ISSN 1870-4069

Price Forecasting: A Data Science Approach

Paul Ricardo Millán González, Félix Mata

Instituto Politécnico Nacional, UPIITA, SEPI,
Mexico City, Mexico

luapmg@outlook.es, mmatar@ipn.mx

Abstract. The research presents a data science approach to analyze prices of products in supermarkets to make price forecasting and discover patterns of behavior in the prices. The dataset comprises a period from 2011 to 2014. The case of study selected in this research is, a price forecasting experiment analyzing data of tuna product in Mexican republic. Data were processed and classified using the methods of linear regression, polynomial regression, regression support vector and neural networks. This research provides preliminary results as a first advance of the approach.

Keywords: Machine learning, data science, prediction.

1 Introduction

Data Analysis is used in this research as an approach to work with inference to derive the conclusion based on previous experience of data behavior. It means find or define mathematical models and algorithms from the data.

The family economy is very important, hence is important to be informed about the behavior prices of basic products principally, such as food, home cleaning products and health care products. Currently exist a public database offered by PROFECO [3] with more than 27 million of records of different products offered in distinct establishments and commercial chains. This database includes attributes like product, presentation, brand, state, municipality, address, date and price.

This research will focus on to analyze the behavior of price for product tuna in Mexican Republic. Our hypothesis is based on the assumption that is possible to create a mathematical model that describe the price behavior of it over the time, it will be achieved using techniques of Machine Learning. This mathematical model can be useful to forecast or discover patterns regarding to price behavior for certain products in the basket basic.

As a first step, it was considered that in a small data sample will be enough to find a model. It was based in a previous analysis when the data was inspected followed the approach of data mining, using tools like RapidMinner (https://rapidminer.com/). It was found, for example, some cases where two supermarkets of the same branch and even in the same city and at the same time presented different prices for a particular product. Therefore, it was thought that every establishment could had your own behavior, and this would add complexity to the experiment because although the database contains

thousands of records these are distributed around all Mexico's territory. Nevertheless, when data were fragmented by city or state the size is decreased, then it was not possible to use them for finding a model.

The rest of the paper is organized as follows: Section 2 describe the related work, in section 3 the methodology used is explained, in section 4 the experiments are done, in section 5 the results are shown and section 6 outlines the conclusions.

2 Related Work

There is a large amount of forms in forecasting economy indices such as stock prices [1], in which the methods used are time series analysis, and the majority are focused on structured data (e.g. stock prices table) [2,3].

Many researchers are doing about the use of machine learning for forecasting, many areas of study can be benefited with this approach. In business it is possible to predict with an acceptable range of error the price behavior of one particular product. Several works use approaches based on machine learning, concretely, it is a matter of creating programs capable of generalizing behaviors from information provided in the form of examples. It is, therefore, a process of induction of knowledge [1].

Other related work is [4] focused on predict the price of natural gas, but not using a data science approach like in our work. While, other works like [5] where is compared support vector machines for regression with Back propagation and RBF networks in stock price data. Finally, the main motivation of our research is the possibility to help, for consumers, companies and government, in the creation of strategies of purchase or public policies.

3 Methodology

The methodology defined consist of three experiments: 1) A Small DataSet (only canned tuna), 2) A Big DataSet with tuna per kilogram and 3) A Big DataSet (only canned tuna).

In the case of experiment 1) with small dataset. The sample was 101 records of the branch Tuny, 140grams in oil, for store Soriana Branch, Mexico City, in Azcapotzalco from 2011 to 2013. The only components in this stage of the study were the date and the price, the date attribute was treated in different ways, first it was transformed the date that came in chain format in three integer values day, month and year, after being transformed each one of these variables in ranges, then the whole date was transformed into rank, finally, the date was transformed into a special continuous float value for regression analysis. The different ways in which the date attribute was transformed did not generate variation in the effectiveness of the model therefore we decided to use the date as a continuous float value as it gives us the advantage that we can print the data easily.

For the case of experiment 2) and 3) big dataset. The sample was 169469 records; this sample data was used to find an acceptable model and a good performance when machine learning algorithms were applied. Data was raised from distinct states and cities around Mexico. The classification consisted of groups such as: state, municipality

and commercial chain. Combinations were performed only with those have more than 100 records. There are two Big datasets, one with canned tuna and tuna per kilograms then this dataset has low prices (canned tuna) and large prices (tuna per kilogram), and another with only canned tuna is that the reason that this only has lower prices.

4 Machine Learning

4.1 Linear Regression with Small DataSet

It is beginning by training a linear regression model: The method used for regression is expected to be a linear combination of the input variables.

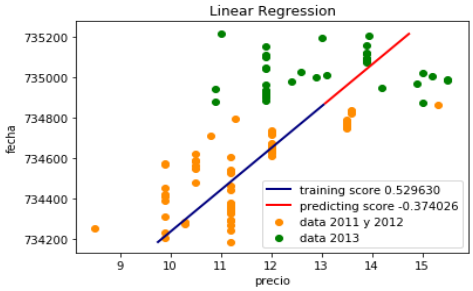

Fig.1. Lineal Regression.

In Fig. 1 it is possible to see a model greeted by lineal regression nevertheless this model does not represent well the behavior of the sample.

4.2 Polynomial Regression with Small DataSet

A new feature matrix is generated consisting of all polynomial combinations of entities with a degree less than or equal to the specified degree.

In Fig. 2 the models represent well the sample, but it is only at the training. While in Fig. 3 can be appreciated that these models do not predict correctly.

4.3 SVR with Small DataSet

Given a set of points, a subset of a larger set (space), each of which belongs to one of two possible categories, an algorithm based on SVM constructs a model capable of predicting whether a new point (whose category we do not know) belongs to one category or the other. The SVM looks for a hyper plane that optimally separates the points of a class from the one of another, that could have been previously projected to a space of superior dimensionality. The support vector machines for regression is a robust technique for function approximation. [5].

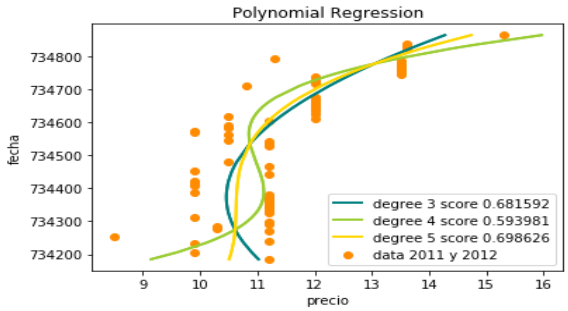

Fig. 2. Polynomial Regression Training.

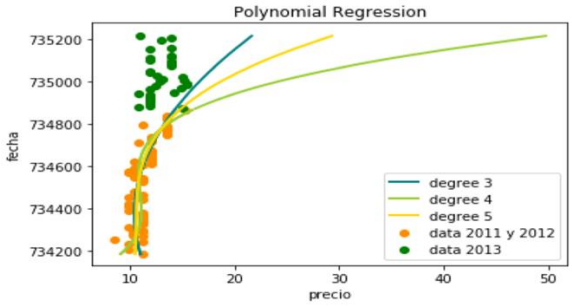

Fig. 3. Polynomial Regression Predicted.

At the moment the algorithm SVR is the one that appears to have the best behavior, but it is not enough. Several tests were done modifying the values of gamma, constant C and epsilon. Fig. 4 shows one of the best results, which as we can see has a negative sign which is very bad since this means that the model does not represent the behavior of the data. From this moment it was considered that the sample was too small and it was proposed to make tests with more data. The Fig. 4 show the results obtained by linear regression and SVR.

Fig. 4. Price forecasting over time.

4.4 Neural Networks with Big DataSet

Broadly speaking, two views exist between practitioners/investors who typically prefer a small in-sample to minimize data holding requirements and researchers/academics who typically chose large in-sample periods [6].

In this stage, the data sample contains 169469 records, it was grouped by state, municipality and commercial chain and took only those that combinations that had more than 100 records. This in reason that was considered that those combinations that had less of 100 could decrease the performance of training as there were not enough samples. The experiment was performed using Microsoft Azure. In Fig. 5 it can be seen the behavior of the Big DataSet with tuna per kilogram over the time. When we talk about the complete sample we refer to the first sample that was selected, which included tuna per kg. Thus, we have low prices of canned tuna and high prices of tuna per kilogram.

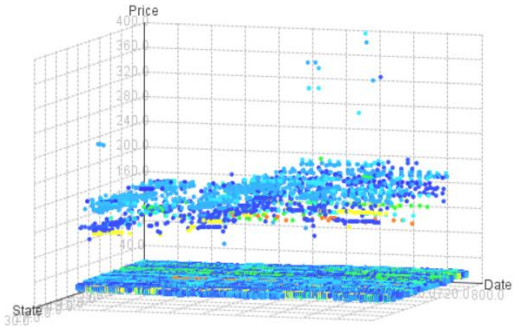

Fig. 5. Big DataSet with tuna per kilogram.

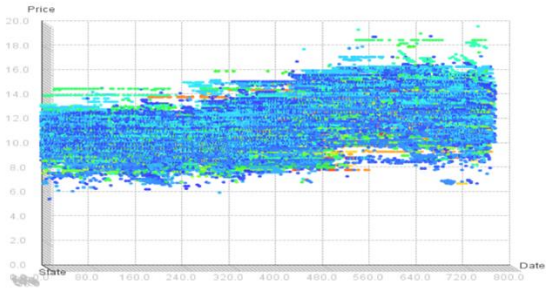

Fig. 6. Big DataSet (only canned tuna).

It was used neural networks because at the beginning it was observed that this algorithm presented the best results. Therefore, the research is centered hereafter applying neural networks. In this sample we have the following attributes: presentation, brand, commercial chain, state, municipality, address all these are string. Date, price: double.

From the previous attributes new attributes were made as day: integer, month: integer, year: integer, day_week: string, month_chain: string. It was tried to use one hot code to binarize the text columns presentation, state, municipality, address, however many columns were created in the process, which complicated the treatment of the sample, therefore this option was declined.

The parameters of Neural Network Regression module were those of default. Create trainer mode: single parameter, Hidden layer specification: fully-connected Number of

hidden nodes: 100, Learning rate: 0.005, Number of learning iterations: 100, The initial learning iterations: 100, The initial learning weights diameter: 0.1, the momentum: 0, The type of normalizer: Min-Max normalizer.

Now there are obvious questions like what would happen if large prices (prices of tuna per kilogram) were removed? And is valid keep them? are these samples helping at the training really? or it is just a coincidence and these samples only confuse the score of performance in the training stage?

Now, large prices (prices of tuna per kilogram) will be removed, and the training will be done one more time, the same columns will be used and the same parameters of the algorithm. In the Fig. 7 is the price behavior over the time of the new sample without large prices. In this figure it can be seen a tridimensional plot of this sample on time, state and price, the colors represent chains of store.

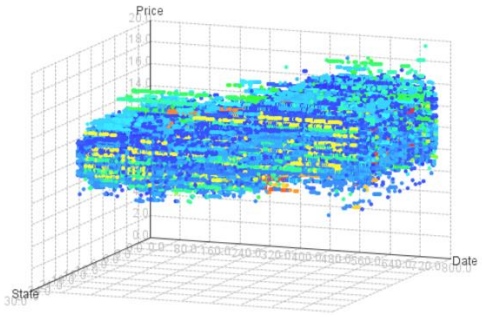

Fig. 7. Big DataSet (only canned tuna) on Date, State and Price.

5 Results

5.1 Results with Big DataSet with Tuna per Kilograms

These are the results of the sample with large prices (with tuna per kilograms), the output provided by Azure with Big DataSet with tuna per kilograms are the followings: Mean Absolute Error = 0.996136, Root Mean Squared Error = 1.727647, Relative Absolute Error = 0.375668, Relative Squared Error = 0.019779, Coefficient of Determination = 0.980221. The Mean Absolute Percentage Error (MAPE) is 7.75%. The maximum error is 201.27% The number of predictions with error above 20% is 1047 of 16947 so 6.17% is outside a range of error that can be considered acceptable.

Fig. 8 Price Behavior Over the Time for Big DataSet with tuna per kilogram.

The Fig. 9 above show the real price behavior over the time against the model provided from our methodology using Neural Network Regression. To help the reader to see better, a zoom is shown in the Figs 10 and 11.

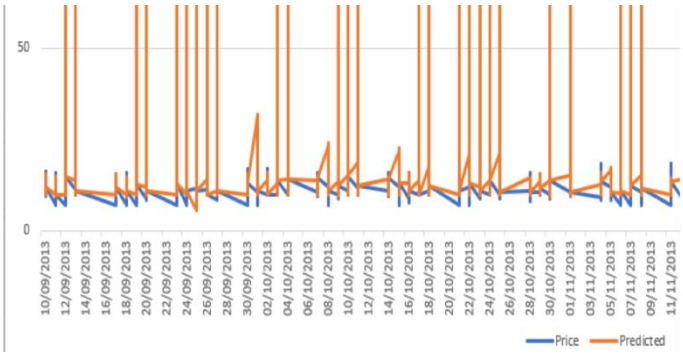

Fig. 9 Zoom of graph in Fig. 8.

The Fig. 9 above is the first part of the zoon of the Fig. 9, here is easier to appreciate the difference between the real price and the predicted price.

5.2 Results with Big DataSet (only canned tuna)

This is the sample without large pri, the result when the large prices of tuna per kilogram are put away is that the performance decline completely. The results provided by Azure for Big DataSet (only canned tuna) are the followings: Mean Absolute Error = 1.000164, Root Mean Squared Error = 1.394822, Relative Absolute Error = 0.707804, Relative Squared Error = 0.71398, Coefficient of Determination = 0.28602. The mean absolute percentage error (MAPE) is 11.52%. The maximum error is 83.01%. The number predictions with error above than 20% are 3023 of 16699 so 18.10% is outside a range of error that can be considered acceptable, which is threefold of the previous result.

Fig. 10. Price Behavior Over the Time of Big DataSet (only canned tuna).

6 Conclusions

Is interesting to see that the large prices of tuna per kilograms can help to obtain better results on prediction, though this is on discussion because it produced bigger errors too.

In my opinion if it is possible to improve the result of predictions using cross data, it would be valid to use them. Many relations can exist between our data and other information available outside. In this case is obviously that exist a relation between tuna and canned tuna, but the relation seems be linear, so is interesting to see how much seem to help at the training more data that seem to behave same, but in distinct range of prices.

More tests are necessary to be able to lower the mean percentage error(MPE) and thus make better predictions. We believe that the model can be improved in such a way that the predicted behavior will be closer to real, there is the possibility that it could be achieved with more samples if we could get an upgrade of our dataset that contain more years or by adding data from other sources of information.

Both tests in the second stage offer good results, that one is better than another depends on the context where you want to apply.

The opinion of the author of this paper is that the experiment of all presentations of tuna with large and low prices together yield better results, because although it has bigger mistakes these are few and the most of predictions are close to real price. However, it is a preliminary study and therefore this is still a supposition, a much larger study is necessary to provide better results and to be able to sustain them.

References

1. Stock, J., Watson, M.: Forecasting output and inflation: The role of asset prices. Journal of Economic Literature, vol. 41, pp. 788–829 (2001)
2. Marcek, D.: Stock price forecasting: Statistical, classical and fuzzy neural network approach. In: V. Torra and Y. Narukawa (eds.), MDAI, ser. Lecture Notes in Computer Science, vol. 3131. Springer, pp. 41–48 (2004)
3. A hybrid ARIMA and support vector machines model in stock price forecasting, vol. 33, no. 3, (2005)
4. Dataset source: https://datos.gob.mx/busca/dataset/quien-es-quien-en-los-precios/resource/9fa38cc3-bcc6-4597-b240-263532532467 [Last access: 10 2017].
5. Čeperić, E., Žiković, S., Čeperić, V.: Short-Term Forecasting of Natural Gas Prices using Machine Learning and Feature Selection Algorithms. Energy (2017)
6. Trafalis, T.B., Ince, H.: Support Vector Machine for Regression And Applications to Financial Forecasting. In: IJCNN '00 Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00), volume 6 (2000)
7. Kambouroudis, D.S., McMillan, D.G.: Is there an ideal in sample length for forecasting volatility? Journal of International Financial Markets, Institutions and Money (2015)