Python for Engineering Data Analysis – from Machine Learning to Visualization
Department of Electrical and Computer Engineering
Technical University of Munich

Submitted by:
Duygu Chousein Nasouf, 03745754, B.Sc. Management & Technology at TUM School of Management
Utku Atay, 03720457, M.Sc. Management & Technology at TUM School of Management
Technical University of Munich, Arcisstraße 21, 80333, Munich

March 16, 2023

Predictive Analysis on Global Warming using Ploty Dash Visualization

1 Abstract

In the last few years, several institutions are providing information about the rising trend of our planet's temperatures. Unfortunately, this trend has nothing positive in it. This results in natural disasters that endanger the lives of humans and animals. Within this report, we will be doing predictive analysis on global warming using machine learning. We attempt to predict the concentration of greenhouse gases in the atmosphere which directly causes a change in the surface temperature of the Earth. These predictions will help us to identify the emergency of the current situation and will hopefully help us to cope with it in the future.

2 Introduction

The main objective of this report is to predict the Earth's surface anomaly temperature and greenhouse gases such as carbon dioxide, methane, and nitrous oxide based on our dataset from "Our World in Data" [1]. To do so, our first step is to analyze global warming by examining the relationship between the greenhouse gases concentration and the temperature anomaly. We do this for each of the gases and after then show a world map starting from the year 1950 to 2014 with the value of the CO_2 emissions. These results should all be to each other. After analyzing the relationship until 2020, we do our prediction for the developments for the next years. After all, we visualize our results in a Graphical representation as a dashboard using plotly's dash.

3 Theory behind our four methodologies

For our purpose, we will compare four different methodologies: linear regression, exponential smoothing, seasonal naive, and prophet in a window of 12 months starting with January 2016. The reason for that is, that we want to find out the methodology with which we want to get a realistic and promising result. First, let us give a short description of the four alternatives.

We use *linear regression*, which is an analysis to predict the value of a variable based on the value of another variable. It, therefore, gives us a relationship between a dependent variable and an independent variable (variable of prediction) [2].

Another alternative is the *seasonal naïve method*, where we set each forecast to be equal to the last observed value from the same season, for example, the forecast for all future values for March is equal to the last observed March values. *Exponential smoothing* forecasts are produced by the weighted averages of past observations, with the weights decaying exponentially as the observation gets older. With this methodology, we can generate quickly and for a wide range of time series [3].

Our last possibility is *Prophet,* which is a forecasting procedure implemented in Python as an open-source library. It is used for forecasts for data with trends and seasonal default structures [4]

4 Experiment Setup

4.1 Understanding the Data

For our analysis, we used four datasets. The *CO_2 emissions dataset* [5], where we get our country-wise CO_2 emission values starting from the year 1751 until the year 2014. As we have nearly all values starting from 1950, we will take the model starting from 1950 into account.

The second dataset we used is the *Mauna Loa CO_2 emission dataset* [6] which contains the CO_2 values as well as the monthly breakdown. The usage is explained in 4.3.

The third and fourth datasets from 'Our World in Data' we used to get the temperature anomaly and the greenhouse gases with countries.

The *greenhouse gases dataset* [7] includes data on the greenhouse gases which are taken into consideration in this report with their corresponding countries as well as specific data on the emissions (per capita, annual). It follows the format of 1 row per location and year. The *temperature anomaly dataset* [8] contains the average temperature anomaly since 1850.

4.2 Cleaning Data and Analyze Correlations

Firstly, we want to check the correlation between temperature anomaly and the change in greenhouse gas emissions. Three gases contribute to greenhouse gas emissions mostly carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O).

Figure 1: *Contributions to greenhouse gas emissions [9]*

CO_2 has the most contribution with 74.4%, followed by methane with 17.3% and nitrous oxide with 6.2 %. To see the correlation in a better way, we plot CO_2-temperature in one plot and the other important greenhouse gases in another plot. The greenhouse gases dataset has breakdowns in the country and years. To get a total emission over years, we sum up the values by summing country values in each year. Since we have regular data after 1950 for temperature data, we filter greenhouse data also after that year. The temperature data has 3 types of entity: Global, North Hemisphere, and South Hemisphere. Since we want to have a general look at the correlation, we get only global data after 1950.

When figure 2 is analyzed, it can be clearly said that there is a strong correlation between CO_2 and temperature anomaly over years.

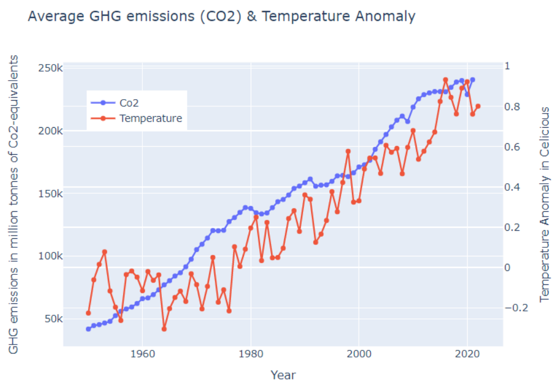

Figure 2: *CO_2 Emission and Temperature Anomaly after 1950*

In greenhouse gas data, there are only proper data for methane and nitrous oxide after 1990. Therefore, we filter greenhouse gas data and temperature data by year after 1990.

Figure 3 shows that there is also a strong correlation between the other greenhouse gases and temperature anomalies. It may be also said that all three values are increasing over years.

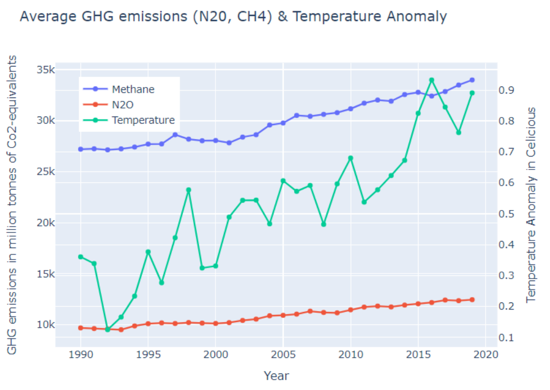

Figure 3: *CH_4 and N_2O Emission and Temperature Anomaly after 1990*

We also plot the country breakdown of CO_2 emission over years on a map by using the *CO_2 emissions dataset* [5] because country names in that were given more appropriately. The map with animation (figure 3) clearly shows that CO_2 emissions are increasing rapidly. Additionally, in the first years, the United States had the lead for emissions, however, in recent years China has taken the lead.

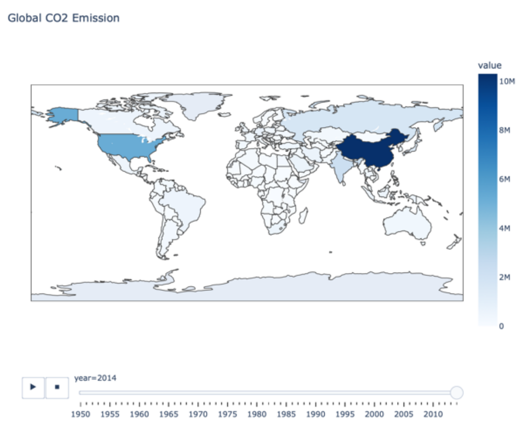

Figure 4: *CO_2 Emission over Years on the world map*

4.3 Creating and Fitting the prediction model

We continue our model with only CO_2 since it has already been noted that CO_2 covers the almost 75% of all greenhouse gas emissions. For the forecasting model, we have used the Mauna Loa CO_2 emission dataset as it contains values with monthly breakdowns.

As we want to create a forecasting model with a time series, we need monthly data. We used the 'Darts' package to create the time series and used 'Plotly' to make dashboards. When the CO_2 emission data is visualized, it can be realized that it has a seasonal pattern. The CO_2 emission reached the local maximum in May and minimum in October. Additionally, there is an upward trend over the years.

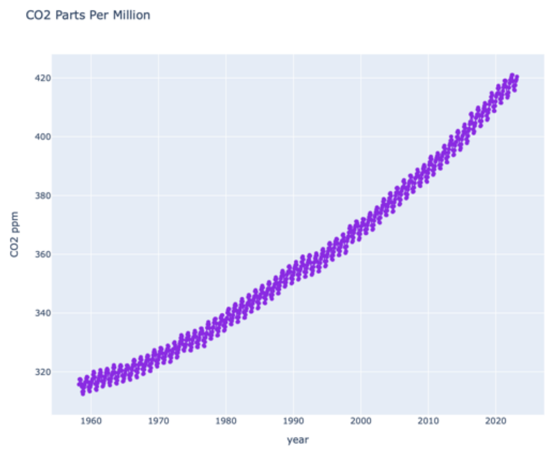

Figure 5: *Monthly CO_2 Emission over the years*

Therefore, we have selected four different models which can be suitable for seasonality and upward trend, which are naïve seasonal, exponential smoothing, linear regression and prophet were already explained in Chapter 3. They make forecasts using historical data and take January 2016 as starting time and a forecast horizon of 12 months for forecasting. Therefore, the models have forecast values from December 2016 to February 2023.

4.4 Data Visualization

For our data visualization, we would like to display the graphs/results that we created in a simple dashboard so that the end user can get a meaningful overview of the whole model. For this implementation, we use the library 'plotly'.

The first model we want to have a look at is Naïve Seasonal. To test the models, we use the historical_forecasts() function, which trains on an expanding window. As we can see in Figure 6 the forecasts are below the actual values, which seem not to be the best fit for our dataset.

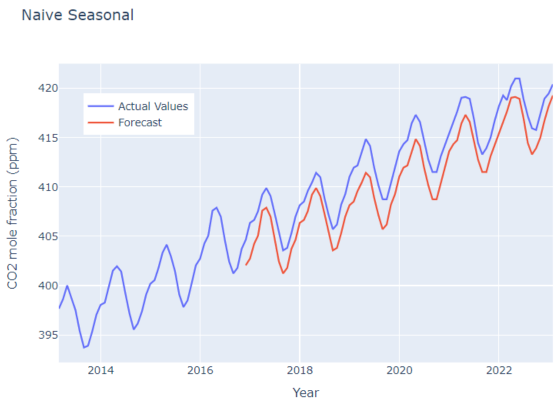

Figure 6: *Forecast of Naïve Seasonal*

The second methodology we used is exponential smoothing, and it predicts very well compared to the Naïve Seasonal, as the actual values are fitting with the forecasted ones, which can be seen in Figure 7.

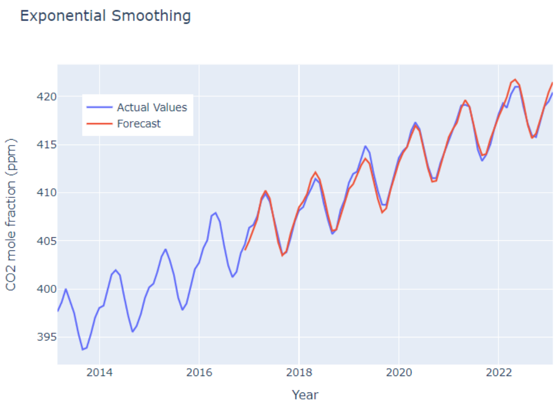

Figure 7: *Forecast of Exponential Smoothing*

The third possibility we want to take into consideration is linear regression, which is working well with CO_2 emission data. Our obtained R^2 value is 0.98.

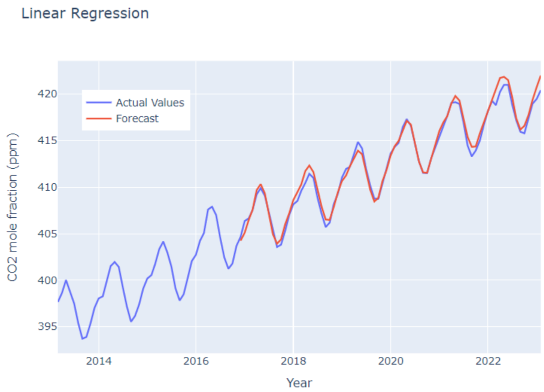

Figure 8: *Forecast of Linear Regression*

Lastly, our prophet model performs better than Naïve seasonal but worse than the others.

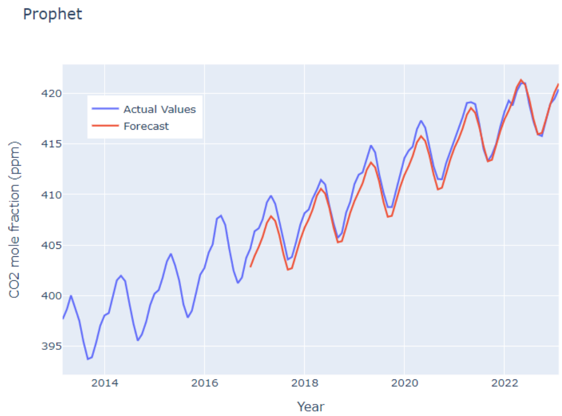

Figure 9: *Forecast of Prophet*

5 Results

We have used 6 different metrics to compare our forecast results of models, which are Root Mean Squared Error (RMSE), Root Mean Squared Log Error (RMSLE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE) and R-squared (R^2). Table 1 shows each model's performance.

	MAE	RMSE	MAPE	SMAPE	RMSLE	R2
Naive Seasonal	2.33	2.39	0.56	0.57	0.01	0.74
Exponential Smoothing	0.44	0.55	0.11	0.11	0.00	0.99
Linear Regression	0.48	0.62	0.12	0.12	0.00	0.98
Prophet	0.96	1.11	0.23	0.23	0.00	0.94

Table 1: *Metrics for models*

We want to have lower error metrics and higher R^2 values. Therefore, we select the exponential smoothing model to predict the next 2 years, as it has the highest R^2.

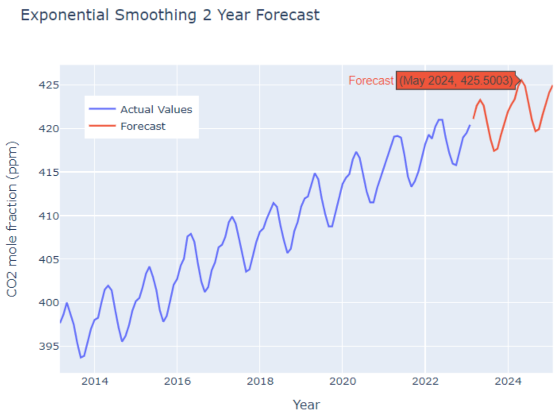

Figure 10: *Forecast of Exponential Smoothing for the next 2 years*

The forecast shows that the CO_2 concentration will be still increasing rapidly and will exceed even 425 ppm at the beginning of 2024 and will continue increasing without any action done.

6 Conclusion

We can see a significant rise in the concentration of greenhouse gases, and if we do not take measures in our daily lives, this will also tend to further increases. Sadly, we discovered that the concentration is expected to increase in the following years too. That should be an urgent signal that we must reduce the emissions, as we have only an increasing line.

There is a direct effect of the increasing temperature of the Earth. With the increasing emissions of greenhouse gases, the problem of global warming is going to increase and affect everyone's lives. With continuous emissions and an increase in the emissions of greenhouse gases, the problem of global warming will increase.

References

[1] Our World in Data (w.Y): CO2 and Greenhouse Gas Emissions, https://ourworldindata.org/CO2-and-greenhouse-gas-emissions (status:01.03.2023)

[2] IBM (2023): Linear Regression. https://www.ibm.com/topics/linear-regression (status: 02.03.2023)

[3] Hyndman, R. and Athanasopoulos, G (2016): Forecasting: Principles and Practice, 2nd edition, https://otexts.com/fpp2/ (status: 02.03.2023)

[4] Brownlee, J. (2020): Time Series Forecasting with Prophet in Python, https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/ (status: 02.03.2023)

[5] Gianatmaja: CO2-Emissions.csv, URL: https://github.com/Gianatmaja/Global-Warming-Dashboard/blob/master/CO2_Emissions.csv (status: 05.03.2023)

[6] Global Monitoring Laboratory (2023): Mauna Loa CO2 monthly mean data, URL: https://gml.noaa.gov/ccgg/trends/data.html (status: 05.03.2023)

[7] Our World in Data (w.Y.): Data on CO2 and greenhouse gas emissions by Our World in Data, URL: https://github.com/owid/CO2-data (status: 03.03.2023)

[8] Our World in Data (w.Y.): Average temperature anomaly Global, URL: https://ourworldindata.org/grapher/temperature-anomaly (status: 03.03.2023)

[9] Ritchie, H. (2020, May 11). CO_2 and Greenhouse Gas Emissions. Our World in Data. https://ourworldindata.org/greenhouse-gas-emissions (status 04.03.2023)