Predicting gasoline prices in Mexico

Francisco Agustín Díaz Vergara | A01204695

Abstract **–** This document provides an overview of a Gradient Descent algorithm implementation that was developed to perform a linear regression aimed at predicting gas prices in Mexico using location and year as input parameters. As well as an analysis of its performance based on training and test sets.

I. INTRODUCTION

The price of gasoline in Mexico is a topic of significant economic and political interest. According to a report by the British Broadcasting Corporation (BBC) gasoline prices increased in price by 48% during the government of former Mexican president Enrique Peña Nieto causing widespread public protests and debate over government policies and fuel markets [1]. I personally remember when gas prices went up in my hometown and how that presented a new challenge for many people I know personally. This is why I decided to develop a Gradient Descent algorithm for linear regression which helped me predict prices of gasoline based on location and year. It's important to note that this project is intended for education purposes and must not be used in any other environment.

II. STATE OF THE ART

The application of machine learning (ML) techniques to predict the prices of various products such as electricity and gasoline is nothing new. ML models can be trained on historical data to predict future prices based on various factors [2]. To achieve this, developers must choose the right ML algorithms and data inputs to ensure accurate predictions. ML learning algorithms are classified into 4 types:

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

However, these types of ML algorithms are classified into even more types such as:

- Linear Regression
- Logistic Regression
- Decision Trees
- Random Forest

- Support Vector Machines (SVM)
- Naive Bayes
- K-Nearest Neighbors (KNN)
- Neural Networks (including Deep Learning)
- Gradient Boosting Machines (GBM)
- Principal Component Analysis (PCA)
- Random forest algorithm

It's important to note that this is not an exhaustive list and there are many variations and combinations of these algorithms as well.

When the appropriate machine learning algorithm is utilized in conjunction with relevant data, it can assist companies and individuals in making better decisions and reducing costs [2]. In recent times for example, predicting prices of energy products has become more crucial due to its association with crude oil prices [3].

III. DATA SET

The data set I used for this Project was retrieved from Kaggle [5]. The data set contains regular gas, premium gas and diesel prices for all gas stations in Mexico and map locations in the form of a longitude and a latitude. It's important to note that while I retrieved the data from Kaggle, the data was extracted by Kaggle from datos.gob.mx which is provided by the CRE (Comisión Reguladora de Energía).

Figure 1.

10707	**PETREOS LAS GLORIAS S.A. DE C.V.**	**PL/9452/EXP/ES/2015**	**20.08398**	**-98.37579**	**21.95**	**23.49**	
3655	GASOLINERA CUDEA SA DE CV	PL/1882/EXP/ES/2015	20.5662	-103.3652	21.65	23.9	23.5
4597	AUTOSERVICIO LA CANDELARIA SA DE CV	PL/2693/EXP/ES/2015	29.13135	-111.0231	21.45	23.69	
9816	ESTACION DE SERVICIO CINCO DE MAYO SA DE CV	PL/8722/EXP/ES/2015	31.2587	-110.9517	17.99	21.19	22.29
8304	ESTACION DE SERVICIO FRAGOSO, S.A. DE C.V.	PL/7298/EXP/ES/2015	18.11682	-94.14303	20.55	22.25	22.59
11360	SUPER SERVICIO ABY SA DE CV	PL/9950/EXP/ES/2015	23.93204	-106.4312	21.9	22.99	23.19
10731	Elda María Gasque Casares	PL/9280/EXP/ES/2015	20.69962	-88.59959	22.32	23.52	23.92
11138	SERVICIO VILLADA, S.A. DE C.V.	PL/1820/EXP/ES/2015	19.37793	-99.00884	20.99	22.79	21.89
4900	SERVICIO OBISPADO S.A DE C.V.	PL/6083/EXP/ES/2015	28.69528	-100.5627	17.4	19.06	21.1
11738	SERVICIO AGUIRRE CASTELLANOS SA DE CV	PL/10342/EXP/ES/2015	20.29818	-103.1901	21.95	23.95	22.9
5287	PRODUCTOS FTLES P/ EXPORTACION DE EL SALTO S.A. DE	PL/5948/EXP/ES/2015	23.73085	-105.68	23.45		24.99
8225	SERVCIOS INTEGRALES SAN MIGUEL, S. A. DE C. V.	PL/6672/EXP/ES/2015	18.26823	-97.16126	21.95	23.59	23.29
9179	E.S.G.E.S. S.A DE C.V	PL/7825/EXP/ES/2015	19.85007	-90.52937	22.99	23.99	
9530	COMBUSTIBLES Y LUBRICANTES RIUZ SA DE CV	PL/8313/EXP/ES/2015	24.20953	-98.49334	21.65	23.3	23.3
14227	Combustibles Olmo, S.A. de C.V:	PL/12680/EXP/ES/2015	20.65189	-103.369	21.99	24.35	23.55
13059	SERVICIO TEPALCINGO S,A, DE C.V.	PL/10929/EXP/ES/2015	18.61637	-98.83538	20.89	23.19	22.99
7208	JOSE RAUL CERVANTES LOPEZ	PL/4369/EXP/ES/2015	19.28881	-97.69137	20.49	22.29	21.99
4351	SERVICIO RIO ELOTA, S.A. DE C.V.	PL/2682/EXP/ES/2015	23.91911	-106.8886	23.5	24.7	25.9
2178	PETROMAX, S.A. DE C.V.	PL/570/EXP/ES/2015	25.76927	-100.273	22.99	25.39	
24229	MEGA GASOLINERAS S.A DE C.V	PL/21071/EXP/ES/2018	20.9238	-100.7779	21.99	24.99	22.99

Figure 1 shows an excerpt of the data in which each column shows a different value. The columns are defined from left to right and their data is defined as follows:

1. Place_id (Id of the the gas station)
2. Name (Name of the gas station)
3. Cree_id (Petroleum products disposal permit)
4. Latitude (Latitude value)
5. Longitude (Longitude value)
6. Regular (Price of regular gasoline)
7. Premium (Price of premium gasoline)
8. Diesel (Price of diesel)

In total the data has 12,958 rows of information, I cleansed the data by doing the following:

- **Shuffle the rows randomly**: The rows are shuffled randomly to avoid any order bias that may affect the analysis.
- **Split the data into training and testing sets**: The data is split into a training set (80% of the data) and a testing set (20% of the data).
- **Extract the relevant features**: The latitude, longitude, year, and gas price values are extracted from each row. If any of these values are missing, the row is skipped.
- **Normalize the latitude values**: The latitude values are normalized using min-max normalization. The minimum and maximum latitude values are computed from the training set, and the same transformation is applied to both the training and testing sets.
- **Store the original and normalized samples**: The original samples are stored in a list called original_samples, and the normalized samples are stored in a list called samples.

Since I decided to predict the price of gasoline using the location of the gas station and year which is why my **features are latitude, longitude, and year** while **the target is the price of regular gasoline**. Choosing longitude, latitude, and year as features can be a good idea in some machine learning applications because they can provide important information about the geographic and temporal context of the data, which can help improve the accuracy of the model.

IV. MODEL PROPOSAL

This project was developed on Python, it's a linear regression using Gradient Descent algorithm to predict gas prices based on a dataset. As mentioned before the model is trained with 80% of the data and tested with 20% of the data. I did this to evaluate the performance of the model on unseen data, since I'm trying to create a model that can generalize well to new unseen data and in this way avoid overfitting.

The learning rate I used for the model is of **0.25**. I consulted online resources to help me arrive at a good learning rate. For example, I found this great graph:

Figure 2.

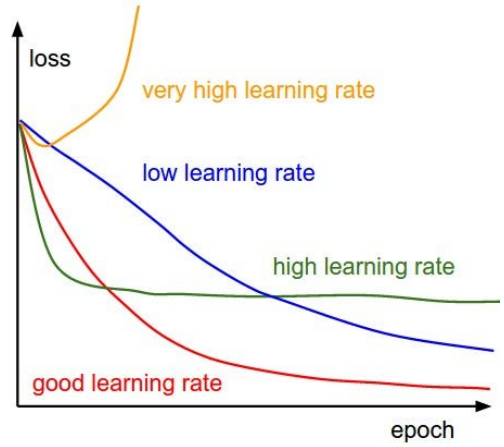
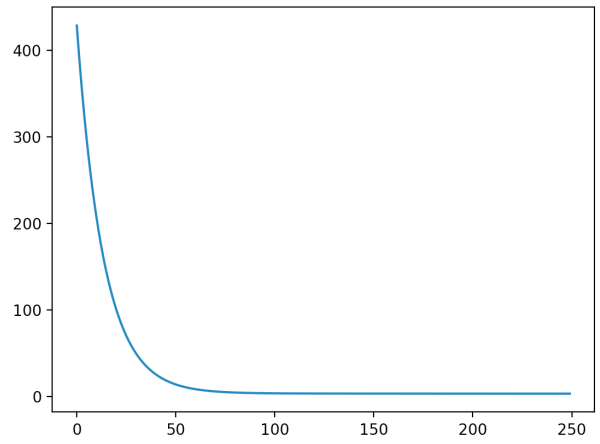


Figure 3.



Even after doing online researching, I arrived at the value of 0.25 by doing **trial and error**. After running the project, I arrived at the plot showed on figure 3. We can clearly see that as the **epochs increase** the **loss decreases** which means the model is **converging**.

To check how well the model was fitting the data I calculated two coefficients of determination, one for the training set and one for the testing set, the results were the following:

R squared test set: 0.8704063949433614
R squared train set: 0.8620442124691561
MSE test error: 3.006805858181015
MSE train error: 3.3890195691447396

Here we can see that the model is doing quite well when using unseen data (test set) since it's not only 87% of the variation in the dependent variable can be explained by the independent variable but it is also slightly higher than the train set which means the model is **not overfitting**. Also, since it's performing well on both the training set and the test set we can conclude that the model is **not underfitting**. I also calculated **train and test errors** by using the **Mean Squared Error**. The mean squared error for the test set is lower than that of the train set which makes sense when comparing them to the R squared for both sets. This further reinforces my conclusions on how well the model fits the data.

The final parameters obtained after running the program were the following:
Final parameters: [15.080756211662584, 5.68760064759406, 7.384211984232345, 0.7873907085865723]

In order from left to right these parameters are the following: bias, latitude, longitude, and year. The coefficients represent the degree to which each input variable contributes to the output prediction. For example, a larger coefficient for latitude would mean that the latitude has a greater impact on the prediction than the other values.

V. CONCLUSIONS

The implementation of the Gradient Descent algorithm for linear regression to predict gas prices in Mexico shows promising results in this project. However, I think that if a model were to be used to predict gas prices in a professional environment the features would have to expand. For example, the model could take into consideration crude oil prices, environmental regulations, political instability, consumer behavior and so many other factors than can affect gas prices. Having said that, this project was a challenge that truly helped me understand all the concepts we've been reviewing during the semester a lot better, and I truly think I learned a lot by doing the code manually and not using a framework since I had to think about what was going on with the model in a much deeper way than before.

VI. REFERENCES

https://www.bbc.com/mundo/noticias-america-latina-38514442 [1]

https://www.emergya.com/es/ideas/prediccion-de-precios-con-machine-learning [2]

https://www.imf.org/en/Blogs/Articles/2014/12/22/seven-questions-about-the-recent-oil-price-slump [3]

https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article [4]

https://www.kaggle.com/datasets/juanagsolano/gas-stations-prices-for-mexico?resource=download [5]

https://towardsdatascience.com/https-medium-com-dashingaditya-rakhecha-understanding-learning-rate-dd5da26bb6de [6]

https://ciep.mx/precio-de-la-gasolina-determinantes-historicos/