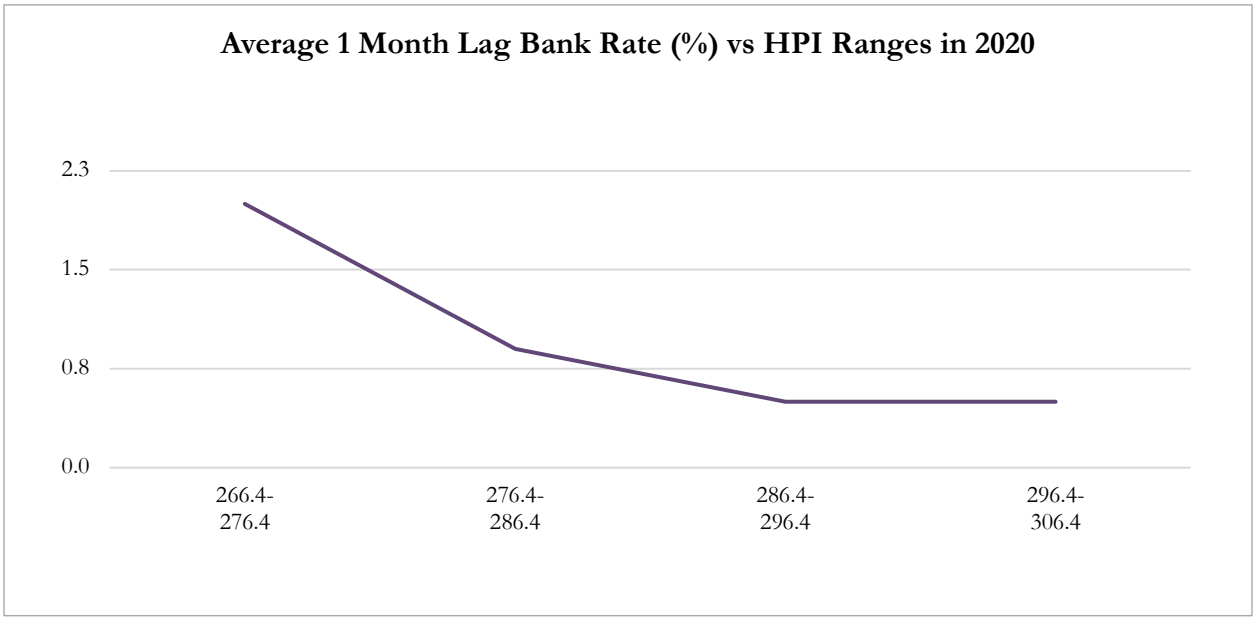
Appendix A: Evaluation of Hypotheses

Hypothesis	Confirmation	Notes
Correlation between growth in Toronto and growth in Vancouver single-family housing indexes (Vancouver lagged 1 month).	True.	Correlation = 0.97.
CPI for shelter (lagged 1 month) correlated with single-family home price index.	True.	Correlation = 0.97.
Seasonality captured in month.	True.	Feature used in model with negative coefficient showing winter months with lower prices.
Decrease in demand and therefore price during extreme weather conditions.	Depends on feature evaluated.	Negative coefficient for temperate in model, keeping in mind that temperature can have positive and negative values. Negative coefficient with month and as month as mentioned above.
Positive bivariate correlation between lagged GDP and target.	True.	Correlation = 0.91.
Construction starts and completions could signify growth in real estate demand and prices.	True.	Main features selected in model.
Construction commencements and population correlation.	False.	Correlation = 0.08.
Home price index and lagged income.	True.	Correlation = 0.95.
Decrease in bank rate increases demand for mortgages.	Relatively true.	Negative coefficient in model which suggests a relationship with prices however not necessarily demand.
Lagged unemployment correlated to target.	False.	Lower than |0.3| correlation.
Positive relationship between new university registrants (all degrees) and average annual income.	False.	Lower than |0.3| correlation. Could consider by degree type as secondary hypothesis.
Median age an interesting variable for modeling generation.	False.	Very low variance, not enough to represent generation or impact of changing age on model.

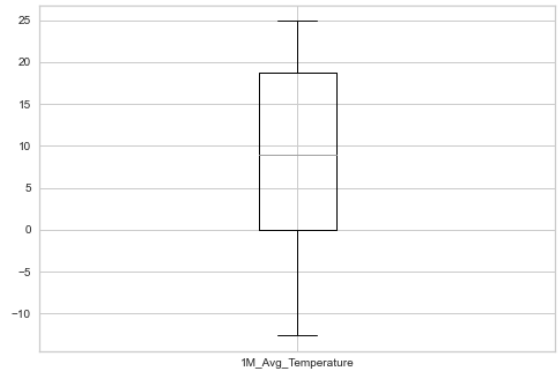
Appendix B: Infographic of Dataset Features and Sources

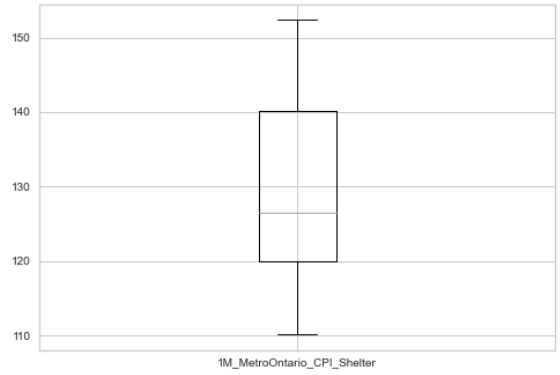


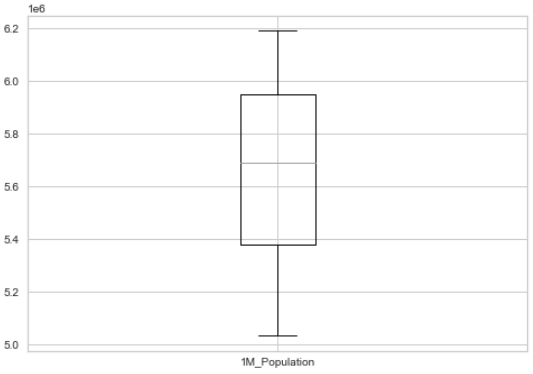
Appendix C: Bank Rate and HPI in 2020

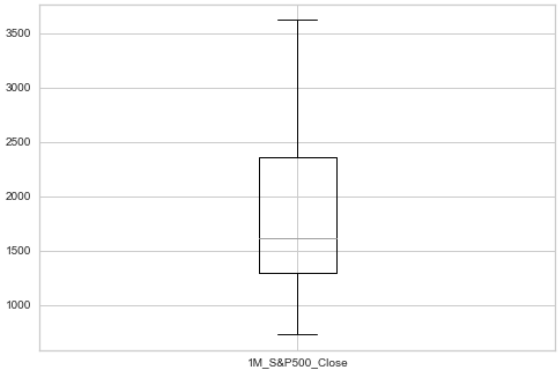


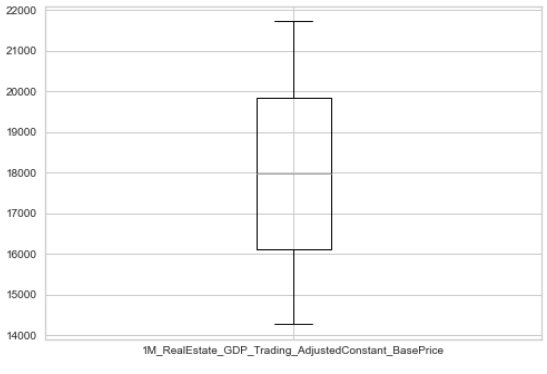
Appendix D: Examples of Boxplots



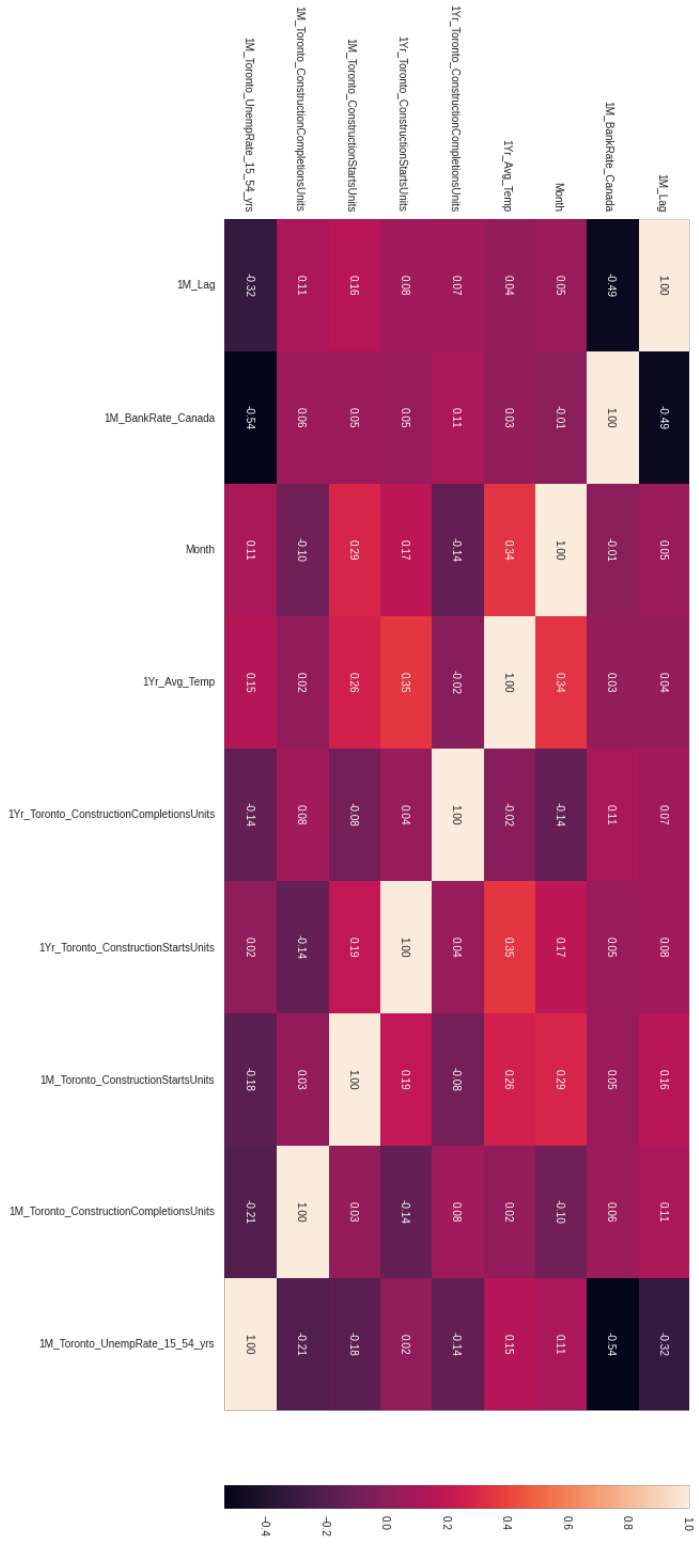




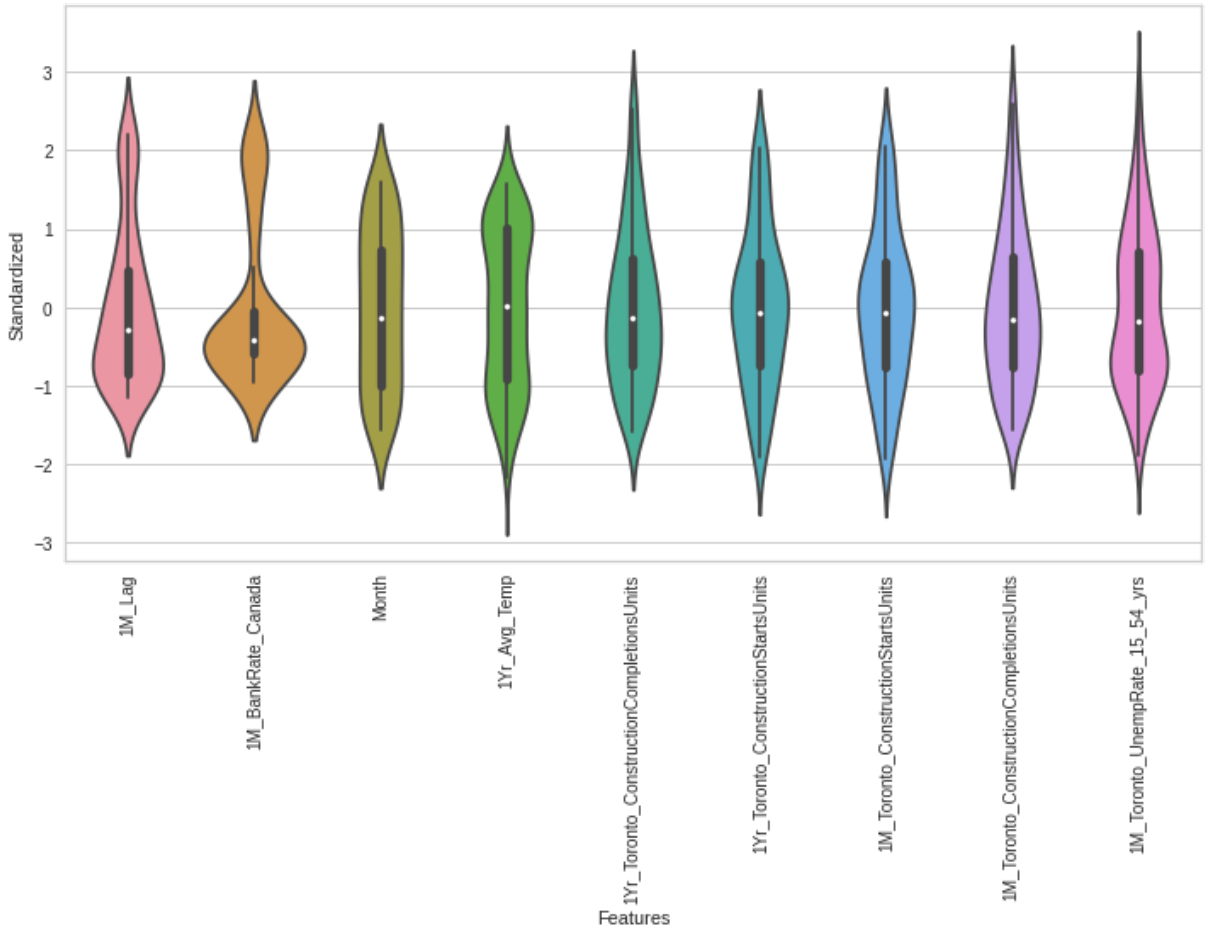




Appendix E: Removal of Multicollinearity in Selected Features



Appendix F: Violin Plots for Selected Features (Standardized)



Appendix G: Coefficients of No Self-Lag Model Non-Standardized Data

```
model.intercept_
```

```
-519.7287189587112
```

```
list(zip(X_train_final.columns,model.coef_))
```

```
[('1M_BankRate_Canada', 8.46055408447196),
 ('Month', -0.5020706116631865),
 ('1Yr_Toronto_ConstructionCompletionsUnits', -0.0005756395153487941),
 ('1M_Toronto_ConstructionCompletionsUnits', 0.0009545324625413516),
 ('1Yr_Avg_Temp', 0.0260529833104508),
 ('1Yr_Toronto_ConstructionStartsUnits', -0.0002612678782395743),
 ('1M_Toronto_ConstructionStartsUnits', -0.00014553575901940025),
 ('1M_Toronto_UnempRate_15_54_yrs', 0.6899041481243967),
 ('1Yr_MetroOntario_CPI_Shelter', 5.350402905473657)]
```