CIS 3715: Principles of Data Science

Final Project: Final Report

Student Name: Krishnarupa Sewsunker

Title: Linear Regression for Greenhouse Gas Concentration Levels and Global Warming Prediction

Introduction

In our present world, we observe more and more shocking instances of environmental tragedies occurring all over the globe. These tragedies are signs pointing to the destructive and unsustainable way in which we as a society live at present, prompting us toward finding solutions before it is too late. Environmental tragedies are the alarms and sirens of the planet urging us and our leaders toward change, and they present themselves in variegated ways – from wildfires, glaciers and ice melting, rise in sea levels, to species extinction – and as they occur with increasing frequency and severity, we continue to become increasingly aware of how closely and directly climate change is impacting our lives and threatening our future.

The motivation of my project was to visualize and make perceivable the extent to which an increasing concentration level of greenhouse gases in the atmosphere directly contributes to the heating if our planet – the phenomenon known as global warming or climate change. Just as the repercussions and symptoms of climate change manifest in diverse ways, the causes contributing to climate change are also numerous, however some causes are unquestionably more serious and detrimental. Out of all these causes, that which is the most severe and liable cause for the rapid rate at which global warming is proceeding is the increasing concentration levels of greenhouse gases in the atmosphere, which is why showing the correlation between an increasing concentration level and the increasing surface temperature of the earth would be an impactful way to show the severity of this problem. The main greenhouse gases are carbon dioxide, methane and nitrous oxide. These three greenhouse gases have the highest concentration in our atmosphere, with carbon dioxide being the leader by far. The problem of our planet's high and rapidly increasing atmospheric concentration of greenhouse gases is anthropogenic, meaning it is generated by humans, and the largest three reasons for greenhouse gas emissions are the burning of fossil fuels for electricity, transportation, and industry.

In this project, I have explored and examined the data on the European Environmental Agency webpage, the Our World in Data website (a University of Oxford-based organization), and the United States Environmental Protection Agency website. My contribution in this project was to firstly collect raw data of concentration levels of the three greenhouse gases respectively and the global temperature increase over a large span of time, pre-process the data by filling missing values, compile the data into one dataset and then produce graphical visualizations that show the correlation between greenhouse gases and global warming, and the unprecedented rate at which this problem is worsening today.

Approach

The idea of the project was to use the supervised learning method of linear regression to map the direct correlation between atmospheric concentration levels of greenhouse gases and global surface temperature. Supervised learning was appropriate in this case because for all samples in the data sets, we know all features and ground truth values. If such a linear regression model is built, it would be able to properly predict how a given concentration level of the three greenhouse gases in a given year would result in a certain increase in surface temperature of our planet.

Foremost in the approach to pursuing this project idea was to research selected topic to be well-informed in the area of investigation, and to develop a deeper understanding of the subject matter in order to be best prepared for my data collection step. This was especially important given that I was not to use an existing data set but was to compile my own data set by collecting multiple sets of relevant data for the independent and dependent variables, and by combining and pre-processing the data.

After extensive research on the topic and reviewing related works and data on the abovementioned and additional websites, I decided I would construct and use a data set that contained five features, all of which were numerical features. Four of the features were to be the independent variables, namely, the year, the annual atmospheric concentration level of carbon dioxide, the annual atmospheric concentration level of methane, and annual atmospheric concentration level of nitrous oxide, and the fifth and final feature of the data set was to be the dependent variable of average global surface temperature.

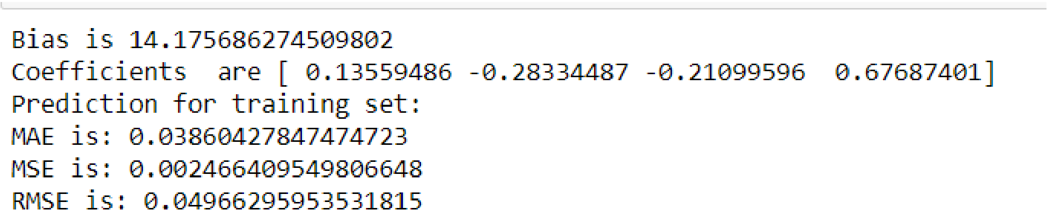
During the course of developing and refining my project, I considered the complexity level of my project and contemplated if I should add more features or variables so that my data set has higher dimensionality and so that I could possibly even consider applying the PCA method in conjunction with the linear regression model, however I decided not to do that in order to preserve the idea of showing the correlation exclusively between greenhouse gases and global temperature increase since this alone is the greatest contributing factor to global warming.

The pre-processing steps required in order to compile my data set was intensive and involved. This was a valuable process to undergo in terms of learning various ways to combine data using the Pandas library. Firstly, the data for each of the greenhouse gases respectively had different starting years and some of the data (particularly the data for carbon dioxide concentration levels) went much further back in time than the rest of the data. Since the data for methane and nitrous oxide began in the starting year 1750, I decided to use this year as the beginning of my data set.

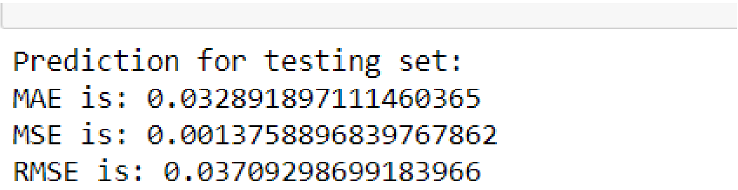
Next, I had to decide what values to fill any missing values with since filling with mean values wasn't appropriate in this case. This is because it was common that the missing values in atmospheric concentration levels for a certain greenhouse gas were most frequent in very early years such as during the eighteenth century and that there were not many (or any) missing values in the table for the more recent years. This meant that, since it is implicit in the problem of global warming that these greenhouse gas concentration levels are substantially higher in present times than historic times, filling missing values with missing values would wrongly skew the data, making it seem that concentration levels were high in historic times when in actual fact, they would have been low as global warming was not as dire a problem in those times. After the missing values were taken care of, I had to combine my individual data sets into one data set. This was the final data set I used to build my linear regression model.

Results

Important metrics for understanding how accurately a linear regression mapping function works for a given data set are the mean absolute error, the mean squared error and the root mean squared error (MAE, MSE and RMSE). If these metrics are very small in magnitude, it shows good predictability of the learned linear mapping function for a given data set, and the smaller these metric values are, the better the accuracy and predictability of the linear regression model. Since these metrics all had very low values, as shown below for the testing set and training set (85:15 split), it was clear that my linear regression model was able to accurately make predictions for this data and that it was an appropriate model for the context of this problem.



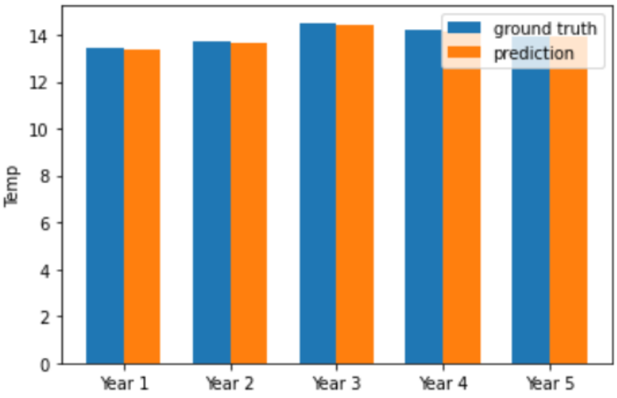
```
Bias is 14.175686274509802
Coefficients  are [ 0.13559486 -0.28334487 -0.21099596  0.67687401]
Prediction for training set:
MAE is: 0.03860427847474723
MSE is: 0.002466409549806648
RMSE is: 0.04966295953531815
```



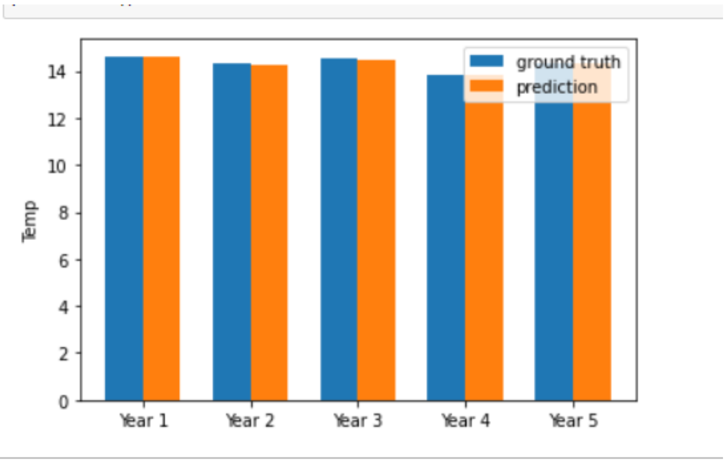
```
Prediction for testing set:
MAE is: 0.03289189711460365
MSE is: 0.0013758896839767862
RMSE is: 0.03709298699183966
```

The two bar graphs on below show the comparison between the ground truth values and the prediction values based on the linear regression model. These graphs depict the ground truth and predictions for two different 5-year samples from the testing set, and in both graphs, the ground truth and prediction values are very close to each other, which further indicates that the linear regression model was accurately able to make predictions in a consistent manner.

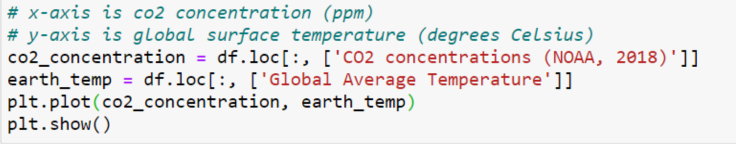
Sample 1:



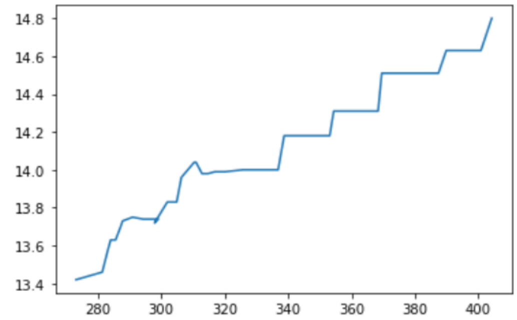
Sample 2:



Below is a line graph showing the correlation between the concentration levels of carbon dioxide in the atmosphere and the average surface temperature of the earth. Analyzing this particular correlation is important and valuable because, as we know from the data, carbon dioxide is the most highly concentrated greenhouse gas in the atmosphere by far. This line graph shows how as the concentration of carbon dioxide increases, so does the global surface temperature.



```
# x-axis is co2 concentration (ppm)
# y-axis is global surface temperature (degrees Celsius)
co2_concentration = df.loc[:, ['CO2 concentrations (NOAA, 2018)']]
earth_temp = df.loc[:, ['Global Average Temperature']]
plt.plot(co2_concentration, earth_temp)
plt.show()
```



Conclusion

In conclusion, my project utilized the supervised learning model of linear regression to make predictions on average global surface temperature based on annual values for the atmospheric concentration levels of the three primary greenhouse gases. Global warming is the phenomena of the earth's increasing temperature as a result of pollution and environmental destruction, and it is common knowledge that greenhouse gases are the greatest contributor to global warming, however this has now been illustrated and underscored by the results of this project. As such, this project illustrates how our means sustaining of energy, transportation and industry are in fact unsustainable for our planet and for all life on the planet. As we continue to head in the direction we are as a society, we will find increasing instances of global natural calamity as climate change continues to be perpetuated, so this project has shown that it is essential that we seek and pursue sound alternatives to the path we have been surging into since the eighteenth century.

Acknowledgements

1. European Environmental Agency [website]:

https://www.eea.europa.eu/data-and-maps/indicators/atmospheric-greenhouse-gas
concentrations-7/assessment

2. Our World in Data [website]:

https://ourworldindata.org/atmospheric-concentrations

3. United States Environmental Protection Agency [website]:

https://www.epa.gov/climate-indicators/climate-change-indicators-atmospheric-concentrations-greenhouse-gases#ref5

4. Current Results: weather and science facts [website]:

https://www.currentresults.com/Environment-Facts/changes-in-earth-temperature.php

5. National Oceanic and Atmospheric Administration: Climate.gov [website]:

https://www.climate.gov/maps-data/datasets/coverage/global/formats/csv