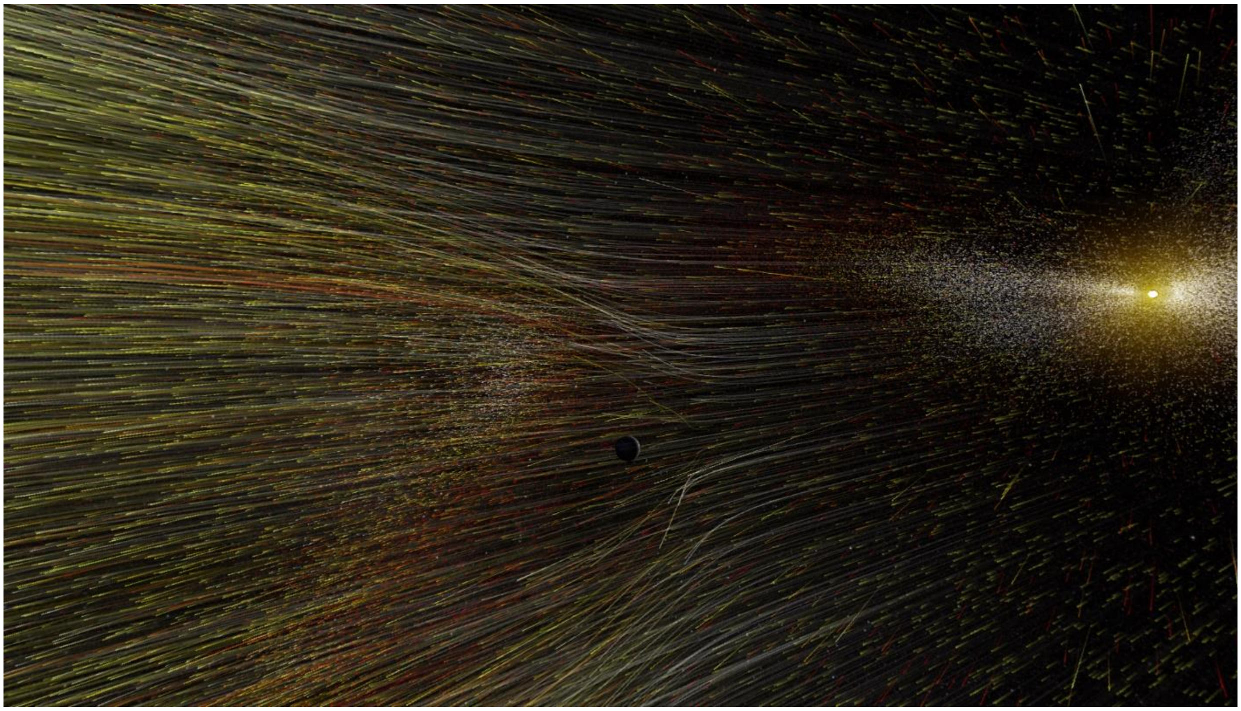

Space Weather: Image Courtesy of NASA

Analysis of Solar Wind Dataset to Predict Disturbance-Storm-Time Index

Semester Project Final Paper

Kari Abromitis

October 14, 2022

Abstract

This semester project analyzes a dataset of solar wind information to discover the basics of solar wind phenomenon and to predict the Disturbance-storm-time (Dst) index over three nonconsecutive periods. Dst is the main indicator of the strength or weakness of Earth's magnetosphere and tends to increase during extreme solar events. Predicting these events is crucial as it can help humans protect vulnerable infrastructure. This project uses the Long Short-Term Memory (LSTM) neural network model to predict Dst values with moderate success. This project also explores the ethical considerations of this analysis and potential future work.

Overview

The data for this project is related to solar wind, which is a phenomenon where charged particles from the sun are released and flow past the Earth's magnetosphere and can potentially weaken our planet's protective layer. The magnitude of these events is measured by the Disturbance-storm-time index or Dst, which gives information about the strength of the earth's magnetosphere at any one time. A negative value means that the magnetosphere is weakened. Our planet's baseline Dst is –20 nanoteslas (nT) and extreme events are generally characterized as –250 nT [1].

The most extreme solar wind event in recorded history was the Carrington Event of 1859 which where Dst was estimated to be at -1600 ± 10 nT [2]. This event was the result of coronal mass ejection (CME) from the Sun and a large solar flare was visible from Earth, as well as aurora borealis as far south as the Caribbean. A geomagnetic induced current also caused widespread telegraph systems failure. An event of this magnitude today would cause major communications disruptions and power outages [3]. Therefore, predicting and protecting infrastructure against these severe events is of crucial importance.

The dataset used for this project is the solar wind dataset obtained from Kaggle [4]. The data is composed of measurements collected from two satellites: NASA's Advanced Composition Explorer (ACE) and NOAA's Deep Space Climate Observatory (DSCOVR). The data is combined from these two satellites to mitigate gaps in the dataset. The data includes 1-minute data for coordinates (Bx, By, Bz) in both GSE and GSM coordinates, interplanetary-magnetic-field component magnitude, and solar wind density, speed, and temperature. Separate datasets provide observed sunspots and Dst.

I am motivated to use this dataset because I am interested in solar wind and data produced by satellite instrumentation generally. At my job, I develop satellite hardware that is ultimately used to collect scientific data. So, I am motivated to work with these types of datasets to help understand how my work is used by the scientific community. The question I want to answer with this project is if the nondirectional features of this dataset predict extreme solar wind events. I plan to address this question by splitting the data into training and test data, creating a model that predicts the training set of data, and test the model on the test set of data.

Related Work

L.Q. Zhang et al utilized the clock angle θ_{CA} (arctan(By/Bz)(-90˚ to 90˚) of the interplanetary magnetic field (IMF) of solar wind, using ACE satellite observations [5]. They found that the IMF decreases with the increase of the temporal scale (i.e., duration) of the solar wind. They found this by using different groupings of wind speeds, and repeatedly found that higher-speed solar wind tends to have larger IMF θ_{CA}. Typically, the IMF θ_{CA} in the solar wind is widely distributed from 90˚ to 90˚, with two peaks at 65˚ and 65˚. They also suggest a monthly variation of this metric, suggesting seasonal variation of IMF θ_{CA}. This research was very enlightening for me and helped me understand the θ_{CA} and Bx, By, Bz columns provided in the dataset much better. My research will be unique from this example as I will focus on solar speed and other features independent of $\theta_{CA.}$ However, this research gave me context for understanding wind speed's relationship with θ_{CA} which may still be useful.

C. Larrodera et al used extreme value theory (EVT), which is used to analyze the likelihood of rare and severe events, to examine ACE satellite data [6]. They looked at solar wind characteristics of magnetic field magnitude and solar proton speed, temperature, and density to characterize the response of the magnetosphere in extreme events. They looked specifically at one in 40 and one in 80 year events. The extreme value distribution was applied to the data to estimate the magnitude of these rare events, and then compared to actual events over a 20 year period beginning in 1997. My work will be unique from this analysis because I will not be using a EVT approach and will not be grouping data in the time periods that Larrodera et. al. used. I will look at many of the same characteristics, however, so this work will be useful to compare my results to.

Upendran et al use deep learning for prediction of solar wind properties [7]. They used preprocessed images of the sun to build a model and test it against the OMNIWEB dataset. This data is taken from multiple satellites in geocentric or L1 orbits. Their model ultimately outperformed benchmark models, obtaining a best fit correlation of 0.55 +- 0.03. They also found that the model accurately predicted higher activation at the coronal holes for fast wind prediction and at the active regions for slow wind prediction. As they did not put this information into their model, it suggests that the deep learning model acquired this knowledge without the physics-knowledge being built in. My work will be unique as I will not be using deep learning on images to create a model and also will be using an entirely different dataset, thus will be unique from this study. It is interesting to see this very different (i.e., image based) approach to a similar problem that I am trying to solve.

Data Acquisition

The data for this project is provided by NOAA, With support from NASA. There are no limits on using or sharing the data as it is publicly available. The data has been preprocessed to a certain extent, as it is aggregated into three non-contiguous periods commonly across datasets.

The features in this dataset are listed below [4].

- bx_gse/gsm - Interplanetary-magnetic-field (IMF) X-component in geocentric solar ecliptic (GSE) and geocentric solar magnetospheric (GSM) coordinates (nT)
- by_gse/gsm - Interplanetary-magnetic-field Y-component in GSE/GSM coordinate (nT)
- bz_gse/gsm - Interplanetary-magnetic-field Z-component in GSE/GSM coordinate (nT)
- theta_gse/gsm - Interplanetary-magnetic-field latitude in GSE/GSM coordinates (degrees)
- phi_gse/gsm - Interplanetary-magnetic-field longitude in GSE/GSM coordinates (degrees)
- bt - Interplanetary-magnetic-field component magnitude (nT)
- density - Solar wind proton density (N/cm^3)
- speed - Solar wind bulk speed (km/s)
- temperature - Solar wind ion temperature (Kelvin)
- sunspots – Number of sunspots observed on a given day, smoothed
- Dst - measure of the severity of a magnetic storm

The target feature in this list is Dst which is the measure of the severity of a magnetic storm and is what scientists are typically trying to predict using the other factors such as density, spend, temperature, and number of sunspots. I will be attempting to build a model that predicts the occurrence and severity of Dst.

Preprocessing

This data is available in three csv files. The first has the majority of the data, with all of the features listed in the previous section except for sunspots and Dst. The data is minutely, and there are over 8 million data points. The second csv file contains only the sunspot data, and it contains time stamps for each point of data, which is typically 10-30 days apart. Thus, this dataset is much smaller. The dataset for Dst is similarly smaller in the third csv, but still more frequent than the sunspot data points with over a hundred thousand data points.

First, looking at summary statistics, I can begin to get a feel for the magnitudes for each of the features, as shown in Table 1. The coordinate features for this analysis are dropped, so the focus is on the variables for bt, density, speed, and temperature, and number of sunspots. The zeros for some of the minimums most likely indicates that there are null values that will need to be removed. Otherwise, there are not many indications of bad or missing data. Dst shows a fairly even spread from -387 to 67.

Table 1: Descriptive Statistics

Metric	bt	density	speed	temperature	smoothed_ssn	Dst
count	8.07E+06	7.71E+06	7.70E+06	7.58E+06	192	139872
mean	5.61E+00	4.42E+00	4.31E+02	1.15E+05	58.09	-11.06
std	3.11E+00	4.33E+00	1.01E+02	1.20E+05	52.518	19.07
min	3.00E-02	0.00E+00	0.00E+00	0.00E+00	2.2	-387
25%	3.64E+00	1.79E+00	3.57E+02	3.98E+04	14.15	-18
50%	4.95E+00	3.34E+00	4.10E+02	7.74E+04	39	-8
75%	6.72E+00	5.71E+00	4.86E+02	1.51E+05	95.13	0
max	8.05E+01	2.00E+02	1.20E+03	6.22E+06	175.2	67

Another useful high level view is looking at the periods seperately for each variable. This will indicate the differences between these groups of data. One major takeaway from Table 2 is that there are about double the data point in Train B and C compared to Train A. This will most likely effect the training ability of this dataset. Another takeaway from these summaries is that Dst is lowest in Period A, and at similar means in Periods B and C. However, the standard deviation in Period A is much higher, suggesting that it is a more intense period. The only other feature that somewhat follows this pattern for Period A is sunspots, suggesting that this feature may be correlated with Dst.

Table 2: Descriptive Statistics by Period

Variable	Metric	Period A	Period B	Period C
bt	**count**	1,575,012	3,084,130	3,407,290
	mean	7.1	5.6	4.9
	std	3.7	2.8	2.8
	min	0.1	0.1	0.0
	25%	5.0	3.8	3.2
	50%	6.3	5.0	4.3
	75%	8.2	6.7	5.8
	max	80.5	42.2	61.0
density	**count**	1,406,866	3,053,066	3,247,498
	mean	4.9	5.7	3.0
	std	4.8	4.7	3.2
	min	-	-	-
	25%	2.3	2.8	1.2
	50%	3.6	4.8	2.2
	75%	5.8	7.2	3.9
	max	184.9	199.7	166.6
speed	**count**	1,406,902	3,048,354	3,247,509
	mean	436.9	429.8	428.6
	std	94.5	96.3	106.7
	min	235.0	228.4	-
	25%	370.8	357.9	349.4
	50%	418.2	410.1	404.2
	75%	484.1	482.0	490.6
	max	1,011.5	1,198.5	1,064.0
temperature	**count**	1,396,624	3,036,895	3,147,033
	mean	105,917.4	136,424.7	98,588.9

	std	100,098.3	149,788.5	89,556.6
	min	10,000.0	1,496.0	-
	25%	43,649.0	37,414.0	40,074.0
	50%	79,238.0	85,524.0	71,521.0
	75%	132,550.0	187,325.0	131,088.0
	max	6,223,700.0	4,206,672.0	5,751,308.0
smoothed_ssn (sunspots)	**count**	40.	72	80
	mean	136.9	51.9	24.3
	std	34.6	39.2	19.0
	min	65.4	3.9	2.2
	25%	108.4	15.3	7.8
	50%	151.5	43.2	20.5
	75%	164.4	91.2	38.5
	max	175.2	116.4	69.5
Dst	**count**	28,824	52,584	58,464
	mean	(16.6)	(9.7)	(9.6)
	std	26.1	16.4	16.5
	min	(387.0)	(223.0)	(374.0)
	25%	(26.0)	(17.0)	(16.0)
	50%	(12.0)	(7.0)	(7.0)
	75%	(1.0)	1.0	-
	max	65.0	59.0	67.0

Visualizations of these variables are shown in Figures 1-6 below for the first 3000 datapoints. This subset was chosen so that the general pattern of the data could be more easily discerned. These charts do not show the data aligned by time scale for sunspots and Dst. There are many noticeable spikes in the data, but they generally fall in line with the magnitudes expected for each of the variables. As noted, extreme Dst events can be as low as -1600 nT, so seeing spikes into the -100s is not surprising.

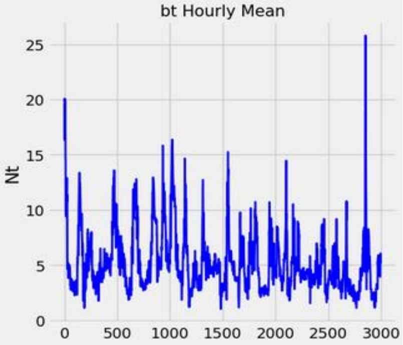

Figure 1: Bt Values

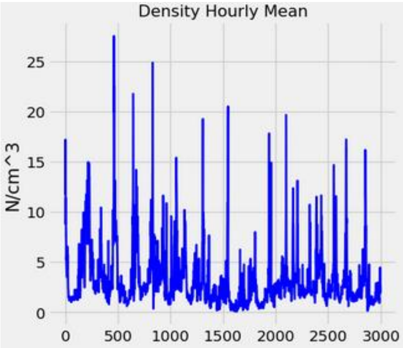

Figure 2: Density Values

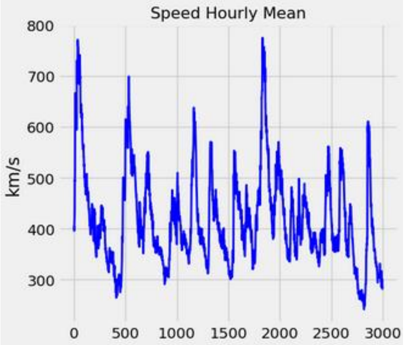

Figure 3: Speed Values

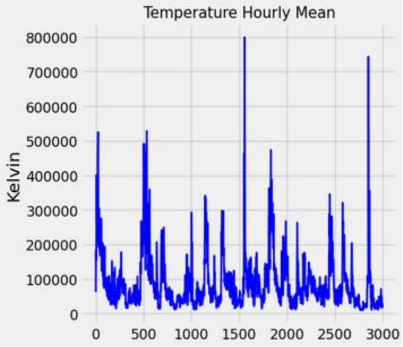

Figure 4: Temperature Values

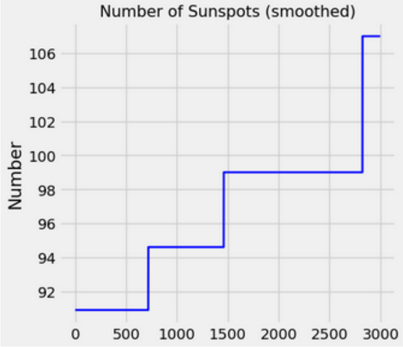

Figure 5: Sunspot Values

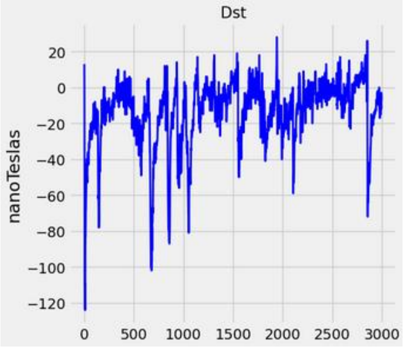

Figure 6: Dst Values

To complete the data pre-processing, the data from the three CSVs is combined dataframe based on the timestamp, null data is removed, then aggregate the data into hourly means and standard deviations. The features data (everything except Dst) is then shifted by one time step so that features$_{t-1}$ aligns with Dst$_t$. This action is taken to approximate the predictive manner of the features into the future Dst value.

For final processing before modeling, all of the X and y data is put into an array with scaled values between 0 and 1. Training and test sets are also assigned for each period. The TrainTestSplit function cannot be used because the data should not be divided randomly, as the sequencing of the data is crucial to the predictive power of the model. The data is instead divided into training and testing groups by timestamp. The data is then ready to implement into a model for fitting.

Model Selection

The goal for this project is to create a model that predicts Dst, a key solar wind indicator, based on several input features. The dataset is a time series, which is different than the model examples used in class. Not only will the predicted Dst factor be dependent on the multiple X factors, but it will also be dependent on their patterns over time. Specifically, I expect that leading up to a solar wind event, the indicators will display a pattern that strongly predicts Dst.

The Long Short-Term Memory (LSTM) neural network model is a strong option for this model, due to the nature of time series events. LSTM is a type of recurrent neural network (RNN) that is different from traditional feed-forward neural networks. RNNs are unique in that they keep past information about inputs for a specified amount of time. This is important for the data presented in this paper as it is expected that there is a gradual buildup of inputs that predict an extreme solar event. LSTM therefore had feedback connections that make it possible for it to process an entire sequence of data. A standard RNN will look at new inputs as well as the previous output to make a prediction (Figure 7). LSTM builds upon this by adding four layers within that cell for each iteration (Figure 8). During this process, LSTMs selectively remembers or forgets information.

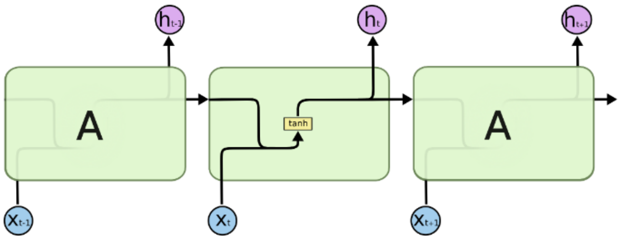

Figure 7: The repeating module in a standard RNN contains a single layer [8]

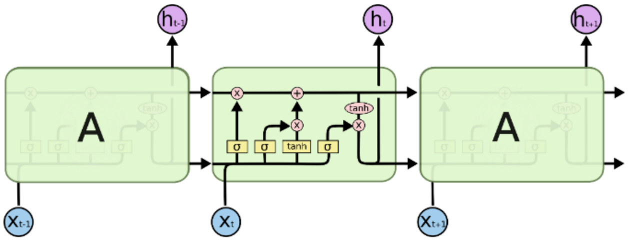

Figure 8: The repeating module in an LSTM contains four interacting layers [8]

The first stop in this function is the "forget gate" where the cell determines which data from previous iterations is remembered. The cell takes in X_t and h_{t-1}, and an assigned sigmoid function outputs a value for each number in the cell state. Numbers closer to 0 will mean that the input is “forgotten” for this state, and numbers closer to 1 will be included.

The next stop is the input gate, which decides which of the new input data will be included in the model. X_t and h_{t-1} is inputted, and a tanh function outputs weights from -1 to 1. At this stop, the sigmoid is multiplied to the tanh function and added to the operation. This allows for new, nonredundant information to be added to the cell.

The final output stop determines what information from the previous functions is most likely to be the correct output. The function once again scales the values using tanh, then filters using the X_t and h_{t-1} values from the previous vector, and finally multiplies these vectors together. This function will diminish values that are less likely, making the output more likely to be the correctly predicted value.

This function is chosen because it will be able to include past observations to fit the predicted Dst value at any one point in the dataset. This will result in a dynamic model that will be able to handle this complicated and large real-world dataset.

Results and Evaluation

This project consisted of building a LSTM model and running it for each of the three periods provided in this dataset. The input variables for this model were hourly means of the following metrics:

Interplanetary-magnetic-field component magnitude (bt) mean and standard deviation (std), proton density mean and std, wind bulk speed mean and std, temperature mean and std, previous Dst value, and number of sunspots. The target variable was the current hourly mean Dst.

The parameters of the LSTM model were adjusted repeatedly and tuned to achieve the lowest root square mean deviation (RSME) between the test y and predicted y datasets. The parameters that were tested and their final specification are listed in Table 3 along with a rationale. RSME was taken as the main validation metric for this model.

Table 3: Parameters Tested and Used for Semester Project

Parameter	Final	Tested	Rationale
Training Set	80% Training 20% Test per period	Range from 30-85% training data tested	80/20 split is standard and showed favorable results in RSME output
Kernels	300	Range from 150-300 tested	Increased kernels generally help the model, and no signs of overfitting were present
Loss Function	Mean Absolute Error (MAE)	Tested MAE and Mean Square Error (MSE)	MSE errors are typically higher than MAE errors
Optimizer	Adam	Adam	Adam stochastic gradient descent appropriate for this model and prevents gradient exploding
Dropout	0%	Range from 0-40% tested	Low rates had no effect and high rates weakened the model
Epochs	30	Range from 5-50 tested	Positive effect of increased Epochs plateaued around 30
Batch size	96	Range from 24-200 tested	96 showed strongest RSME effects and is a reasonable time period for prediction
Shuffle	False	False	False appropriate as data is sequential
Dense Layer	Included	Tested with and without layer	Dense layers typically improve neural net models

Models were created separately for each period, as the LSTM model relies on sequential modeling of times series data. If the periods had been combined into one dataset, there would have been incorrect overlap where the datasets met in the data frame. Further, the periods in this dataset are nonconsecutive and it was not shared when they took place, so it would even be difficult to be confident that they are in order from Periods A to C.

The Figures below show the outcome of the final model. Period A had the highest RSME at 6.49, compared to 3.12 and 3.02 for Periods B and C. Period A had had about half the samples of B and C, and this seems to be the main cause of the higher RSME. Including more samples in the training set had a large effect on the accuracy for all datasets, so it is logical that this hurt Period A's performance.

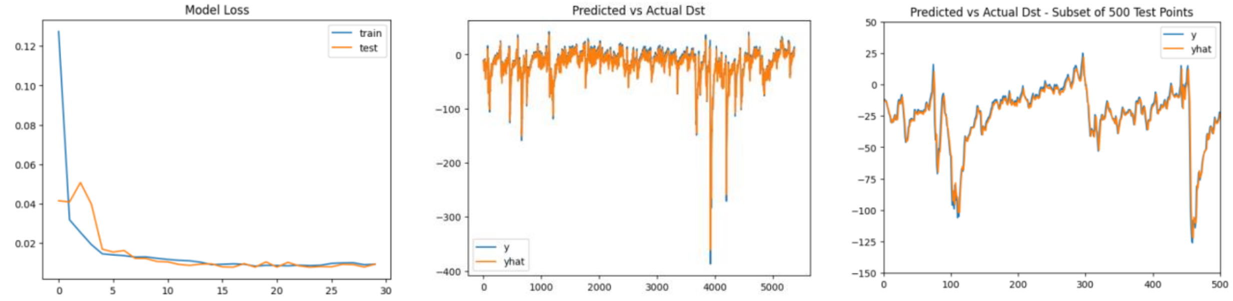

Figure 9: Period A. RSME = 6.49

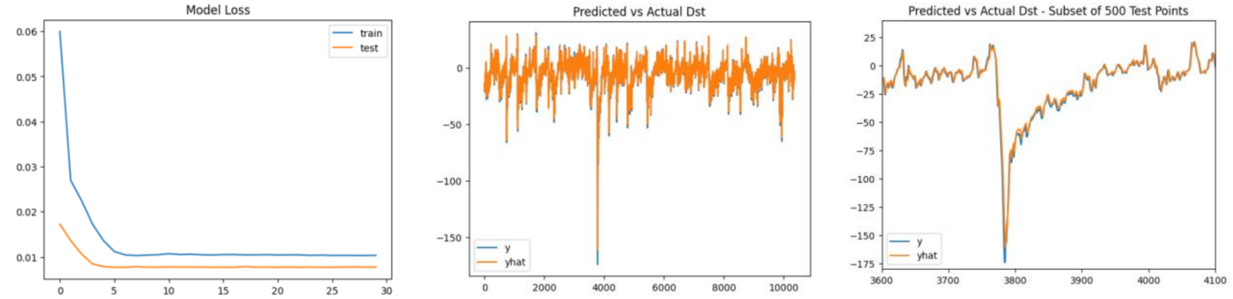

Figure 10: Period B. RSME = 3.12

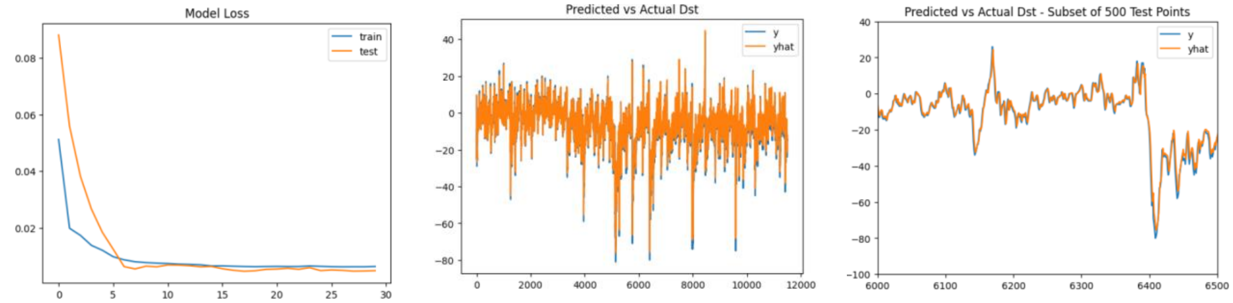

Figure 11: Period C. RSME = 3.02

Each of the Figures above includes a model loss chart during the epoch iterations, predicted vs actual Dst, and a zoomed in view of a subset of the predicted vs actual Dst to help visualize the accuracy of the model. Although Period A had a slightly poorer performance, all of the models were decently accurate in prediction, and did not show major signs of overfitting. The subsets in the far right graphs were chosen as they show significant dips in Dst, which is associated with strong solar storms. The models were very accurate in predicting these dips, although it generally underestimated the severity slightly.

Ethics

Because this dataset is looking at physical phenomenon, there are likely not ethics considerations to consider. However, there are certainly biases that are present in the dataset that need to be considered. For example, the data is preprocessed into three time periods, which were deemed by the processors to be good representations of Dst cycles, and thus good for training. Undoubtedly, assumptions about what makes a good representation were made, and thus the user cannot escape those assumptions.

Any downstream uses of this model have ethical considerations. For example, predicting a severe magnetic storm would be used to create early warning systems for humans. The methods in which those warnings were distributed and targeted could be of some ethical concern. For example, they could favor certain industries or regions that are deemed crucial or with the most to lose from a storm.

Future Work

Given more time and experience with these models, there are many more opportunities for continued exploration with this dataset.

For simplicity, the directional variables from the dataset (B_x, B_y, B_z, B_θ) were removed. The directional analysis required substantial data preprocessing and were a different type of data compared to the other metrics (direction vs magnitude). Adding this information into the model would cause complications and would likely make the neural network implementation more difficult. However, after reviewing research performed by other data scientists, there are strong indications that the direction is predictive of the strength of Dst, so including these variables in the model could be helpful for the overall fit.

Based on the results of my analysis, there are a few items that warrant further analysis and work. The first is adjustment of some of the parameters, specifically the dropout rate. This function should be helpful with reducing noise in the model by dropping out unnecessary data from the LSTM cell. However, it was unimpactful at low levels (<10%) and hurt the model at higher levels (10-40%). I believe this result warrants additional investigation because one of the most determinant parameters of model success was Batch Size, or the number of previous datapoints that are considered in the LSTM model. Generally, higher was better, up till the 100s of batch size. This is a significant batch size, and the final 96 that was chosen represents 4 days of previous data. It may be the case that a higher batch size with a drop out layer would be favorable for the model.

A final area of work is in combining Periods A, B, and C and testing models across periods. I chose not to combine the periods for this model because although it would increase the training and testing dataset size, the periods are not consecutive, and the sequential nature of LSTM would create errors and incorrect analysis along the points where the model met. Additional investigation and improvements of the structuring of the model should allow all three periods to be used in the training of a single model. Additionally, another option would be to use another source for this data where it is not pre-separated into periods.

Sources

[1] O'Callaghan, J. (2019, September 24). New studies warn of cataclysmic solar superstorms. Scientific American. Retrieved October 14, 2022, from https://www.scientificamerican.com/article/new-studies-warn-of-cataclysmic-solar-superstorms/

[2] Tsurutani, B. T. (2003). The Extreme Magnetic Storm of 1–2 September 1859. Journal of Geophysical Research, 108(A7). https://doi.org/10.1029/2002ja009504

[3] Maynard, T., Smith, N., &amp; Gonzalez, S. (2013). Solar Storm Risk to the North American Electric Grid. Atmospheric and Environmental Research.

[4] NASA, NOAA. NASA and NOAA Satellites Solar-Wind Dataset. Retrieved September 28, 2022, from https://www.kaggle.com/datasets/arashnic/soalr-wind?resource=download.

[5] Zhang, L. Q., Wang, C., Wang, J. Y., & Lui, A. T. Y. (2019). Statistical properties of the IMF clock angle in the solar wind with northward and southward interplanetary magnetic field based on Ace Observation from 1998 to 2009: Dependence on the temporal scale of the Solar Wind. *Advances in Space Research*, *63*(10), 3077–3087. https://doi.org/10.1016/j.asr.2019.01.023

[6] Larrodera, C., Nikitina, L., & Cid, C. (2022). Estimation of the solar wind extreme events. https://doi.org/10.5194/egusphere-egu22-730

[7] Upendran, V., Cheung, M. C., Hanasoge, S., & Krishnamurthi, G. (2020). Solar wind prediction using Deep Learning. *Space Weather*, *18*(9). https://doi.org/10.1029/2020sw002478

[8] Understanding LSTM networks. Understanding LSTM Networks -- Colah's blog. (n.d.). Retrieved October 8, 2022, from http://colah.github.io/posts/2015-08-Understanding-LSTMs/