Everett Johnson

March 2, 2022

DSC 680

Machine Learning Model – Climate Change Analysis

Business Problem

Climate change has been a popular subject in the last ten years. Scientists, politicians, and everyday individuals have their opinions on what climate change is and how it is affecting our planet. Although there are many different opinions on the extend of climate change and how it can affect us in the future, we have plenty of data today to study the phenomenon and come up with our own conclusions.

Background/History

Data has been collected for many years regarding the average temperatures of the Earth's surface. Data scientists can use this data to form models that could give a good indication of how fast and even if the Earth's surface temperature is increasing.

Data Explanation

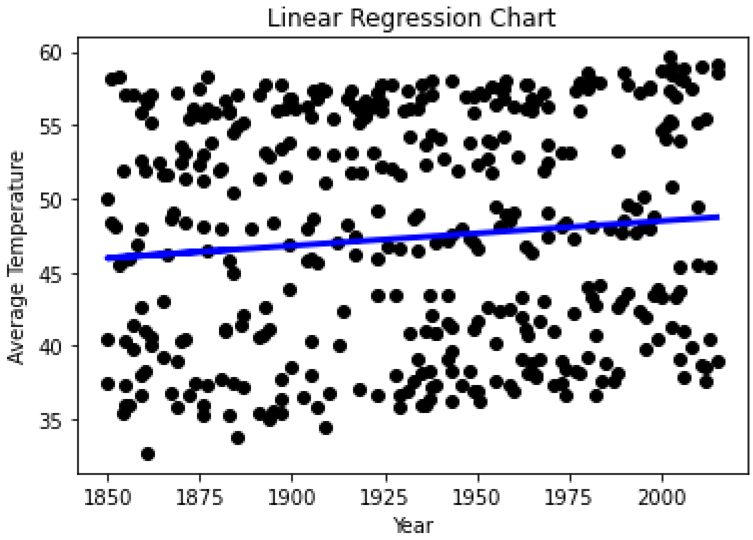
The data used within this model is downloaded from Kaggle.com. According to the website, the data is downloaded and collected from Berkley Earth, "We have repackaged the data from a newer compilation put together by the Berkeley Earth, which is affiliated with Lawrence Berkeley National Laboratory. The Berkeley Earth Surface Temperature Study combines 1.6 billion temperature reports from 16 pre-existing archives." (Kaggle.com). There are multiple files available but the main file that will be used in this model is the "Land Average Temperature in Celsius" file. The timeframe for this data is the 1750's to the present.

Methods

The methods used in this model will be a linear regression model using the Sci-Kit Learn library to study and predict how the average temperature of the Earth's surface may progress. Since the temperatures could be considered continuous in nature, then the linear regression model is a good starting point for experimentation. During the cleaning and pre-processing stage, the data is checked for any missing values that could otherwise skew the data. For the model, the data was split into training data and test data – a standard 80/20 split was used.

Analysis/Conclusion

The model indicates that the overall trend of the average temperature of the Earth is in an upward trend as seen below. The overall model accuracy is showing to be about 85%, but the parameters, such as MAE and MSE are showing to be low and higher at 6.87 and 58.7 respectively. Due to this further analysis should be used to see if there are other models that may produce similar results.



Assumptions

The main assumption with the data is that it is truthful and accurate. Also, that the data shows an average of the entire Earth's surface.

Limitations

This information can be useful to some extent, but a more granular approach using specific regions of the Earth, such as Antarctica, could be analyzed to determine more specific information and trend patterns.

Challenges

There are a few main challenges to this type of data and model. The first would be to ensure the data accuracy and see how the temperatures were collected. This may not be an issue in current times with the available technology, but ensuring data accuracy from certain timeframes, such as the late 1800's, would be a challenge. Most of the data was missing from the 1750's, which was removed from the model but ensuring the data quality of the data that is provided could be a challenge.

Future Uses

Future uses for this type of model could include a use for climate researchers to see how the Earth's surface temperature has changed of the last few centuries. Further analysis could be completed on different features to see what correlations there are with the temperature increases. It can also be used as a starting point to predict how future temperatures may look over time.

Recommendations

Recommendations would include using other types of data in the model, such as carbon dioxide emissions and deforestation data to see if these items could correlate with the average surface temperature over time.

Implementation Plan

Implementation plan could include adding source files and code to a website for future climate researchers to study.

Ethical Assessment

Ethics should play a part in any discussion regarding climate research. There should be a disclaimer when providing the results of a study, such as this that this is only one small part of the bigger picture. Climate change is a complicated process with many data points to consider. Providing the results of any one study without context could influence people’s interpretation and be misleading, especially when the topic is discussed in a political context.

References:

Chugh, A. (2020, December 8). *Mae, MSE, RMSE, coefficient of determination, adjusted R squared‑which metric is better?* Medium. Retrieved March 2, 2022, from https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e#:~:text=The%20Mean%20absolute%20error%20represents,the%20residuals%20in%20the%20dataset.

Linear Models. scikit. (n.d.). Retrieved March 2, 2022, from https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares

Earth, B. (2017, May 1). *Climate change: Earth surface temperature data*. Kaggle. Retrieved March 2, 2022, from https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data