PREDICTING FOOD PRICES WITH PYTHON

Abstract

Food and shelter are the primary expenditures of Canadian households and due to rising housing costs more Canadians are becoming food insecure leading to rising food bank visits. Based on the work done for Canada Food Price Report of 2017 this research project attempts to use historical econometric and financial data to predict the future of food products contained in the Canadian Consumer Price Index using Python Sci-kit Learn linear regression models coupled with mutual information feature selection.

Patrick Walter
pt365049@dal.ca
Dalhousie University

Table of Contents

Introduction

Consumer Price Index

The Canadian Consumer Price Index is an indicator of the changes in prices experienced by Canadians obtained by comparing the rising or falling cost of a fixed basket of good and services purchased by Canadians. A fixed basket means that it contains goods and services of the same quantity and quality and therefore the changes in cost are a true reflection of the pure price changes (Statistics Canada, 2018). The Consumer Price Index is released monthly for each province and is divided into eight categories which include: food, shelter, household operations and furnishings, clothing and footwear, transportation, health and personal care, recreation, education and reading, and alcoholic beverages and tobacco products. There are also three special aggregates which include: all items excluding food, all items excluding energy, and energy. These categories are then divided into subcategories which are then divided into even more refined values, some of which represent specific foods such the different meats, seafood, and dairy products.

The main purpose of the CPI is to be an indicator of the changes in the level of consumer prices which reflects the rate of inflation. Since inflation is a direct factor of purchasing power the CPI is useful to all Canadians who can compare the changes in their own personal income with that seen by the CPI to assess their personal financial situation. Most governments and economists believe that a small positive value for inflation is optimal for allowing producers to increase the quality of products and services (Harris, 2017).

Canada Food Price Report

Beginning in 2010, The Canada Food Price Report has set out to be a tool used to forecast Canadian food prices and focus on the factors that will be affecting the future of consumer food prices over the following one-year period. The report began at University of Guelph by Dr. Sylvain Charlebois and Dr. Francis Tapon and is now a publication by Dalhousie University. The report combines information from several food-related reports and data to identify the key fundamental drivers that will impact food prices for the year ahead. These drivers are categorized into macro, sectorial, and domestic (Charlebois, et al., 2017).

In 2017, the report employed a machine learning model to supplement the panel of domain experts' advice (Charlebois, et al., 2017). It leveraged a combination of different machine learning methodologies including linear regression and support vector machines to forecast components of the Canadian Consumer Price Index. This was built on a research thesis done by Jabez (Jay) Harris which set out to use compare different machine learning algorithms ability to make econometric models to forecast major food group categories listed in the Canada Consumer Price Index against benchmark models commonly used in financial and econometric forecasting (Harris, 2017). In the report there were over twenty independent variables identified as potential inputs to the machine learning models and of these only ones that were highly correlated with a food categories price were used. These independent variables included household income, immigrant income, income distribution, international aid, population, unemployment, commodity futures, fuel prices, crude oil prices, energy indexes, CDN exchange rate, U.S. overnight lending rates, global agricultural production, global rainfall, commodity prices and global temperatures (Charlebois, et al., 2017).

Research Objectives

The main objective of this research was to attempt to predict the overall food category of the CPI using a linear regression model implemented in python using Sci-kit Learn. This lead to subsequent experiments conducted on how to apply feature selection using mutual information to reduce the dimensionality of the model to make more accurate predictions. To test this implementation, individual models were constructed to predict the monthly average food price for the twelve-month period of each year between 2012 to 2016. The target values for these predictions were sourced from Statistics Canada and the attributes dataset was a subset of the dataset produced by Jay Harris for research for the Canada Food Price Report 2017. Secondly this research attempted to predict twenty-one other subcategories of the CPI for the years of 2016 and 2017. In this experiment the goal was to identify the specific subcategories of the food price that were most difficult to predict accurately with this dataset using feature selection for a linear regression model.

Research Methodology

Dataset

The dataset of this work was adopted from the work done by Jay Harris in a thesis titled A Machine Learning Approach to Forecasting Consumer Food Prices. The targets dataset included twenty-two different targets of which seventeen were food related products. Some targets are specific food products such as eggs and coffee, while other are aggregates such as food from restaurants and food from stores. All the targets are part of the CPI produced by Statistics Canada. This historic dataset dates to January of 1985 until the most recently released month, which at time of writing is February 2018.

The attributes dataset contains 280 independent features. This is significantly less than the dataset used in the previous thesis which consisted of 476 features before feature selection. The starting record for this dataset is also January 1985, however not all features have values in the early years. The dataset has 391 records which conclude at July of 2017. However, not all features have values for records near the end date.

In the Canada Food Price Report 2017 the historical data was capped at January of 1999. This decision was two-fold. Firstly, as mentioned some of the features do not have records for earlier years, and those that do may not be reliable. Secondly in 1997 the Canada Harmonized Sales Tax was implemented which had an effect of consumer prices in Atlantic provinces (Harris, 2017). In this work the entire historic data set was used for feature selection but like the Food Report only records after January 1999 were used in modeling.

Machine Learning in Python

Scikit-learn is a machine learning library for the python programming languages which includes many implementations of popular machine learning algorithms including regression. Scikit-learn was used for it's machine learning and modeling capabilities for this research. Other libraries used in this research include pandas and matplotlib. Pandas is an open source data structures and data analysis library used for data wrangling and structuring (Pandas, 2017). Matplotlib is a 2D plotting library used for data visualization (Hunter, Dale, Firing, & Droettboom, 2012).

Data Preprocessing

The original dataset contained all features and targets together in a comma-separated values formatted file. This was divided into two comma-separated values files, one for the features and one for the targets. The timestamp feature was copied to both files. Panda's dataframes were used to structure the datasets for modeling. A dataframe is two-dimensional tabular data structure with labeled axes which will allow for the time series of each feature and target to be selected by name. (Pandas, 2017) The two CSVs containing the attributes and features datasets were imported into two dataframes for their entire matrices. These dataframes contained records beginning at January 1985. Reduced versions of these dataframes were constructed beginning at January 1999. The complete dataframes were to be used by feature selection, while the reduced dataframes were to be used for constructing the linear regression models.

As mentioned previously, some of the records had missing values for some of there features in later years. To deal with missing values in records, any feature that had missing values after January 1999 was removed. This was a simple solution but only affected a few features. Other solutions such as predicting future values using linear regression or averaging were discussed for the future experiments.

Further data preprocessing was done using these four dataframes for each of the two experiments conducted in this research project. The first was to divide the attributes dataframes into test and train sets for different years. The second was to divide the targets dataframe into individual time series for each of the targets. These processes will be described in detail in the experiment section of this report.

Linear Regression

Linear regression is a method of predicting a target value which is expected to be a linear combination of weighted input variables where the weights are represented by coefficients and an intercept. The ordinary least squares implementation of linear regression creates a linear model with a vector of weights or coefficients that minimizes the residual sum of squares between targets in the dataset and predicted targets (Scikit Learn, 2017). In mathematical notation the predicted target ý:

$$\hat{y}(w, x) = w_0 + w_1 x_1 + ... + w_p x_p$$

Feature Selection

In this project an implementation of mutual information feature selection was used to select a set number of features from the set features. Mutual information measures the dependency of two variables and gives a non-negative value. Scikit-learn has implementation of mutual information for feature selection for both regression and classification. The regression mutual information feature selection uses an estimate of for continuous target variable (Scikit Learn, 2017). This implementation uses mutual information estimation based on work by Alexander Kraskov, Harald Stögbauer, and Peter

Grassberger which uses entropy estimates from k-nearest neighbor distances to estimate mutual information (Kraskov, Stögbauer, & Grassberger, 2004). Unlike other estimates of independence or correlation between variables, mutual information also picks up dependences that are not seen in linear correlation coefficients because they are not shown in covariance.

K Best

In this project mutual information feature selection was used as a scoring function. Each of the features was given a score based on its mutual information score with the target variable. These scores are then used by the SelectKBest feature selection function. This takes the top *k* scoring features, where *k* is some integer between one and the total number of features in the dataset to select from. Finding the optimal value for *k* was part of the experimentation of this project. The scoring can be based on different methods of evaluating correlation and in this research mutual information was used.

Experimentation

Predicting the Average Food Price for a Five-Year Period

The first objective of the experimentation in this research project was to predict twelve-month period of an aggregate target labelled *Food17*, which includes all food products captured in the Canada Consumer Price Index, for each year in a five-year period between 2012 and 2016. These years were selected as they were the last five years with complete twelve-month records in the attributes dataframe.

To achieve this, the attributes dataframe was divided into a testing set and training set for each of the fives years in the five-year period. This was done for each year as follows. The training set contained the historical data up until December of the year before the target year while the testing set contained the twelve-month records of the target year. For example, the training set for the year 2016 was from January 1999 to December 2015, and the training set contained the twelve months from January 2016 to December 2016. This was done for each of the five years to produce ten dataframes, five training sets, and five testing sets.

Feature selection was conducted using the entire attributes dataframe and the *Food17* target dataframe. This used K-Best selection based on mutual information scores. Values in the range of 5 to 35 were used for the value of *k.* For some years, such as 2013 and 2014, values of *k* less than 10 yielded the best results in terms of mean absolute error and variance scores. However, for other years, specifically for 2016, much greater number of features were needed to make accurate predictions. In summary, for higher values of *k* years that preformed well with few features preformed worst however years that were harder to predict gave better results. What was different about the year 2016 was that it was the only year in the five-year period in which the aggregate food price target had a downward trend. Based on these observations, further work on how to predict irregular years, or crashes, were proposed.

Results from this experiment were varied by year, however each model was able to predict the twelve-month period of the average food price target *Food17* with mean absolute errors between zero and three.

Predicting Each Individual Target

Next experimentation was to predict each of the other twenty-two target attributes included in the attributes dataset. The goal here is to shed light on which products prices can be predicted accurately using features selected from the total attributes dataset. In this experiment the process of feature selection was done using mutual information of each of the random variables and the different target variable.

This experiment was first preformed for the year of 2017 after the target dataset was expanded using the up-to-date data from Statistics Canada. However, the much larger attributes dataset was not able to be updated. The attributes data of 2016 was used to attempt to predict each of the individual targets for the 2017 year. This might be more representative of real world application when a future year is being forecasted based on values from this last year.

The targets dataframe was divided into one dimensional time series for each of the twenty-two targets. These individual targets were then used as the target for different experiments. The K-Best feature selection using mutual information scores was again used. This time each experiment had it's own feature selection process with the different targets. For each of the targets the mutual information score was computed between that target and each of the attributes in the attributes dataframe. The K-Best selector was used again to select the top *k* scoring features. For this experiment *k* was set to 10 and was increased for targets that had poor accuracy.

A linear regression model was constructed for each of the targets with the test set being all records between January 1999 and December 2016 using the features selected by the feature selection process. Some observations include that many of same features were selected for each of the very different individual and aggregate targets. The aggregate targets were easier to be predicted with less features, where as individual products such as eggs and chicken. Some targets were able to be predicted accurately using 2016 attributes data while other had large variance between the predicted values and the actual target values.

This same experiment was repeated for the year 2016. Results for 2016 were significantly worst than those of 2017 which is again attributed to the downward trend of many of the targets that year. Again, a solution to this problem was to raise the value of *k* to attempt to capture the downward trend of the year.

Further Research

Future Research Objectives

The main objectives of the future research of this project are to get better accuracy for the average food price target and the twenty-one other targets for the years of 2016 and 2017. Further research in how to better optimize the feature selection process for each year and how to better evaluate the model based on the features selected. Other scoring methods for evaluating the correlation between a feature and target will be explored and how to better optimize the number of features selected for each target in a model will be further researched.

A key observation from this research was that some years which seem for follow a normal behavior are easy to predict using a few features and a short history of data while other years in which anomalies happen a much greater number of features are needed to get accurate forecasts. Discussion on how to use the past anomaly years, or crashes, could lead to developments in implementing a model to keep track of the features used to predict those years.

Conclusion

Every month Statistics Canada releases aggregate data for prices of a fixed basket of goods and services for every Canadian province. A large portion of this basket contains food related products and services and their price changes are captured monthly. This research set out to develop an implementation of linear regression modelling for the Canadian Consumer Price Index components related to food. The main target for modeling was the overall food price which captured all food related products and services in the basket. Attempting forecast this indicator was an attempt to forecast real changes in food prices as they are seen by Canadian consumers.

The models developed in this research deployed linear regression models implemented using Python Programming Languages and a machine learning library called Sci-Kit Learn. The models were developed using a large dataset of attributes sourced from a thesis titled A Machine Learning Approach to Forecasting Consumer Food Prices by Jay Harris at Dalhousie University. Feature selection using sci-kit implementation of mutual information as a scoring method was used to reduce the dimensionality of the model to increase the accuracy of the predictions. Further work on this project will attempt to increase the accuracy of the models and better predict the average food price and prices of other individual and aggregate food related targets in the basket.

References

Charlebois, S., Jabez, H., Tyedmers, P., Bailey, M., Keselj, V., Conrad, C., . . . Chamberlain, S. (2017). *Canada's Food Price Report.* Halifax: Dalhousie University.

Harris, J. (2017, August). A Machine Learning Approach to Forecasting Consumer Food Prices. Halifax, Nova Scotia, Canada.

Hunter, J., Dale, D., Firing, E., & Droettboom, M. (2012). *Matplotlib*. Retrieved from Matplotlib: https://matplotlib.org/

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating Mutual Information. *Physical Review, E*(69).

Pandas. (2017, December). *Python Data Analysis Library*. Retrieved from Pandas: https://pandas.pydata.org/

Scikit Learn. (2017). *Generalized Linear Models*. Retrieved from Scikit Learn: http://scikit-learn.org/stable/modules/linear_model.html

Scikit Learn. (2017). *sklearn.feature_selection.mutual_info_regression*. Retrieved from Scikit Learn: http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html#sklearn.feature_selection.mutual_info_regression

Statistics Canada. (2018, March). *Consumer Price Index (CPI)*. Retrieved from Statistics Cnada: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=2301