Capstone Project Report: Sunspot Forecasting

Siddhanth Kumar

Abstract

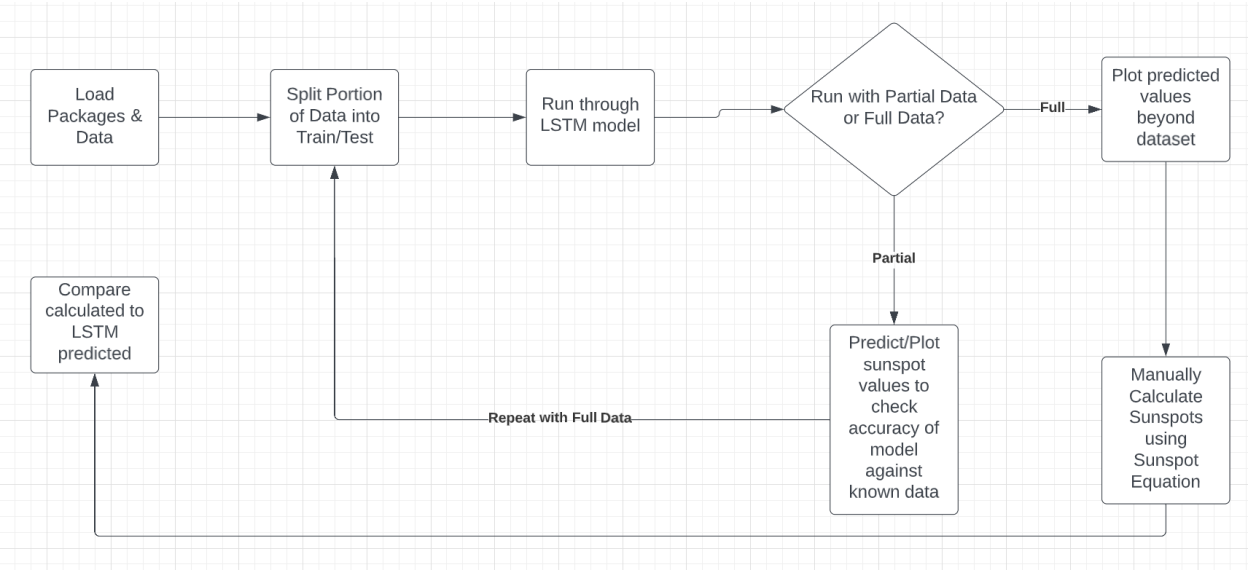
This project aims to create a machine learning model to accurately forecast the monthly mean sunspot number. This was accomplished in 3 high-level parts. First, datasets were gathered for the training and testing of an LSTM Recurrent Neural Network for sunspot number predictions. Next, the sunspot number equation was used to validate the accuracy of the created model. Finally, the historical dataset, Recurrent neural network data, along with the sunspot number equation calculated data were plotted on one graph.

Introduction

This machine learning model was inspired by the significance of the sunspot number, a numerical representation of the Sun's "spottedness" due to individual sunspots and sunspot groups. Sunspots are dark spots on the photosphere of the sun due to a given area being cooler than the surrounding regions. Besides solely quantifying the number of groups and spots on the sun, the greater importance of sunspots comes from their ability to evaluate the solar cycle. The frequency of visible sunspots indicates solar activity throughout the 11-year solar cycle. Additionally, sunspots are where solar flares appear on the photosphere and where magnetic fields are the strongest. Understanding the importance of sunspots and their solar implications led to the creation of a neural network to accurately forecast future sunspot numbers. Moreover, to validate the accuracy of the model, it was compared against historical data from the Solar Influences Data Analysis Center of the Royal Observatory of Belgium along with data calculated from the "sunspot number equation", an equation to calculate the sunspot number based on the number of individual sunspots and sunspot groups. This all culminated in a detailed graph with all the data.

Calculations

The implementation required several steps.



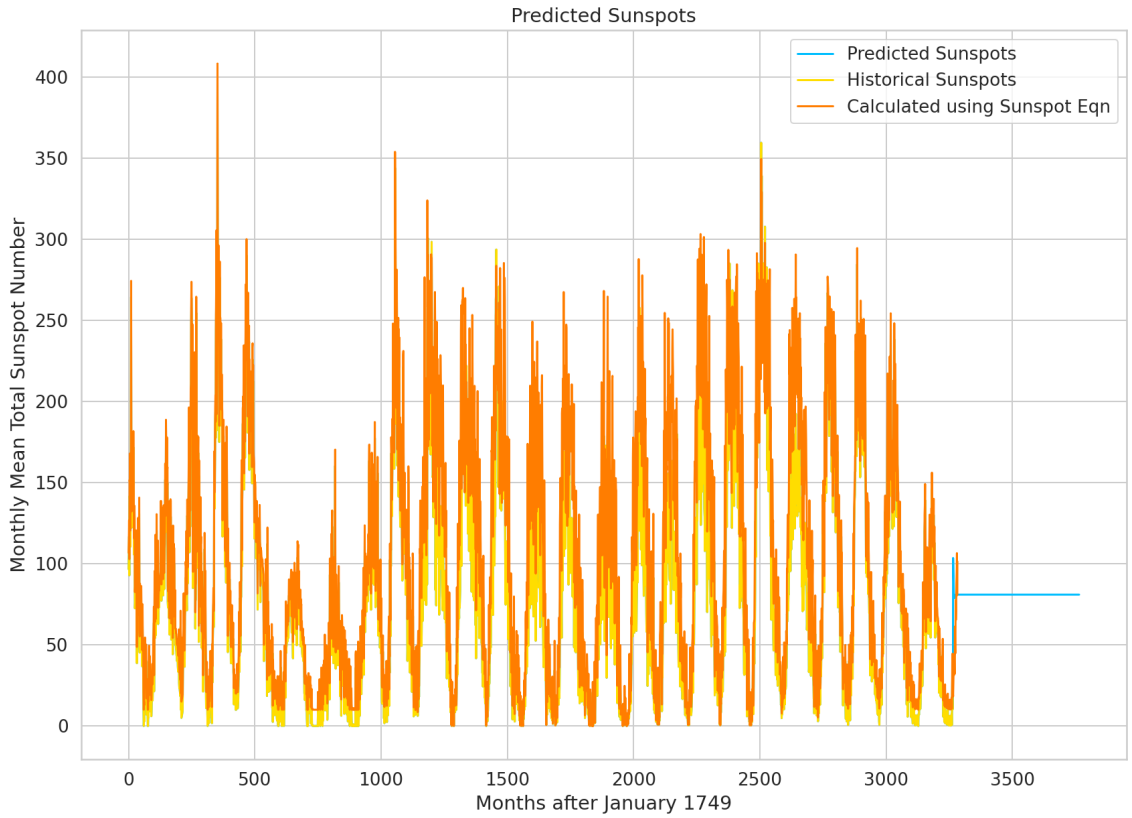
First, all the required packages (NumPy, pandas, tqdm, seaborn, matplotlib, sklearn, torch) were imported along with downloading the data that was used for training and testing the neural network. Secondly, after formatting the data to be ready for testing, the data passed through a training function to train the model. The model was trained on 100 epochs; one epoch is a complete pass through the training dataset. We chose 100 epochs to minimize underfitting (the model is not trained well enough to accurately predict values, training loss > test loss) and overfitting (the model becomes too accustomed to training data, test loss > training loss). Every epoch also returned a "train loss", signifying how every epoch reduced the loss of the model, bringing us closer to the desired accuracy for the model. Training loss signifies how well the model is accustomed to the data, while test or validation loss signifies how well the model can adapt to new data. For a successful model, both the training and test loss should be similar and close to zero. Afterward, we trained the model on all training data because we didn't need to split the dataset into training and testing portions. We predicted 100 data points from the end of the dataset. Afterward, the sunspot number over time is described by applying the "Boulder Sunspot Equation" derived by Swiss astronomer Rudolf Wolf below (Eq. 1):

$$R = k\,(10g + s) \quad (1)$$

R is the sunspot number, **k** is a correction factor (meant to normalize data from the 1700s and 1800s when telescope technology was not as advanced), **g** is the number of sunspot groups at a given time, and **s** is the number of individual sunspots at a given time. Using two datasets from the Solar Influences Data analysis center from the Royal Observatory of Belgium again, we acquired a dataset for the mean number of individual sunspots per month and the mean number of groups per month. This allowed us to calculate the sunspot number for a given month, R.

Graphing

Matplotlib.pyplot was used to graph all the plots in this project. For the final plot, three distinct lines were plotted: historical sunspot number data, sunspot number equation data, and machine learning predicted data. To create an array for the sunspot number equation data, a function was created called “calculate”. This performed the equation stated above.



Conclusion

As seen in Graph 3, the machine learning prediction values (blue line) match up closely with the calculated data initially. However, looking forward, the flat blue line is caused by our model's limitation being a small dataset. It could also be the nature of RNN models that they cannot predict accurately after a certain number of values. Overall, the model was able to predict results effectively. A larger training dataset would likely have enabled our model to predict more months in the future accurately, so for future projects, we hope to implement larger datasets to reduce loss and increase validation accuracy. One success of this project is machine learning and its initial accuracy. However, a drawback of this project is the lack of sizable datasets in many portions of the project including the actual training dataset along with the sunspot group count dataset used for the sunspot equation. Overall, this project was a satisfying way to culminate the knowledge learned from the class.