Predictive Analytics: House price prediction in Canada

CKME136 - Data Analytics: Capstone Crs - W2020

Student: Agnieszka Rutkowska

Abstract

Prices of houses in Canada seem to be constantly rising. Using datasets from Open Canada data repositories, this project aims to answer two questions:

1. How does employment ratio, average income and vacancy ratio impact housing prices?
2. How are the Canadian housing prices going to change in the next 5 years?

After identifying main predictors, causality analytics models and time-series models are developed and compared. Models are used to predict housing prices in the next 5 years (years 2020 - 2025).

Analysis is performed using Python language.

The source code for the project is stored at GitHub: https://github.com/aarutkowska/CKME136

Table of Contents

1. Introduction

The current situation on the housing market is referred to as Canadian property bubble. Real estate prices rose up to 337% between 2003 and 2020, with a partial correction in mid 2017 (Wikipedia, 2020).

As of February 2020, the housing market in the major agglomerations looks as follows:

- In Toronto low inventories led to further acceleration of price increase. If the supply remains low in the next months, prices can spiral upwards as they did in 2016 and early 2017;
- Vancouver agglomeration is a seller market and the decrease of inventory should keep it through the remainder of 2020;
- Ottawa and Montreal are reported as hottest (demand wise) and tightest markets in the country. Hight activity levels and low inventories continue;
- In Calgary and Edmonton the earlier elevated inventories are being drawn down and prices are recovering. Further reductions should stabilize prices later in the year. (Royal Bank of Canada, 2020)

The largest deterioration in housing affordability occurred between 2015 and mid 2016 because of rapid house price inflation in British Columbia and Ontario. Between 2015 – 2018 the factors influencing housing resale were mortgage rates and rules, house price growth, strong labour market and robust migration (Khan & Webley, 2019).

The below table shows home price forecast, annual percent change (Royal Bank of Canada, 2019)

Region	Year 2020	Year 2021
Canada	4.2	4.6
British Columbia	2.0	4.5
Alberta	2.4	4.0
Saskatchewan	2.5	4.0
Manitoba	3.5	3.1
Ontario	5.5	5.0
Quebec	4.8	4.1
New Brunswick	4.0	2.4
Nova Scotia	2.0	2.2
Prince Edward Island	4.3	3.1
Newfoundland & Labrador	1.2	1.5

2. Data description

Statistics Canada is the source of all datasets used for the analysis. All information is provided under Open Government Licence – Canada. (Government of Canada, 2019)

2.1. Introduction to New housing price index

New housing price index measures the difference in selling prices of new residential houses over time, where specification of the compared houses remain the same. The index is a source of information used by economics, academics and general public to monitor and analyse market trends. Data sampling is performed by surveying residential builders. Response to the survey is mandatory. Data is stratified by metropolitan area. The observed population in New Housing price index is limited to 27 metropolitan areas, representing all Canadian provinces. (Statistics Canada, 2020)

Appropriate methods are taken to detect errors and monitor data quality, making New housing price index a reliable source of information.

2.2. Dataset: New housing price index

Source: New housing price index, monthly (Statistics Canada, 2019)

Column name (Statistics Canada, 2019)	**Description**
REF_DATE	Date: range 1981-01-01 to 2019-12-01, format yyyy-mm-dd
GEO	Region (country, province or metropolitan area), factor (40 values)
DGUID: Dissemination Geography Unique Identifier	Region ID, factor (40 values), paired with GEO, note: blank for "Saint John, Fredericton, and Moncton, New Brunswick" region
New housing price indexes	Index type: factor (3 values): "Total (house and land)", "House only", "Land only"
UOM: Unit of measure	Index base, 1 value: "Index, 201612=100"
UOM_ID	1 value: "347", paired with UOM
SCALAR_FACTOR	1 value: "units"
SCALAR_ID	1 value: "0", paired with SCALAR_FACTOR
VECTOR	120 values, format v\d{9}
COORDINATE	120 values, numeric (one decimal point), paired with VECTOR
VALUE	1. Value, one decimal point, note: multiple blanks, the index base period, for which the New Housing Price Index (NHPI) equals 100, is December 2016.
STATUS	Data quality, factor with 4 values: "..", "x", "E", blank

	Legend Extract *".." not available for a specific reference period* *"E" use with caution* *"x" suppressed to meet the confidentiality requirements of the Statistics Act*
SYMBOL	blank
TERMINATED	blank
DECIMALS	1 value: "1"

2.3. Dataset: Employment and average weekly earnings

Source: Employment and average weekly earnings (including overtime) for all employees by province and territory, monthly, seasonally adjusted (Statistics Canada, 2019)

Column name (Statistics Canada, 2019)	**Description**
REF_DATE	Date: range: 2001-01-01 to 2019-10-01, format yyyy-mm-dd
GEO	Region: country or province, 14 values
DGUID: Dissemination Geography Unique Identifier	Region ID, factor (14 values), paired with GEO
Estimate	Index type, 2 values: "Employment for all employees", "Average weekly earnings including overtime for all employees"
NAICS: North American Industry Classification System	Factor, 28 values
UOM: Unit of measure	Factor, 2 values: "persons", "units"
UOM_ID	2 values: "81", "249", paired with UOM
SCALAR_FACTOR	1 value: "units"
SCALAR_ID	1 value: "0", paired with SCALAR_FACTOR
VECTOR	742 values, format v\d{9}
COORDINATE	742 values, numeric (one decimal point), paired with VECTOR

VALUE	Value, numeric, integers for UOM = "persons", two decimal points for UOM = "units", multiple blanks
STATUS	Data quality, factor with 9 values: "..", "A", "B", "C", "D", "E", "F", "x", blank *Legend Extract* *".." not available for a specific reference period* *"A" data quality: excellent* *"B" data quality: very good* *"C" data quality: good* *"D" data quality: acceptable* *"E" use with caution* *"F" too unreliable to be published* *"x" suppressed to meet the confidentiality requirements of the Statistics Act*
SYMBOL	blank
TERMINATED	blank
DECIMALS	2 values: "0", "2" matching UOM

2.4. Dataset: Canada Mortgage and Housing Corporation, vacancy rates

Source: Canada Mortgage and Housing Corporation, vacancy rates, apartment structures of six units and over, privately initiated in census metropolitan areas (Statistics Canada, 2020)

Column name (Statistics Canada, 2019)	**Description**
REF_DATE	Date, range 1971-01-01 to 2019-01-01, format yyyy
GEO	Region (country, province or metropolitan area), factor (37 values)
DGUID: Dissemination Geography Unique Identifier	Region ID, factor (35 values), paired with GEO, note: blank for GEO values "Census metropolitan areas", "Montréal excluding Saint-Jérôme, Quebec"
UOM: Unit of measure	1 value: "Rate"
UOM_ID	1 value: "257", paired with UOM
SCALAR_FACTOR	1 value: "units"
SCALAR_ID	1 value: "0", paired with SCALAR_FACTOR
VECTOR	37 values, , format v\d{6-9}

COORDINATE	37 values, integer (1 - 37), paired with VECTOR
VALUE	value, one decimal point
STATUS	blank
SYMBOL	blank
TERMINATED	2 values: "t", blank *Legend Extract* *"t" terminated*
DECIMALS	1 value: "1"

2.5. Data source constrains and assumptions

Changes introduced to the New Housing Price index in December 2016 (Statistics Canada, 2019):

- Thunder Bay no longer included
- The index for Ottawa-Gatineau (Quebec part) begins
- The index for Oshawa begins
- Guelph is included in the provincial aggregate for Ontario
- Separate indexes are published for the census metropolitan areas (CMAs) of Toronto, Oshawa, Ottawa-Gatineau (Ontario part), Ottawa-Gatineau (Quebec part) and Greater Sudbury
- The census metropolitan areas (CMAs) of Ottawa-Gatineau (Quebec part), Trois-Rivières and Sherbrooke are included in the provincial aggregate for Quebec
- The census metropolitan area (CMA) of Kelowna is included in the provincial aggregate for British Columbia.
- For historical continuity the Ottawa-Gatineau (Ontario part) index is linked to the previously combined Ottawa-Gatineau index, the Toronto index is linked to the previously combined Toronto and Oshawa index, the Greater Sudbury index is linked to the previously combined Greater Sudbury and Thunder Bay index, which were published until December 2016

For the vacancy rates dataset, geographical areas are modified every 5 years to reflect most recent census definitions, therefore, data are not strictly comparable historically. (Statistics Canada, 2020)

New Housing price index and employment and average weekly earnings datasets contain data with quality rating of E and lower. That data needs to be used with caution or excluded from the analysis.

Vacancy rates dataset contains records marked as terminated. These need to be excluded from the analysis.

Datasets differ in data ranges and frequency data is reported. When querying data, this needs to be considered and addressed.

All data sets use Dissemination Geography Unique Identifier (DGUI), which will be used to join tables in the queries. The areas included in the datasets differ. When querying data, this needs to be considered and addressed.

3. Literature review

This project uses Cross-industry standard process for data mining: CRISP-DM (Larose & Larose, 2015, p. 7).

1. Business/research understanding phase
2. Data understanding phase
3. Data preparation phase
4. Modeling phase
5. Evaluation phase
6. Deployment phase

3.1. Data pre-processing and exploratory analysis

Data pre-processing includes data cleaning and data transformation activities. Attention should be given to fields that are obsolete or redundant, missing values, outliers, data in a form not suitable for the data mining models and values not consistent with policy or common sense (Larose & Larose, 2015, p. 20). To start with, data description should be provided and variables need to be split into features and target variable (Roman, 2019).

Missing data can be handled by omitting the records from analysis or replacing missing values. Omitting values is potentially dangerous, as the pattern of missing values might be systematic, and deleting the records might lead to a biased subset of the data (Larose & Larose, 2015, p. 21), however it is often used, e.g. dropping columns which have more null values than assumed (e.g. 8) and then removing rows containing null values (Peixerio, Project 2 - Predict air quality with Prophet, 2019). Missing data can be replaced with a constant, mean, mode, random value or imputed value (Larose & Larose, 2015, p. 21), e.g. with an average considering only the positive values (Peixerio, Project 2 - Predict air quality with Prophet, 2019). Duplicated records need to be removed (Larose & Larose, 2015, p. 45). If needed, data types might be converted, e.g. from text to float (Peixerio, Project 2 - Predict air quality with Prophet, 2019).

Datasets need to be checked for possible misclassifications and outliers. Misclassifications can be identified by analyzing frequency distribution. Tools for identifying outliers include histograms, two dimensional scatter plots (graphical approach) (Larose & Larose, 2015, pp. 25-27) and z-score calculation (numerical approach) (Larose & Larose, 2015, p. 38). Examples include visualizing the location of the houses based on latitude and longitude (Raghavan, 2017), creating scatter plots to look into how common factors (such as house size, number of bedrooms, area represented by a zip code) affect the target variable house price (Raghavan, 2017).

Central measures are used to see if data set is skewed. Calculating data ranges and standard deviation enables understanding data distribution (Larose & Larose, 2015, pp. 28-29). If data is distribution is not symmetrical transformation can be applied. Normal probability plot can be used to verify if data distribution is normal (Larose & Larose, 2015, p. 32), as well as plotting histograms is a good visual aid to understand the distribution (Kim, 2019). Skewness and kurtosis parameters for the distribution need to be analyzed (Kim, 2019). Common data transformations include natural logarithm transformation, square root transformation and the inverse square root transformation (Larose & Larose, 2015, p. 38), e.g. by using *log1p* function to reduce skewness and kurtosis. *Log1p* function can also be represented as *log(1+x)* (Kim, 2019).

Exploratory analysis includes getting to know datasets by understanding variable types, examining the interrelationships among the attributes, identifying interesting subsets for observation and developing the initial idea of possible associations amongst predictors, as well as between the predictors and the target variable (Larose & Larose, 2015, p. 54).

Correlation matrix is used to identify features of interest, i.e. the ones with high correlation with the target variable (Roman, 2019). To further understand these variables, box plots can be used. Remove outliers helps to increase correlation score (Kim, 2019).

For time series analysis autocorrelation plot is used to understand if there is seasonality in the data or if time series is stationary (i.e. mean and variance are constant and covariance is independent of time). Dickey-Fuller test should be used to verify if the series is stationary. H_0 hypothesis for the test states that unit root is present. If $p > 0$, process is not stationary; if $p = 0$, H_0 can be rejected and process can be treaded as stationary (Peixerio, The Complete Guide to Time Series Analysis and Forecasting, 2019). Plotting the data over the entire period of time (e.g. stock prices) is a visual method used to verify if process is stationary and if seasonality in the data can be noticed (Peixerio, Project 1 - Predicting stock price, 2019). For a smoother trend, data can be aggregated (e.g. by day or week) (Peixerio, Project 2 - Predict air quality with Prophet, 2019).

3.2. Dimension reduction

To guard against multicollinearity, which might lead to instability in the solution space and possible incoherent results, dimension reduction can be applied. Retaining too many variables may lead to overfitting. Also, analysis solely at the variable level might miss the fundamental underlying relationships among the predictors. Dimension reduction techniques include principal component analysis (PCA), factor analysis and user-defined composites (Larose & Larose, 2015, pp. 92-93). Data should be standardized (mean set to 0 and standard variation set to one) and prior to reduction (Larose & Larose, 2015, p. 94). PCA is used to substitute a smaller number of uncorrelated components for the original variables (Larose & Larose, 2015, p. 110). Factor analysis represents the model of the data. It is used to apply factor rotation (Larose & Larose, 2015, p. 111). User-defined composite combines several variables together into a simple composite measures. They are know as summated scales. Compared to the use of individual variables, user-defined composites provide a way to diminish the effects of measurement error (Larose & Larose, 2015, pp. 117-118).

3.3. Data modeling

Data mining methods can be categorized as supervised and unsupervised. The majority of machine learning problems fall under supervised. The goal of these methods is to approximate the mapping function so well that new input data can predict the output variables for that data. Supervised data mining methods include recommendations and regression problems, such as time series analysis. In unsupervised learning there are only input data and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. Unsupervised data mining methods include clustering and association (Brownlee, 2019).

3.3.1. Multiple regression and decision trees

Before building the model, data has to be split into training and test subsets (Roman, 2019), e.g. 90% for train and 10% for test set (Raghavan, 2017). Typically, data is also shuffled into a random order to remove bias in the ordering of the dataset (Roman, 2019).

K-fold cross validation is a technique used for making sure that model is well trained, without using the test set. It consists of splitting data into *k* partitions of equal size. For each partition *i*, we train the model on remaining *k-1* parameters and evaluate it on partition *i*. The final score is the average of the k scores obtained (Roman, 2019).

Underfitting is caused by high bias and occurs when the model can't predict even the outcomes of training set (didn't learn well). Overfitting is caused by high variance and occurs when model did learn well on data to the point of memorizing it and is not able to generalize on a new data set. When the balance between bias and variance is just right, the model is able to predict correctly (Roman, 2019).

To minimise the risk of overfitting on a test set another part of the dataset can be held out as a validation set, however reducing the size of training set can result is underfitting. By using k-fold validation we allow to train the model even if little data is available (Roman, 2019).

Multiple regression is used to explore the relationship between the target variable and two or more predictor variables (Larose & Larose, 2015, p. 236). An example is applying multiple linear regression on factors such as number of bedrooms, number of bathrooms, square feet area, number of floors etc. to predict sales prices of the house (Raghavan, 2017).

Decision tree models need to have maximum depth defined. Looking at the example of 4 graphs for a decision tree model with different maximum depths – each graph visualizes the learning curves of the model for both training and testing as the size of the training set is increased for maximum depth values of 1, 3, 6 and 10. At depth 3, for this particular model, we can notice that as the number of training points increase, the training score decreases; In contrast, the test score increases. Also, training and testing scores tend to converge, so having more training points will not benefit the model. By plotting complexity curves, the optimal maximum depth value can be determined. Complexity curve is a graph for a decision tree model that has been trained and validated on the training data using different maximum depths. The graph produces two complexity curves – one for training and one for validation sets. The curves represent bias – variance trade off. For this model, the best max depth is 4, as it yields best validation score. For more depth, the training score decreases, which is the sign of overfitting (Roman, 2019).

To fit a model we use the optimal maximum depth parameter and grid search technique. In this example 10 shuffled sets were created and for each 20% of data will be used as the validation set. (Roman, 2019). Popular models used in Python are: lasso, elastic net, kernel ridge, gradient boosting, XGBoost and Light GMB regression (Kim, 2019). Grid search technique exhaustively generates candidates from a grid of parameter values specified, which is a dictionary with the values of the hyperparameters to evaluate. The example shows two variants of grid explored: one of linear values of C parameter, the second one with RBF kernel as cross products of C parameter and γ parameter, where C in {1, 10, 100, 1000}, γ in {0.001, 0.0001}. When fitting it on a dataset all possible combinations of parameter values are evaluated and the best combination is retained (Roman, 2019). Hyper-parameters are parameters that are not directly learnt within estimators. They are passed as arguments to the constructor of the estimator classes. Typical examples include C, *kernel* and γ for Support Vector Classifier, α for Lasso (Pedregosa, et al., 2011). RBF kernel, also called Gaussian kernel, stands for radial basis function kernel. It is one of positive-definite kernel used in operatory theory and is a generalization of a positive-definite function or a positive-definite matrix. Examples of other positive-definite kernels are linear kernel, polynomial kernel, Laplacian kernel (Wikipedia, 2020).

3.3.2. Time series analysis

Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. (NIST/SEMATECH, 2013, p. 6.4.). We can differentiate between univariate and multivariate time series models.

In the beginning, data should be split into a train and test sets. Test set should be created by holding out the last entries for prediction and validation, e.g. 30 records. (Peixerio, Project 2 - Predict air quality with Prophet, 2019)

Basic techniques include single moving average and centered moving average, but these are not able to cope with a significant trend (NIST/SEMATECH, 2013, pp. 6.4.2.1-6.4.2.2.). Using moving average requires adjusting window size parameter (Peixerio, The Complete Guide to Time Series Analysis and Forecasting, 2019).

Exponential smoothing assigns exponentially decreasing weights as the observation get older, this way recent observations are given relatively more weight in forecasting than the older observations (NIST/SEMATECH, 2013, p. 6.4.3.). Single, double and triple exponential smoothing models can be applied. Double exponential smoothing is used when there is a trend in a data series (Peixerio, The Complete Guide to Time Series Analysis and Forecasting, 2019). Triple exponential smoothing is used when data shows both trend and seasonality (NIST/SEMATECH, 2013, p. 6.4.3.5.). Exponential smoothing requires adjusting smoothing factor α (smoothing set to 0 approaches moving average model), double exponential smoothing uses same factor α and trend smoothing factor β, triple exponential smoothing uses factors α, β, seasonal smoothing factor γ and length of the season L. α, β

and γ can take values between 0 and 1 (Peixerio, The Complete Guide to Time Series Analysis and Forecasting, 2019). In an example moving average of stock prices was smoothed by 30 days and 60 days. In the next example exponential smoothing was used smoothed with α factors 0.05 and 0.3, to show how changing parameters affects the curve. Next example involves double exponential smoothing with (α, β) pairs as follows: (0.9, 0.9), (0.9, 0.02), (0.02, 0.9), (0.02, 0.02). In summary of those examples, it is advised to experiment with different curves to observe their behaviours. (Peixerio, Project 1 - Predicting stock price, 2019)

Univariate time series models assume that the data are stationary. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time (NIST/SEMATECH, 2013, p. 6.4.4.2.). Dickey-Fuller test is used to see if the process is stationary (Peixerio, Project 1 - Predicting stock price, 2019).

If the time series is not stationary, we can often transform it to stationarity. Although seasonality also violates stationarity, this is usually explicitly incorporated into the time series model. (NIST/SEMATECH, 2013, p. 6.4.4.2.). The following techniques can be used to detect seasonality: run sequence plot, a seasonal subseries plot, multiple box plots, the autocorrelation plot can help identify seasonality (NIST/SEMATECH, 2013, p. 6.4.4.3.).

Autoregressive (AR) Models, are constructed using regression of the current value of the series against one or more prior values of the series (NIST/SEMATECH, 2013, p. 6.4.4.4.). These models can be analyzed using standard linear least squares techniques. Example result of Dickey-Fuller test $p=0.54$ means that time series is not stationary. Autocorrelation plots show autocorrelation is very high and have no clear seasonality. To get rig of high autocorrelation and to make process stationary time series is subtracted from itself with a one day lag. This ay time series with $p=0$ is obtained, with low autocorrelation and partial autocorrelation (Peixerio, Project 1 - Predicting stock price, 2019).

Other common approaches to univariate time series include frequency based methods, i.e. modeling a sinusoidal type data, where spectral plot is the primary tool and Box-Jenkins Approach, which combines the moving average and the autoregressive approaches. Box-Jenkins models are considered very powerful. They also require long observations series of minimum 50 observations (NIST/SEMATECH, 2013, p. 6.4.4.4.). ARMAV model (AutoRegressive Moving Average Vector model) is the multivariate form of the Box-Jenkins univariate model (NIST/SEMATECH, 2013, p. 6.4.5.).

Seasonal Autoregressive Integrated Moving Average model (SARIMA) is a combination of simpler models that can model time series exhibiting non-stationary properties and seasonality. Before applying SARIMA, we must apply transformations to our time series to remove seasonality and non-stationary behaviours. The following models are the part of SARIMA(p,d,q)(P,D,Q,s)

- Autoregression model AR(p), to identify lag parameter p
- Moving average model MA(q) to identify q – biggest lag, after which other lags are not significant on the correlation plot
- Order of integration I(d), where d is number of differences required to make the series stationary
- Seasonality S(P,D,Q,s), where s – season length, $P=p$, $Q=q$, D – order of seasonal integration representing the number of differences required to remove seasonality from the series (Peixerio, The Complete Guide to Time Series Analysis and Forecasting, 2019)

To find a best performing model, all combinations of the parameters should be checked (Peixerio, Project 1 - Predicting stock price, 2019).

One of powerful modeling tools is Prophet (Peixerio, Project 2 - Predict air quality with Prophet, 2019). Prophet is a procedure for forecasting time series data based on an additive models where non-linear trends are fit with yearly, weekly and daily seasonality, implemented in R and Python (Github, n.d.).

3.4. Evaluating models

Models can be too complex or too simple to sufficiently generalize the data. Algorithms might not be appropriate for the structure of the data given. Data could be too noisy or contain too few samples to allow the model to adequately capture the target variable (Roman, 2019), so performance of models needs to be evaluated. For estimation and prediction models, we are provided with both estimated or predicted value $\hat{y}$ of the numeric target variable and the actual value y (Larose & Larose, 2015, p. 452).

Mean square error *MSE* and standard error of the estimate s is used to evaluate model performance. Mean absolute error *MAE* minimizes the influence of outliers, but is not available in all statistical packages (Larose & Larose, 2015, p. 454).

Mean average percentage error *MAPE* is also used as the metric to evaluate the error of the model. In the example using SARIMA model to predict stock prices, the final model has *MAPE* of 0.79% and when we use test data set, the model does not predict well as predictions are flat and all predictions are below the actual price, which indicates that model does not work well (Peixerio, Project 1 - Predicting stock price, 2019). In another example, for model created using Prophet tool to predict air quality, MAPE of 13.86% was achieved, which is considered a good result, especially without any fine tuning of the model, which shows only downward trend and had not identified any seasonality (Peixerio, Project 2 - Predict air quality with Prophet, 2019).

Coefficient of determination R^2 is another measure of goodness of regression models. R^2 represents the proportion of the variability in the response that is accounted for by the linear relationship between predictors and response (Larose & Larose, 2015, p. 453). R^2 describes how good the model is at making predictions. If $R^2 = 0$, the model is not better than always predicting mean (negative R^2 would indicate that the model is even worse than always predicting mean), $R^2 = 1$ would stand for perfect prediction, R^2 value between 0 and 1 indicates what percentage of target variable can be explained by the feature (Roman, 2019).

4. Report

4.1. Data cleaning

Purpose of this stage is to get familiarised with data and prepare data for next steps of the analysis.

3 csv files are used for the analysis: New housing price index, Employment and average weekly earnings, Canada Mortgage and Housing Corporation, vacancy rates. Data is loaded from local drive as one of the source files is too big to upload to GitHub (but compressed version of the file has been uploaded to GitHub repository).

4 data frames are the output of this stage, ready for further processing.

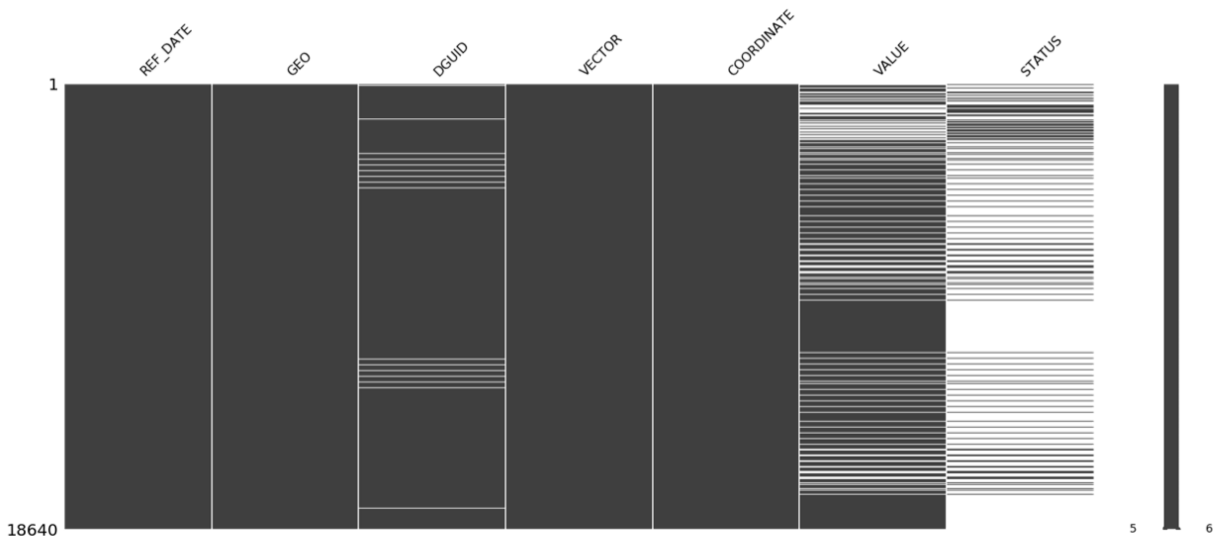
The target variable is VALUE for in New housing price index.

4.1.1. New housing price index

From New housing price index rows with values other than 'House Only' are dropped (i.e. rows with value 'Land Only' and 'Total (house and land)'), as only house value is in scope of the analysis. Columns which are blank or do not add information are dropped. Data types are set as follows:

```
REF_DATE      datetime64[ns]
GEO                 category
DGUID               category
VECTOR                object
COORDINATE           float64
VALUE                float64
STATUS              category
dtype: object
```

Based on information in STATUS column values with insufficient data quality are identified – and rows containing those values are dropped. Distribution of missing values looks as follows:

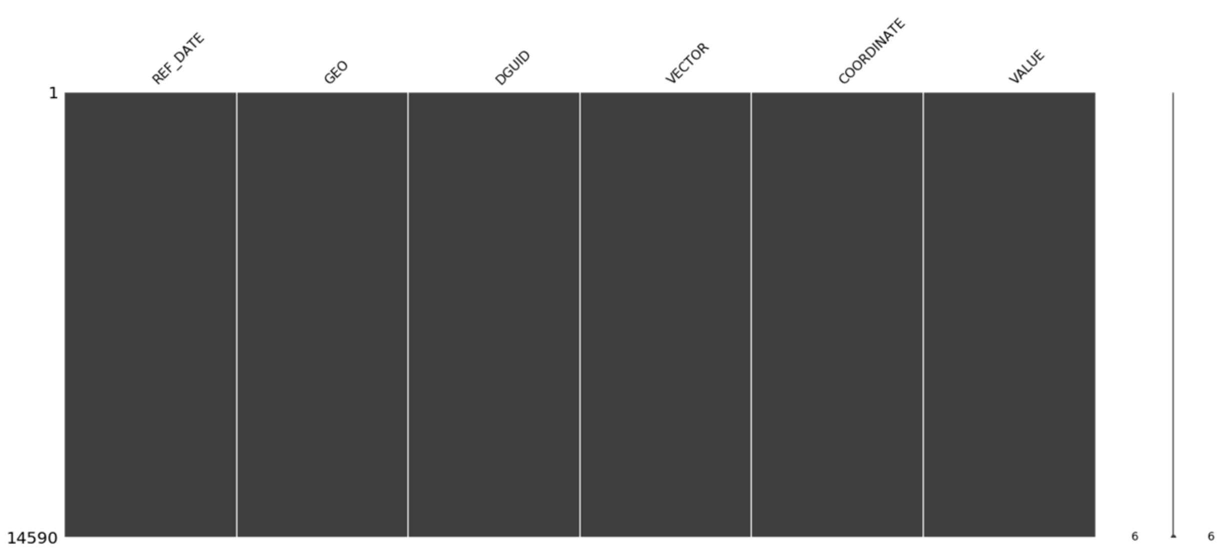


Attempt to replace missing DGUID values based on GEO value was not successful. Looking on the GEO values for which DGUID is blank:

```
GEO
Saint John, Fredericton, and Moncton, New Brunswick    468
Name: REF_DATE, dtype: int64
```

It can be observed that there is no corresponding GEO value to fill in blanks (above GEO value does not have DGUID assigned), so rows with blank DGUID values are dropped. DGUID is needed to join with other tables to be able to analyze data by regions.

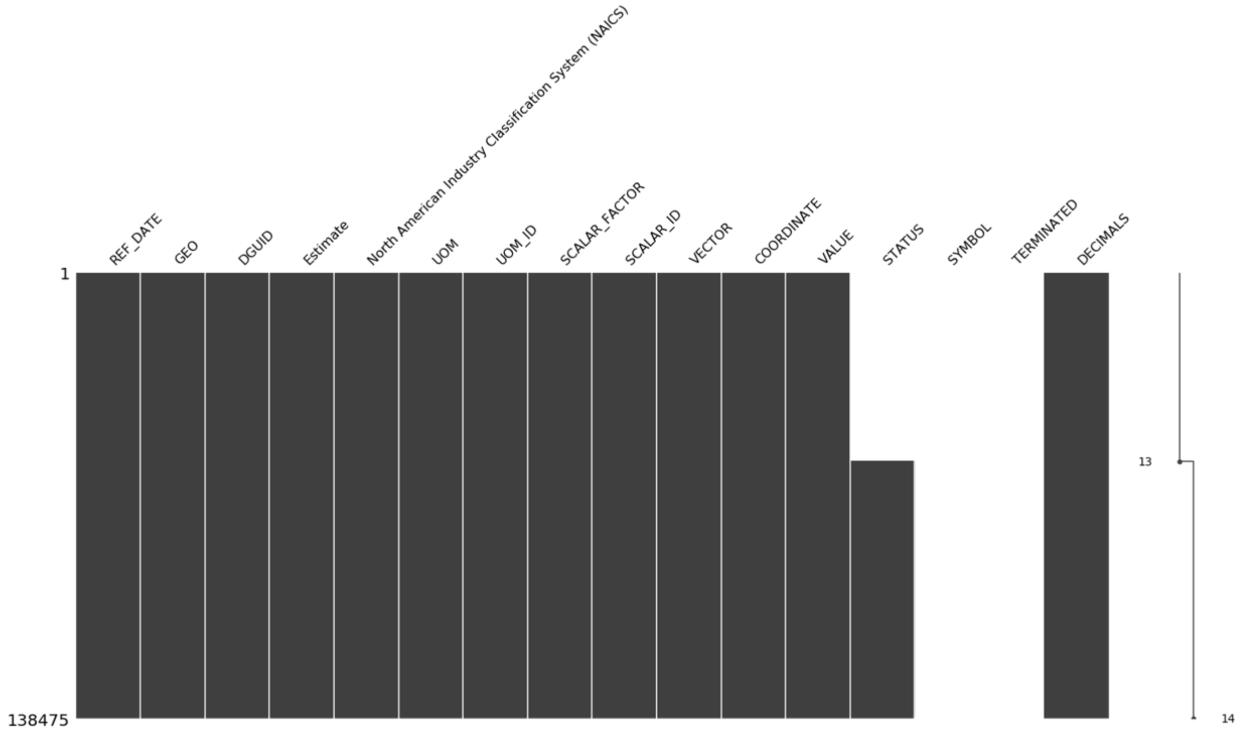
Distribution of missing VALUE on GEO, DGUID is verified, it can be observed that the missing values occur for before region was added to the report (as indicated in the data description), so as no values are available rows with missing VALUE are removed – rows with missing values are also indicated in status column. As status column is now empty, it is dropped. Another check for missing values is performed:



14590 rows with no missing values remain.

4.1.2. Employment and average weekly earnings

Based on information in status column rows with bad data quality are identified and dropped. Rows with missing values are previewed. There are no missing values identified.



Columns which are blank or which do not contain usable information are dropped. Data types are adjusted as follows:

```
REF_DATE                                                  datetime64[ns]
GEO                                                             category
DGUID                                                           category
Estimate                                                        category
North American Industry Classification System (NAICS)             object
UOM                                                             category
VECTOR                                                            object
COORDINATE                                                        object
VALUE                                                            float64
dtype: object
```

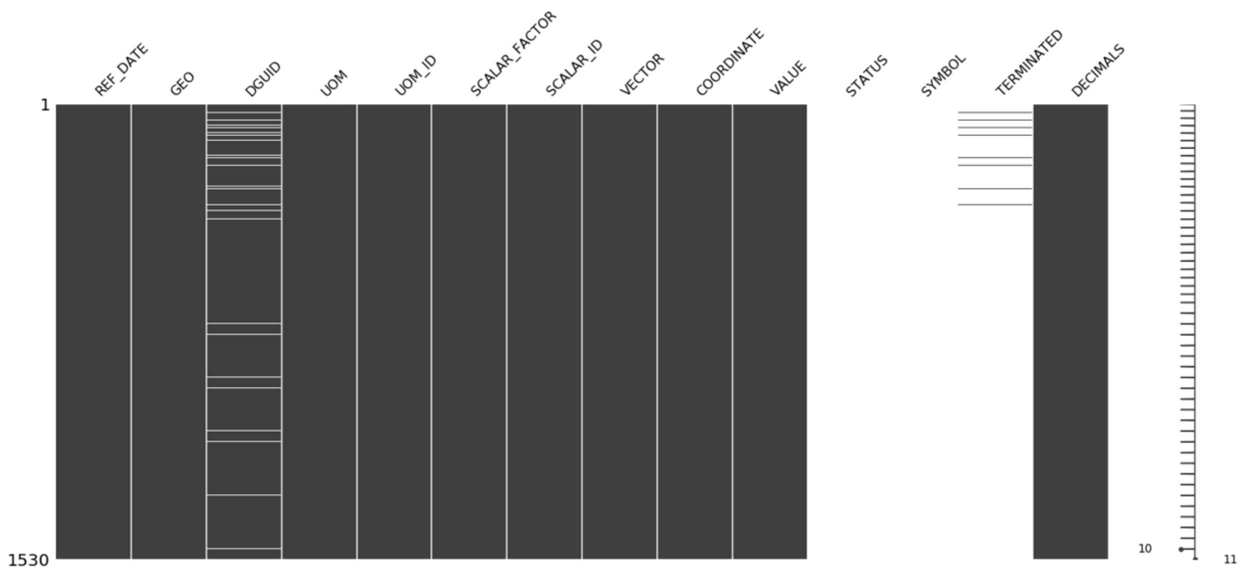
Estimate column contains 2 values. Looking at count of value by Estimate and UOM:

```
Estimate                                                 UOM
Average weekly earnings including overtime for all employees  Dollars    66466
Employment for all employees                             Persons    72009
Name: REF_DATE, dtype: int64
```

Data frame is separated into two data frames for both estimate types: average earnings (dollars) and employment (persons).

4.1.3. Canada Mortgage and Housing Corporation, vacancy rates

Based on information in status column rows with bad data quality are identified and dropped. Rows with missing values are identified and dropped.



Columns which will not be used (are blank or contain no added information) are removed, data types are adjusted as follows:

```
REF_DATE      datetime64[ns]
GEO                 category
DGUID               category
VECTOR                object
COORDINATE             int64
VALUE                float64
TERMINATED          category
dtype: object
```

4.2. Exploratory analysis

Data ranges and frequency of 4 input files were reviewed:

```
housing_index3

count                      14590
unique                       468
top          2019-09-01 00:00:00
freq                          39
first        1981-01-01 00:00:00
last         2019-12-01 00:00:00
Name: REF_DATE, dtype: object
```

```
employment_persons

count                      72009
unique                       226
top          2019-04-01 00:00:00
freq                         332
first        2001-01-01 00:00:00
last         2019-10-01 00:00:00
Name: REF_DATE, dtype: object
```

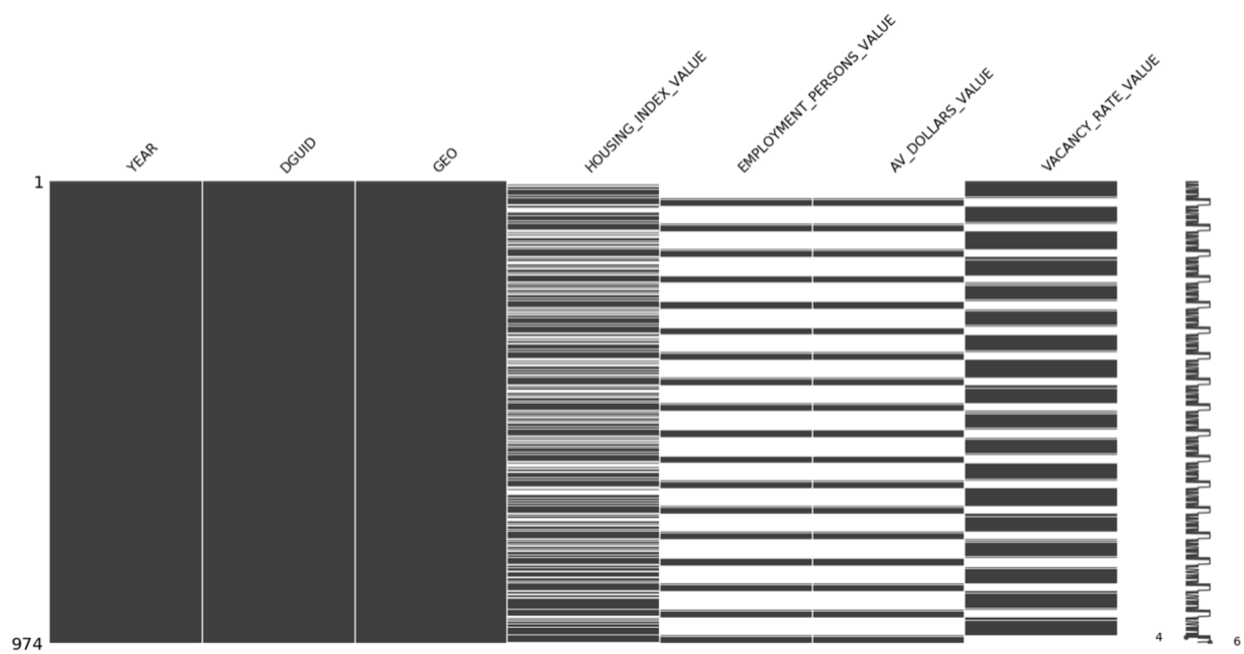
```
average_dollars

count                      66466
unique                       226
top          2003-06-01 00:00:00
freq                         302
first        2001-01-01 00:00:00
last         2019-10-01 00:00:00
Name: REF_DATE, dtype: object
```

```
vacancy_rate

count                       1454
unique                        49
top          2019-01-01 00:00:00
freq                          35
first        1971-01-01 00:00:00
last         2019-01-01 00:00:00
Name: REF_DATE, dtype: object
```

The first full year for all 4 tables is 2001, the last full year for all 4 tables is 2018. Data needed to be aggregated to annual (which is how data on vacancy rate is delivered) so that tables could be merged. Tables were merged on year, DGUID, GEO.



A province was assigned to each GEO and all values were aggregated by province. Preview of results looks as follows:

	YEAR	Province	HOUSING_INDEX_VALUE	EMPLOYMENT_PERSONS_VALUE	AV_DOLLARS_VALUE	VACANCY_RATE_VALUE
0	2001	British Columbia	62.533333	4.660451e+04	651.153958	1.900000
1	2001	New Brunswick	59.079167	2.202550e+06	715.439583	0.850000
2	2001	Ontario	62.211364	3.094027e+05	645.782085	2.518182
3	2001	Quebec	62.708333	1.455693e+05	712.294208	1.040000
4	2001	Saskatchewan	57.591667	1.634547e+05	649.716091	4.400000
...	...	...	...	...	...	...
85	2018	British Columbia	101.879167	5.612521e+04	1065.639001	3.400000
86	2018	New Brunswick	103.470833	2.585179e+06	1120.981378	1.600000
87	2018	Ontario	101.303472	3.610820e+05	1024.418376	3.063636
88	2018	Quebec	107.630000	1.889555e+05	1073.019382	2.000000
89	2018	Saskatchewan	100.655556	2.022395e+05	1083.494323	4.200000

90 rows × 6 columns

Correlation for values was calculated:

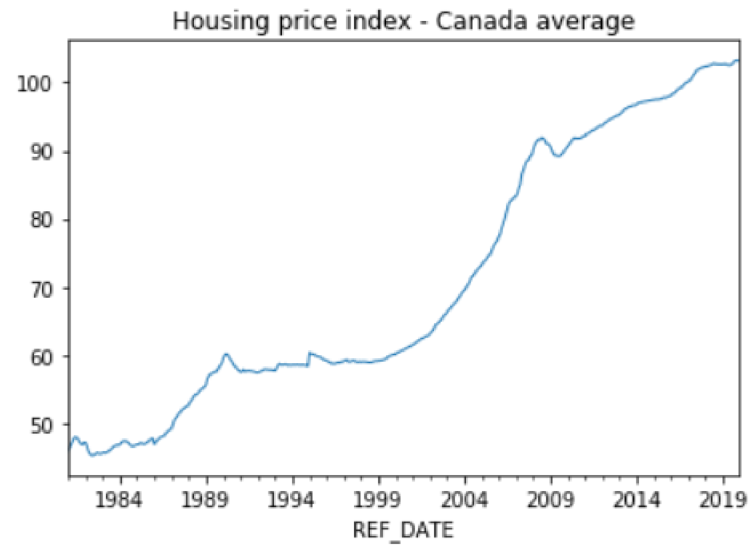
	HOUSING_INDEX_VALUE	EMPLOYMENT_PERSONS_VALUE	AV_DOLLARS_VALUE	VACANCY_RATE_VALUE
HOUSING_INDEX_VALUE	1.000000	0.011146	0.916782	0.394644
EMPLOYMENT_PERSONS_VALUE	0.011146	1.000000	0.203625	-0.192798
AV_DOLLARS_VALUE	0.916782	0.203625	1.000000	0.381146
VACANCY_RATE_VALUE	0.394644	-0.192798	0.381146	1.000000

It can be observed that there is strong positive correlation between New housing price index and average weekly earnings (0.916), whereas vacancy rate is only moderately correlated with New housing price index (0.394). Employment rate has a weak positive correlation of 0.01 with New housing price index.

Further analysis could identify whether house prices rise faster than earnings. That could be done by creating a linear regression model.

4.3. Time series analysis

Time series was prepared by calculating average of regions per time period (month) for New housing price index. The plot of the time series looks as follows:



4.3.1. ARIMA

Augmented Dickey Fuller test was performed for the time series, the results are as follows:

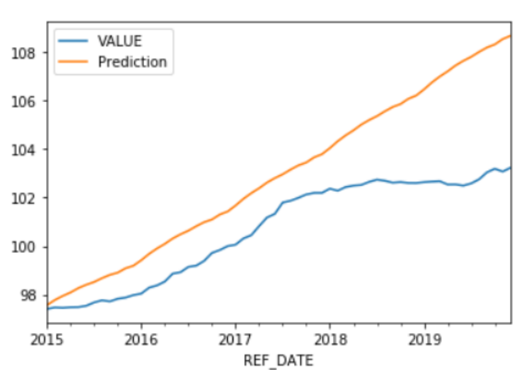
```
ADF Statistic: -0.085376
p-value: 0.950876
```

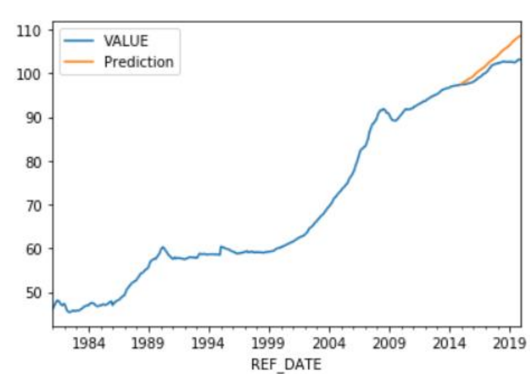
P-value is greater than the significance level (*α = 0.05*), so null hypothesis that the time series has a unit root and is non stationary cannot be rejected.

Auto ARIMA stepwise model is used. Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) are two different model selection criteria, **auto_arima** model used selects fit parameters based on AIC, although it also shows BIC score. Best fit parameters are identified for lowest value of AIC = -139.373.

```
Fit ARIMA: (1, 1, 1)x(1, 1, 1, 12) (constant=True); AIC=-139.373, BIC=-114.652, Time=8.027 seconds
```

Data is split to train and test sets, test set contains last 60 periods (5 years). Predictions are calculated. The value vs prediction charts look as follows:





It can be observed that the prediction value is always higher than the actual value, for the proposed model. Measures of the model performance are as follows:

```
The Mean Squared Error of ARIMA forecasts is 7.31
The Root Mean Squared Error of ARIMA forecasts is 2.7
```

```
Mean absolute percentage error is 2.22%
```

4.3.2. Random forest

Same as for ARIMA model, 60 periods set to evaluate the random forest model. **RandomForestRegressor** model was used. Predictions for the test period are flat (presented below).

```
array([97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303,
       97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303,
       97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303,
       97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303,
       97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303,
       97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303,
       97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303,
       97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303,
       97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303,
       97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303,
       97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303,
       97.3830303, 97.3830303, 97.3830303, 97.3830303, 97.3830303])
```

Measures of the model performance are as follows:

```
The Mean Squared Error of random forest forecasts is 15.09
The Root Mean Squared Error of random forest forecasts is 3.88
```

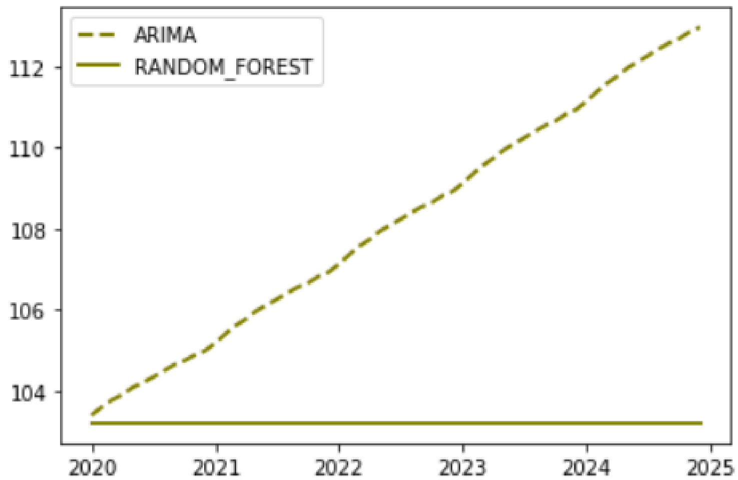
```
Mean absolute percentage error is 3.23%
```

The reason for the model not prediction correctly (making flat predictions) is because it can't predict outside the range of data it haven't seen. To solve this, predictions should not be made based on time series component (using other variables in multi-variate analysis) or, for univariate time series, pre-processing would be necessary to make time series stationary, including:

- Statistical transformations (Box-Cox transform, log transform, etc.)
- Detrending (differencing, STL, SEATS, etc.)
- Time Delay Embedding
- Feature engineering (lags, rolling statistics, Fourier terms, time dummies, etc.) (Tilgner, 2019)

4.3.3. Model comparison

Future predictions (next 5 years) produced by both models look as follows:



Below table shows the summary of model performance metrics:

Model	MSE	RMSE	MAPE
ARIMA	7.31	2.7	2.22%
Random forest	15.09	3.88	3.23%

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Mean absolute percentage error (MAPE) presents the normalized value of the error (Rosemarin, 2018). MAPE is the most widely used evaluation measure, however it can be biased (when used to select among competing prediction methods it systematically selects those whose predictions are too low) (Tofallis, 2015).

Performance measures are better for ARIMA model, however both models are not satisfactory – ARIMA was predicting all values above the actual values for the test data set. Predictions for random forest are flat and, it can be assumed, below the actual values. Although the error values for both models are low, the models are not predicting correctly and should not be implemented.

4.4. Summary and conclusions

It has been observed that New housing price index and average weekly earnings are strongly correlated.

Out of two developed models (ARIMA and decision trees), the ARIMA model shows smaller error on predictions. ARIMA model has predicted the value of New housing price index to be 104.98 in December 2020. As the value was 103.23 in December 2019, which stands for 1.695% annual increase. RBC report predicted 4.2% annual increase. Predictions made by the developed model do not seem satisfactory to support successful decision making.

5. References

Brownlee, J. (2019, 8 12). *Supervised and Unsupervised Machine Learning Algorithms*. Retrieved from Machine Learning Mystery: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

Geng, N. (2018). Fundamental Drivers of House Prices. International Monetary Fund. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=32&cad=rja&uact=8&ved=2ahUKEwiD_NPk89TnAhULX80KHdH1AgQ4HhAWMAF6BAgKEAE&url=https%3A%2F%2Fwww.imf.org%2F~%2Fmedia%2FFiles%2FPublications%2FWP%2F2018%2Fwp18164.ashx&usg=AOvVaw28OZ-461NgES5t6PDKlI

Github. (n.d.). *Forecasting at scale*. Retrieved from Facebook open source: https://facebook.github.io/prophet/

Government of Canada. (2019, 06 18). *Open Government Licence - Canada*. Retrieved from Open Canada: https://open.canada.ca/en/open-government-licence-canada

Khan, M., & Webley, T. (2019, 12). *Disentangling the Factors Driving Housing Resales.* Retrieved from Bank of Canada: https://www.bankofcanada.ca/wp-content/uploads/2019/04/san2019-12.pdf

Kim, E. (2019, 08). *Predicting House Prices with Machine Learning*. Retrieved from Kaggle: https://www.kaggle.com/erick5/predicting-house-prices-with-machine-learning

Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics.* New Jersey: Wiley.

NIST/SEMATECH. (2013, 10 30). *Introduction to Time Series Analysis*. Retrieved from Engineering Statistics Handbook: https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, É. (2011, 10 12). *Tuning the hyper-parameters of an estimator*. Retrieved from Scikit-learn: Machine Learning in Python: https://scikit-learn.org/stable/modules/grid_search.html

Peixerio, M. (2019, 08 07). *Project 1 - Predicting stock price.* Retrieved from Towards Data Scence: https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775

Peixerio, M. (2019, 08 07). *Project 2 - Predict air quality with Prophet.* Retrieved from Towards Data Science: https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775

Peixerio, M. (2019, 08 07). *The Complete Guide to Time Series Analysis and Forecasting*. Retrieved from Towards Data Science: https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775

Raghavan, S. (2017, 06 17). *Create a model to predict house prices using Python*. Retrieved from Towards Data Science: https://towardsdatascience.com/create-a-model-to-predict-house-prices-using-python-d34fe8fad88f

Roman, V. (2019, 01 20). *Machine Learning Project: Predicting Boston House Prices With Regression*. Retrieved from Towards Data Science: https://towardsdatascience.com/machine-learning-project-predicting-boston-house-prices-with-regression-b4e47493633d

Rosemarin, R. (2018, 01 13). *Why we use Root mean square error (RMSE), Mean absolute and mean absolute percent errors for forecasting time series models?* Retrieved from Quora: https://www.quora.com/Why-we-use-Root-mean-square-error-RMSE-Mean-absolute-and-mean-absolute-percent-errors-for-forecasting-time-series-models

Royal Bank of Canada. (2019, 12). *Home resale and price forecast.* Retrieved from Canadian Housing Reports: http://www.rbc.com/economics/economic-data/pdf/home-resale-fcst_can.pdf

Royal Bank of Canada. (2020, 02 07). *Are Toronto home prices sky-bound again?* Retrieved from Canadian Housing Reports: https://royal-bank-of-canada-2124.docs.contently.com/v/housing-market-commentary-february-2020?utm_medium=internal&utm_source=website&utm_campaign=housing

Statistics Canada. (2019, 12 20). *Employment and average weekly earnings (including overtime) for all employees by province and territory, monthly, seasonally adjusted*. Retrieved from Open Canada: https://open.canada.ca/data/en/dataset/8dccf1db-a127-45f7-9aea-0acdc5d19cc9

Statistics Canada. (2019, 11 03). *Full Table Download (CSV) User Guide*. Retrieved from Open Canada: https://www.statcan.gc.ca/eng/developers/csv/user-guide

Statistics Canada. (2019, 12 20). *New housing price index, monthly*. Retrieved from Open Canada: https://open.canada.ca/data/en/dataset/324befd1-893b-42e6-bece-6d30af3dd9f1

Statistics Canada. (2020, 02 09). *Canada Mortgage and Housing Corporation, vacancy rates, apartment structures of six units and over, privately initiated in census metropolitan areas*. Retrieved from Open Canada: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410012701

Statistics Canada. (2020, 01 21). *New Housing Price Index (NHPI)*. Retrieved from Statistics Canada: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=2310

Tilgner, M. (2019, 09 25). *Time series forecasting with random forest*. Retrieved from Startworx: https://www.statworx.com/at/blog/time-series-forecasting-with-random-forest/

Tofallis, C. (2015). A better measure of relative prediction accuracy. *Journal of the Operational Research Society*, 66. Retrieved from http://vuh-la-risprt.herts.ac.uk/portal/files/9313801/A_Better_Measure_of_Relative_prediction_accuracy_for_model_selection_and_fitting_Preprint_JORS_.pdf

Wikipedia. (2020, 02 01). *Canadian property bubble.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Canadian_property_bubble

Wikipedia. (2020, 02 20). *Positive-definite kernel*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Positive-definite_kernel