# IBM Coursera

## Applied Data Science Capstone

2020

---

Predicting Inflation Report

---

Oscar Holguín

Date: 21 de octubre de 2020

# 1. Introduction

## 1.1. Background

Inflation is defined as the rise of prices of most goods and services of daily or common use. This services and goods include food, clothing, housing transport recreation among others []. The inflation index measures the change in prices of the basket of commodities and services over a period of time. This period of time is usually per month, or yearly but it can vary. When measuring inflation, you do so in percentage and it represents the measure of the purchasing power of a given population who uses a specific currency. The purchasing power of the population is indirectly proportional to inflation, thus when inflation rises it means people are able to afford less goods and services than before. One common problem in different countries is that sometimes inflation rises more than wage increments.

## 1.2. Problem

The main problem with the inflation index is that inflation is highly volatile and thus difficult to predict. Inflation usually gets most of the people unprepared and results in a painful strike to their economy. It is the aim of this project to try to use different machine learning approaches to choose the best model that is able to predict this index. This would help people to be prepared, or save money when inflation is about to rise.

## 1.3. Interest

This project would result interesting to the general public, since everyone wants to be prepared for inflation changes. Economists might also be interested in this approach for their specific applications. Lastly this will also be of special interest to the government because they can also be prepared and use this information to make new policies, or efficient decision making.

# 2. Data

## 2.1. Data Acquisition

For this project, the data will be of Mexico's inflation. The data considered will be mainly the national inflation and maybe some major cities or cities which inflation usually varies such as , Monterrey, Gudalajara or border city Cd. Juárez. The data is provided by an official institution called Banco de Mexico and INEGI (Instituto Nacional de Estadística y Geografía), both institutions provide this information, INGEI is the one currently in charge, but historical data is found in Banco de Mexico. A third institution, Banxico, gathers all the data provided by the aforementioned institutions and displays it in their official website, with a more understandable and clean way. Therefore in this project the data will be acquired from scraping `https://www.banxico.org.mx/tipcamb/main.do?page=inf&idioma=sp`

The data of inflation used, is called IPC (Indice de Precios al Consumidor) Index of Prices to the consumer. The IPC is calculated with the following equation.

$$IPC = \frac{NewPrices * NewProdcuts}{OldPrices * OldProducts} \tag{1}$$

The Inflation time series is monthly data from January 1969 until September 2020.

## 3. Methodology

The first step is to clean the data and pre-process it. We keep only the relevant columns in the dataframe which are Date and Inflation data. After doing so, the initial data is plotted to take a first look at the information

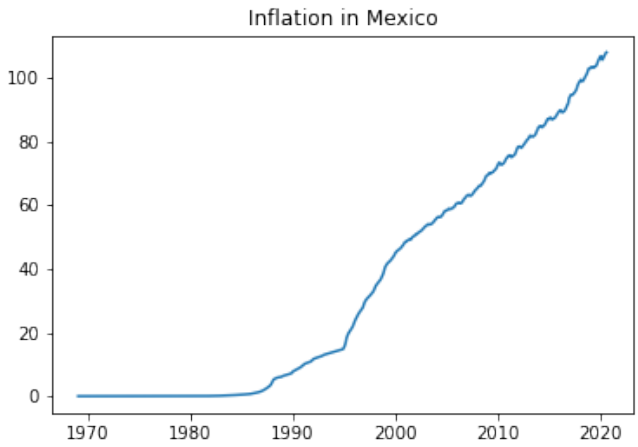


Figura 1: Inflación en México de 1969-2020

As we can see the inflation has a clearly tendency to rise over time, and no seasonality can be detected at first sight. The latter, can be deducted but it is better if tested to prove it. To do this, the KPSS Stationary Test is performed. In econometrics, Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests are used for testing if a time series is stationary around a deterministic trend (i.e. trend-stationary) in which the null hypothesis assumes that the series is stationary. The p value obtained by the test is 0, which is smaller than the assumption p is 0.05.

$$p = 0 < 0.05 \tag{2}$$

According to this test, stationary data can be assumed, if it resulted in non-stationary data differentiation will be made in order to perform ML models on this differentiated data.

### 3.1. Linear Regression

Linear Regression is a method that approaches the relationship between dependent and independent variables, linearly. In this project the linear regression function from scikit learn will be used. This library uses the ordinary least squares Linear Regression, which fits a linear model with coefficients to minimize the residual sum of squares between the targets in the dataset and those predicted by this approximation

First the data is splitted into test and train data, train data will be 80 % of the total and the rest, test data. The model is fitted with the train data and tested with the test data to evaluate the prediction. In this case the error of the predicted with respect to the test data, is evaluated with the RMSE (Root Mean Squared Error), which definition is the following equation:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2} \tag{3}$$

The lower this value is, the closer the predicted value is to the tested.

### 3.2. SGD

Stochastic gradient descent is an iterative method for optimizing a target function with specific properties (such as differentiable or subdifferentiable). It is basically a stochastic approach of a gradient descent since it uses a calculated randomly subset of data estimation instead of the gradient calculated by the dataset.

Stochastic Gradient Descent Regression in scikit learn library, implements the stochastic gradient descent learning routine to fit linear regression models. According to documentation this method is well suited for problems with a large number of training samples.

In this method data must be transformed, so after following the same first steps, splitting the data, one must scale the input before training and predicting it.

### 3.3. SVM

Support Vector Machine is a supervised learning model that analyzes the data used for regression analysis. It is one of the most robust prediction algorithms for datasets under 10,000 samples. In the scikit library some parameters can be modified, such as the kernel, which specifies the type of kernel to be used in the algorithm. In this project the kernels used were:

- Linear kernel
- Polynominal kernel
- RBF kernel

The other parameters that will be modified is the regularization parameter, which can be tuned with the C value. The strength of the regularization is inversely proportional to C. In this case different C values will be used to determine the best approach for each kernel. .

## 4. Results and Discussion

### 4.1. Linear Regression Results

In this case, the mean squared error MSE, was obtained MSE=127.09. To get the RMSE we simply get the squared root obtaining:

$$RMSE = \sqrt{MSE} = \sqrt{127.09} = 11.27 \tag{4}$$

When comparing the predicted values with the tested ones, the following plot is produced:

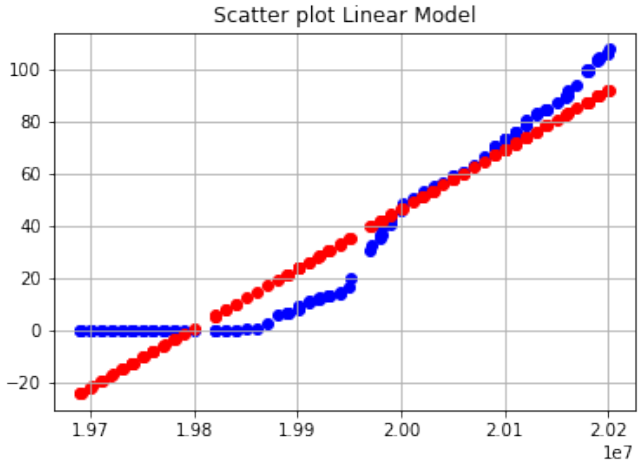


Figura 2: Linear Regression Model

As we can see the linear Regression is a good approach since the inflation data is pretty linearly behaved (Inflation rises over time constantly), but it can be better

## 4.2. SGD Results

For this method the MSE yields 115.17

$$RMSE = \sqrt{MSE} = \sqrt{115.17} = 10.73 \tag{5}$$

As we can see here SGD doesn't perform as well because it is well suited for data $\geq$ 10,000 values according to documentation. In this case we can see that SGD performs quite similar to Linear Regression, as a matter of fact it is just a little better according to the RMSE.

When plotting the values, the following figure is produced:

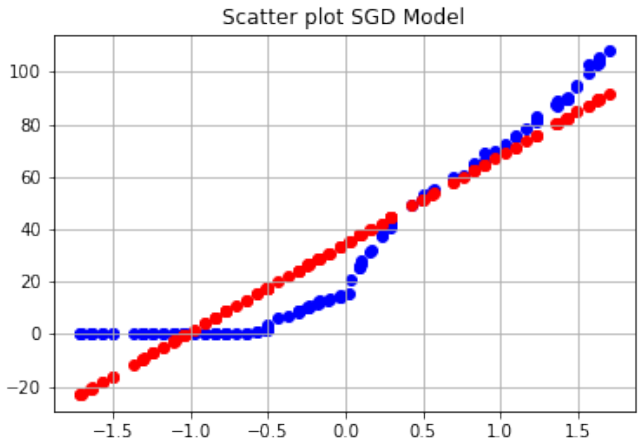


Figura 3: SGD predicted vs tested

### 4.3. SVM

Support Vector Machine was performed with different kernels, linear, polynominal and rbf kernel. In this case we got the following MSE's

$$RMSE_{linear} = \sqrt{324185489427984} =$$

(6)

$$\begin{aligned} RMSE_{poly} &= \sqrt{323.45} \\ &= 17.98 \end{aligned}$$

(7)

$$\begin{aligned} RMSE_{rbf} &= \sqrt{7.97} \\ &= 2.82 \end{aligned}$$

(8)

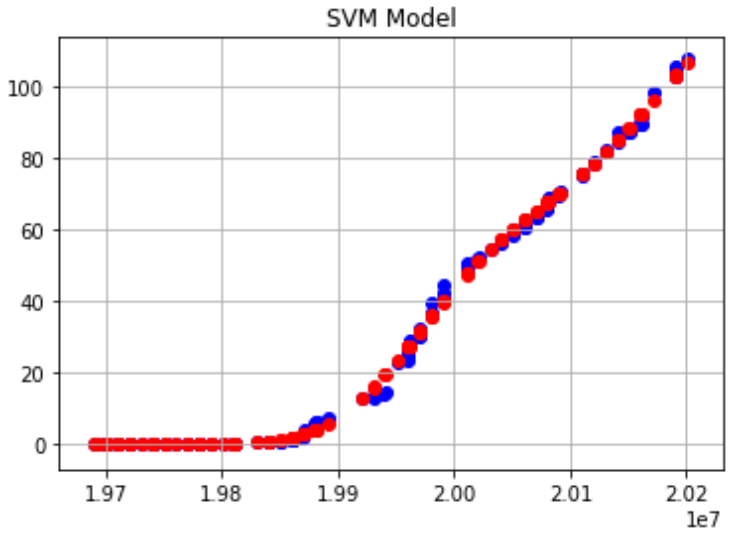


Figura 4: SVM RBF kernel predicted vs tested

It is relevant to note that the best was the rbf kernel with a C value of 100. Since this parameters can be modified, one might get a better prediction configuring the different available parameters

such as C, and $\gamma$ (which was set to "*scale*" in this case). This combinations yielded a good and very low RMSE values when compared with the other approaches. The plot also confirms that this method effectively predicted the tested data.

## 5. Conclusion

In conclusion even though data seemed pretty linear, it was not completely linear. The best model was found to be the Support Vector Machine but with a non linear Kernel (RBF). The parameters had to be tunned to get the most optimal result. It is remarkable that the parameters have a strong influence on the result. For future analysis it will be of special interest to predict other type of data or in a smaller time of window to predict non stationary or seasonal data. It is also necessary to try different algorithms such as the ridge model, decision tree, or even the SGD with different loss functions.