Solar wind analysis

Abstract

The OMNI2 dataset, comprising 289,296 records of near-Earth solar wind and magnetic field data from January 1, 1988, to December 31, 2020, provides hourly measurements of magnetic field components (Bx, By, Bz), plasma density, plasma bulk velocity (V), and the geomagnetic index (Dst). Despite 13% of data being missing, exploratory analysis revealed that most variables had near-normal distributions. Correlation analysis indicated low correlations among most variables, except a -0.47 correlation between plasma bulk velocity and Dst, showing that increased solar wind velocity amplifies geomagnetic disturbances. Autocorrelation patterns, excluding Bz, exhibited peaks at a 650-hour lag, reflecting the Sun's rotational period.

Linear regression models yielded an R^2 of 0.314, suggesting limited predictive power for geomagnetic storm prediction. Principal Component Analysis (PCA) identified two components that explained over 56% of the variance. Magnetic storm classification models—Naïve Bayes, Linear Discriminant Analysis (LDA), and Logistic Regression—were trained and evaluated. Logistic Regression performed best with an AUC score of 0.839, surpassing Naïve Bayes (0.784) and LDA (0.758). This analysis highlighted logistic regression as a best predictive model for magnetic storm classification.

Dataset

The OMNI2 dataset consists of 289296 records of near-Earth solar wind and magnetic field data, including measurements of components of magnetic field (Bx, By, Bz), plasma density (density), plasma bulk velocity (V) and geomagnetic index (Dst). The data was collected hourly from multiple spacecraft missions during January 1, 1988 and December 31, 2020. The data is stored in a CSV format, featuring a timestamp for each record. Occasionally there are missing values in the dataset which have almost correlation perfect correlations among variables and mostly located during 1988 and 1994 and on average amount for 13% of the data records.

Analysis

The analysis started with general inspection of the dataset. The dataset consists of 289296 rows x 7 columns, which indicate date stamps and the values for different solar wind and magnetic field metrics. The exploratory analysis was performed on missing values.

Missing values of magnetic field components (Bx, By, Bz) represent 12.55% of the total dataset and account for approximately 36300 records. For density, missing values represent 14.52% and account for 42004 records from the dataset. For plasma bulk velocity (V), missing values represent 12.91% of the dataset and accounted for 37349 records. Dst column was found to have no missing values. After plotting missing values for each variable as a function of year, distributions of

missing values were found to be similar among variables, with majority of missing values located in period between 1988 and 1994. Missing values were defined as missing at random (MAR) and likely appeared due to the faultiness of the measurement systems, as the data is missing in batches for all metrics [1]. Missing data was deleted, as it represents relatively small amount of the dataset, and it will not have substantial impact on the consequent analysis.

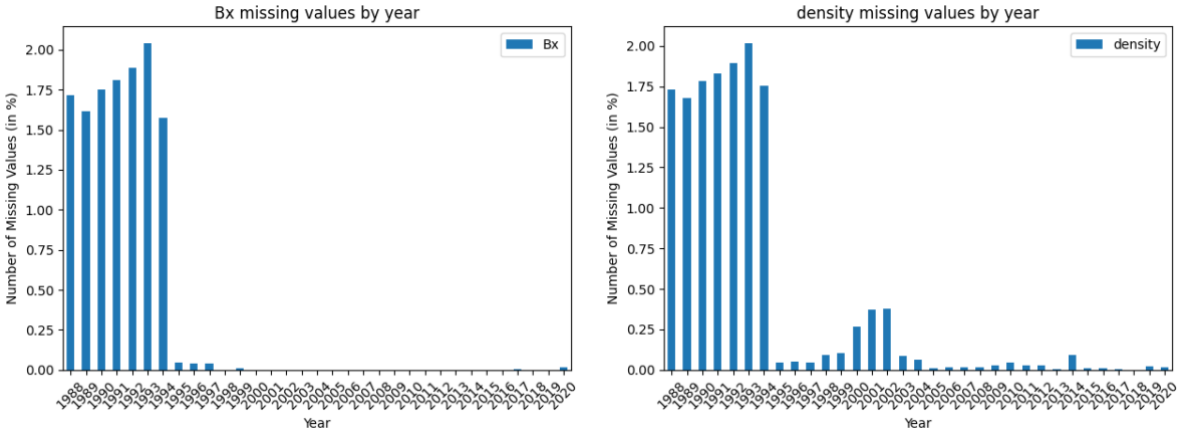

Figure 1: Missing values distribution for Bx and density In the period between 1988 and 2020.

Histograms for each metric of the dataset were created to analyse general distribution of each variable. As can be seen from figure 2, almost all metrics have distributions which are close to normal and centre around 0, except for plasma bulk velocity (V) and density. Distribution of these two metrics have visible skew to the right, which indicates the asymmetries of distribution.

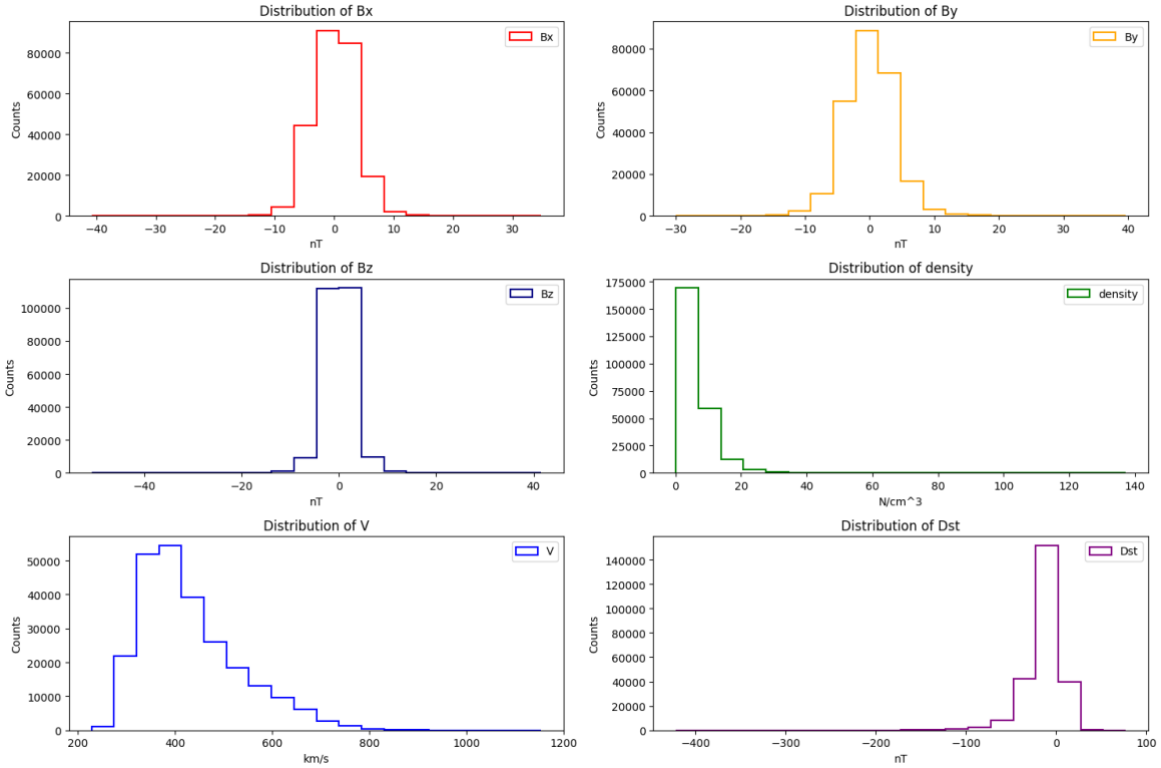

Figure 2: Distribution histograms of each metric in the dataset.

Key metrics such as mean, standard deviation, and quartile values were calculated for each metric, which gives more detailed information about distribution of variables.

	Bx, in nT	By, in nT	Bz, in nT	Density, in N/cm^3	V, in km/s	Dst, in nT
Mean	0.0082	0.0032	-0.0339	6.4424	431.5562	-12.9904
Median	0.0	0.0	0.0	5.0	408.0	-9.0
Mode	-2.7	-2.1	0.0	2.7	376	-4.0
Standard Deviation	3.52	3.83	3.02	5.19	102.11	20.27
1st Quartile (25%)	-2.6	-2.4	-1.5	3.2	356.0	-21.0
2nd Quartile (50%)	0.0	0.0	0.0	5.0	408.0	-9.0
3rd Quartile (75%)	2.5	2.4	1.4	8.0	486.0	-1.0
Interquantile Range (IQR)	5.1	4.8	2.9	4.8	130.0	20.0

Table 1: Key metrics including mean, median, mode standard deviation, quartile values and interquartile range for each variable in the dataset.

Correlation heatmap was then plotted to analyse the relationship between variables.

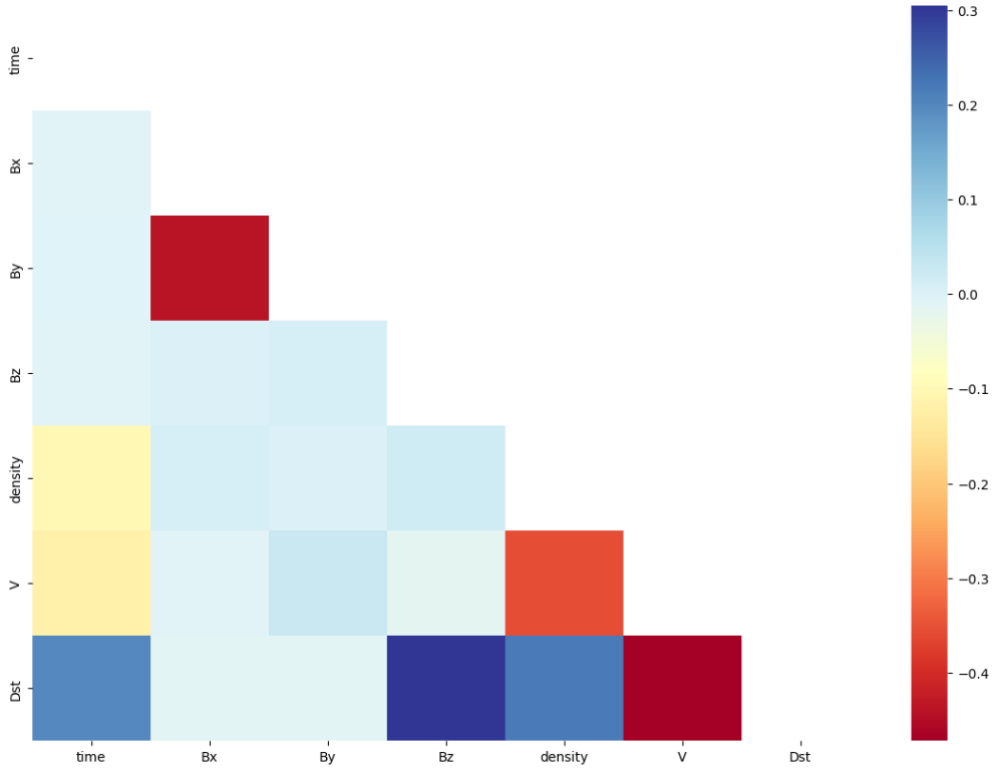

Figure 3: Correlation heatmap indicating relationships between variables.

There are no particularly high correlations among variables, meaning they usually don’t follow same trends. Plasma bulk velocity (V) and geomagnetic index (Dst) have the highest negative correlation among variables with corresponding value of -0.47. Increase in plasma speed means more disturbance to Earth’s magnetic field, which is indicated by geomagnetic index [2]. High negative geomagnetic index values represent stronger disturbances to magnetic field. Higher bulk plasma velocity means lower value of Dst, hence negative correlation.

There is a correlation value of 0.30 between vertical component of magnetic field Bz and geomagnetic index (Dst). When the Bz component of the interplanetary magnetic field (IMF) is negative (pointing southward), it aligns with Earth's magnetic field, allowing solar wind particles to penetrate Earth’s magnetosphere more easily. This interaction can increase geomagnetic activity, leading to lower (more negative) values of the Dst index, which measures geomagnetic disturbances.

Conversely, when Bz is positive (northward), it opposes Earth's magnetic field, reducing the penetration of solar wind particles and leading to quieter geomagnetic conditions and higher (less negative) Dst values. The correlation is only moderate (0.30) because other factors, such as solar wind velocity and density, also significantly influence geomagnetic activity [3]. Other variables have near zero correlation values, indicating almost non-existent correlations between variables.

Autocorrelation analysis was then performed for each of the variable in the dataset and result was plotted between in range between 1 and 1000 hours. Autocorrelations for all variables except vertical component of magnetic field (Bz) follow the same pattern and have distinct peaks at approximately 650 hours of time lag. This happens because Sun has rotation period of approximately 27 days, which correspond to the autocorrelation peaks observed in variables. Because of the rotation, measurements phases coincide approximately every 27 days and have higher similarities between each other, hence higher correlation.

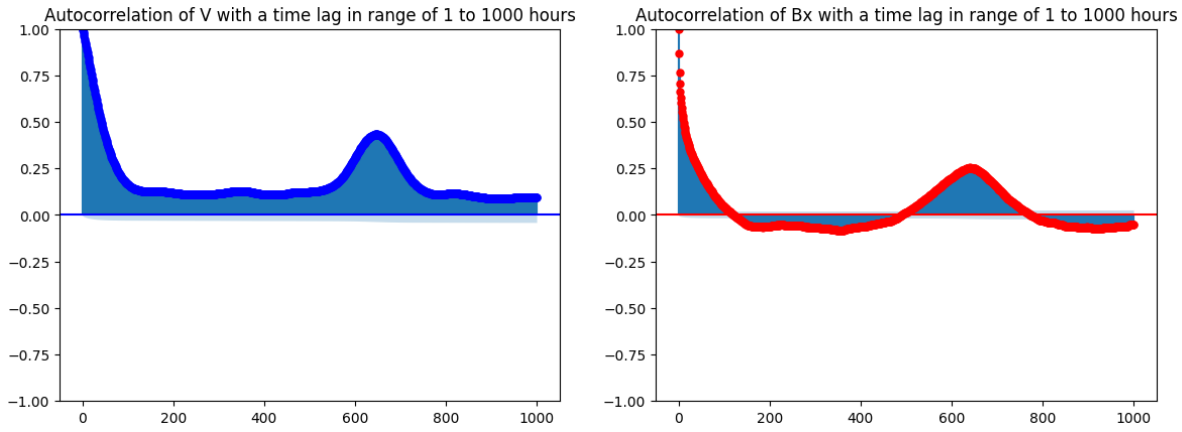

Figure 4: Autocorrelation of plasma bulk velocity (V) and horizontal component of magnetic field (Bx) in range of 1 to 1000 hours.

Linear regression model was then fit to data to assess if the nature of the data is linear. All variables were then fit to linear regression except geomagnetic index (Dst), which was the variable to be predicted by linear regression. After linear regression was fit, the following key metrics were obtained, shown on table 2.

R-squared	F-statistic	Log-Likelihood	AIC	P-value (const)
0.314	2.261e+04	-1.0450e+06	2.090e+06	0.000

Table 2: Key metrics for linear regression model with magnetic field components (Bx, By, Bz), solar wind density (density) and plasma bulk velocity (V) as predictors and geomagnetic index (Dst) as predicted variable.

The linear model doesn't have great predictive power in the given example, as it has R-value of 0.314. This R-squared value means this model accounts for 31.4% of the variation in the independent by the dependent values, which is considered low [4]. High F-statistic value represents that the result of linear regression occurred not due chance. The probability mass function (PMF) or probability density function (PDF) multiplied at each data point evaluation yields the log-likelihood. Low log-likelihood indicates that linear model doesn't have a great fit to the data. Akaike Information Criterion (AIC) is used to compare models by evaluating how well they fit the data and penalising for the complexity of the model. The lower the AIC, the better. F-statistic, log-likelihood and AIC needs to be used in comparison to linear regression models performed on the same data to be effectively compared. Finally, p-value shows the significance of the given experiment, with measurements with value below 0.05 considered significant. Based on the results it was determined that the data is likely of non-linear nature.

The influence plot was created to analyse outlier data points which could potentially impact the fit of the linear regression model. It was created using 3 metrics mentioned below.

1. Standardized Residuals: These are residuals (errors or deviations of predicted values from observed values) divided by their standard deviations. They are plotted on the y-axis. High residuals (positive or negative) indicate that a point deviates significantly from the predicted trend.
2. Hat Values (Leverage): These measure the influence of each observation on the predicted values, based on its location in the predictor space. They are plotted on the x-axis. High leverage points are data points that can disproportionately influence the model's predictions.
3. Cook's Distance: This is a measure of the overall influence of a data point, considering both the residual and the leverage. The size of each bubble in the plot is proportional to Cook's distance. Points with larger Cook's distance are potentially influential observations that could affect the model's fit.

Out of all points produced on the plot, there is one clear influential point with large radius, indicating large Cook's distance. Number next to the bubble indicates the index of the datapoint, showing its location in the dataset. After analysing data around the specified index, it was discovered that the time of recording (November 1991) coincides with great geomagnetic storm at the time. During this storm plasma bulk velocity (V) exceeded 900 km/s and it is considered the largest geomagnetic storm yet associated with the eruption of a quiescent filament [5].

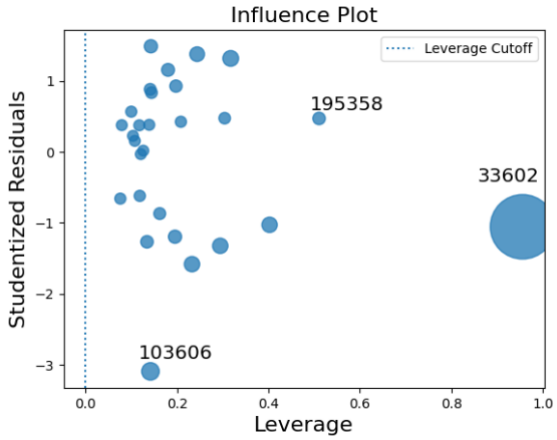

Figure 5: Influence plot with hat values (leverage) shown on the x axis and studentized residuals on y axis. Size of a bubble represents the Cook's distance and indicates the extent to which point affects linear regression fit.

By using backward elimination, unnecessary variables were deleted from the predictors to maximize performance of the model as a function of predictors. By trying different combinations of predictors, it was discovered that linear regression model with vertical component of magnetic field (Bz), solar wind density (density) and plasma bulk velocity (V) has the best ratio of fit results as a function of variables used. The key statistics for this model are shown on the table below.

R-squared	F-statistic	Log-Likelihood	AIC	P-value (const)
0.312	3.596e+04	-1.0108e+06	2.022e+06	0.000

Table 3: Key metrics for linear regression model with vertical magnetic field component (Bz), solar wind density (density) and plasma bulk velocity (V) as predictors and geomagnetic index (Dst) as predicted variable.

It has lower R-squared value compared to previous model, but slightly improved F-statistic, log-likelihood and AIC values, which can be considered a good trade-off between the resulting performance of the model and amount of information used for the fit.

Principal Component was performed on OMNI2 data. Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of large datasets while preserving as much of the data's variation as possible. It helps simplify the data without losing critical information by transforming the original variables into a new set of variables called principal components. These principal components are orthogonal (meaning they are at right angles to each other, indicating no correlation among them) and are ordered so that the first few retain most of the variation present in all the original variables [6].

A scree plot shows the variance in the data as explained or captured by each principal component shown as eigenvalues on y axis against principal components on the x axis. It is typically used to determine the number of principal components to retain to efficiently summarize the data. According to Kaiser criterion, components with eigenvalues greater than 1 are chosen to be retained for the analysis [7]. Such

eigenvalues suggest that more variance than a single variable is explained by the corresponding component, given that the variable accounts for a unit of variance. In this case, principal components indicated by 0 and 1 on the graph could be retained for the analysis, and principal component indicated by 2 has eigenvalue of approximately one, suggesting it might be marginally useful depending on the specific needs for variance explanation. Principal components have ratio weights of 0.289, 0.270, 0.199, 0.129 and 0.111 respectively. First two and first three components represent approximately 56% and 75.9% of the total component weights, respectively.

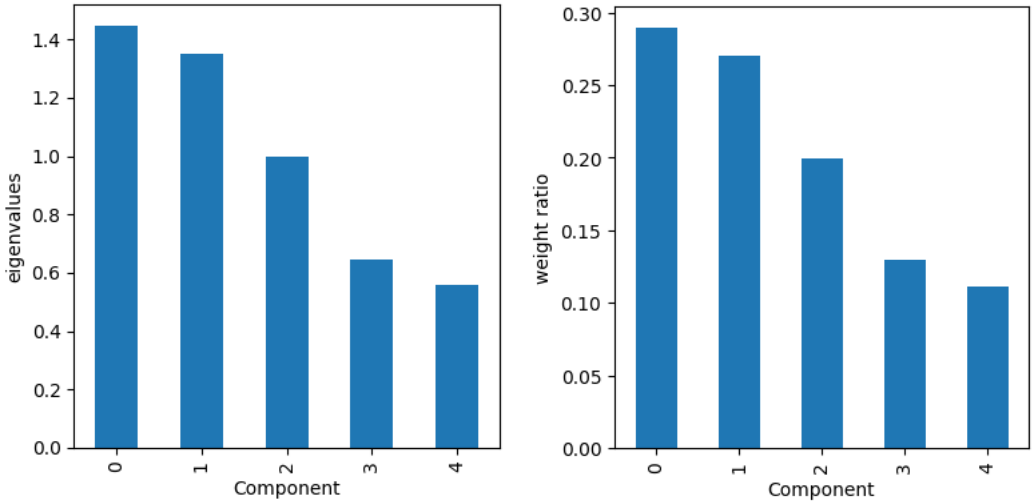

Figure 6: Scree plots showing principal components on the x axis and their associated eigenvalues shown on the y axis on the left, and principal components with corresponding weight ratio on the right.

In PCA loadings represent weights assigned to each original variable in the dataset, which help define the composition of principal components. If the dataset is represented as a matrix, X, and the weights as a vector, w, for a principal component, then the principal component can be calculated as a product Xw. This operation projects the original data onto a new axis defined by the weights [8]. The magnitude of the weights indicates the strength of the influence of each variable on the principal component. A positive weight suggests that as the variable increases, the principal component also increases, assuming other variables are constant. Conversely, a negative weight suggests an inverse relationship with the principal component.

Principal component	Bx	By	Bz	density	V
0	-0.685	0.686	-0.002	-0.161	0.185
1	-0.175	0.170	0.070	0.687	-0.681
2	-0.020	0.001	-0.997	0.044	-0.052
3	0.112	0.089	-0.008	0.700	0.700
4	0.698	0.702	-0.012	-0.097	-0.104

Table 4: PCA loadings for each principal component of every variable in the dataset.

In principal component 0 there is strong negative loading for Bx (-0.685) and strong positive loading for By (0.686). This indicates that Bx and By vary inversely with each other in this component. Other variables have relatively low influence on this component, with the next significant being density (-0.161) and V (0.185).

For principal component 1 there is a similar situation as in the previous case, with density and plasma bulk velocity (V) having positive and negative loadings of 0.687 and -0.681 respectively, meaning they tend to vary inversely with each other in given component. In principal component 2 the Bz variable is almost solely responsible for driving variance in this component because of its dominant negative loading (-0.997). Deconstructing principal components this way helps gain understanding of relationships between variables and allows for more in depth analysis.

Three magnetic storm classification models (Naïve Bayes, Linear Discriminant Analysis and Logistic Regression) were then built and compared using accuracy and AUC scores. The new column was added to the dataset, classifying measurements with geomagnetic index (Dst) below -20 as a geomagnetic storm. Dataset was then split into training and testing sets, with measurements from 1988 to 2019 as training and measurements from 2020 as testing datasets.

Data from 1988 to 2019 was then analysed to understand distribution between instances where data was classified as magnetic storm or not. There were 176918 measurements classified as not a magnetic storm and 60840 measurements classified as a magnetic storm, which accounted for 25.59% and 74.41% of the training dataset respectively.

Most classification models require even distribution between classes, so that there will be no bias present in the model. If number of measurements in one class is prevalent, model can learn to classify majority of measurements contributing to prevalent class and automatically have higher accuracy [9]. To avoid this, the dataset have to be rebalanced. Because the dataset is quite large, random measurements from the prevalent class can be deleted to balance the dataset. After deleting 65% of the prevalent dataset at random, we get nearly even value distribution with 61921 measurements classified as not magnetic storm and 60840 as magnetic storm, representing nearly 50% of measurements for each class.

After training models, they were compared using data from 2020 by plotting ROC curves and calculating AUC scores. AUC score associates with general predictive ability of classifier model. AUC represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance [10]. Logistic regression model was found to be the best model among three classifier models with an AUC score of 0.839, compared to Naïve Bayes AUC score of 0.784 and linear discriminant analysis (LDA) AUC score of 0.758.

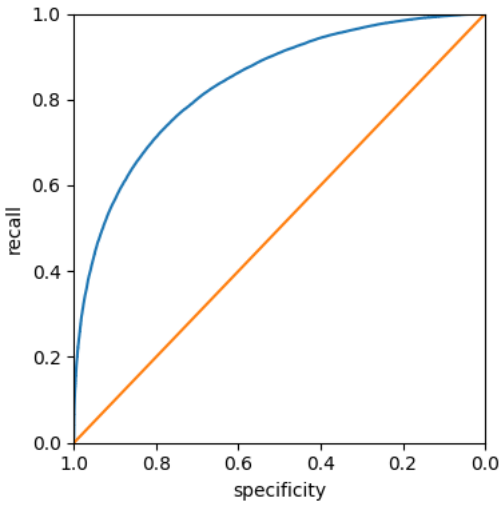

Figure 7: Reciever operating curve (ROC) of Logistic regression curve with AUC score of 0.839.

Conclusion

This paper examined the OMNI2 solar wind dataset to identify trends and to develop predictive models for magnetic storm classification. The exploratory data analysis revealed that the dataset had approximately 13% missing values, mostly between 1988 and 1994. These missing values were removed from the analysis. The majority of metrics exhibited approximately normal distributions, while plasma bulk velocity (V) and density displayed a right-skewed distribution.

Correlation analysis found that most variables had low correlation values. The highest negative correlation of -0.47 was observed between plasma velocity and the geomagnetic index. This indicates that as solar wind velocity increases, geomagnetic disturbances grow stronger. Autocorrelation analysis showed that all variables except the vertical magnetic field component (Bz) had distinct peaks at a lag of 650 hours, which corresponds to the Sun's rotational period of roughly 27 days.

Linear regression models demonstrated limited predictive power, with an R^2 of 0.314 for predicting the geomagnetic index. This suggests that the relationship between solar wind components and geomagnetic storms is not linear. Principal Component Analysis (PCA) identified two principal components that explained over 56% of the data variance. These components primarily showed inverse relationships between Bx and By, and between density and plasma bulk velocity (V).

Three magnetic storm classification models (Naïve Bayes, Linear Discriminant Analysis, and Logistic Regression) were trained and tested on the data from 1988 to 2019 and 2020 respectively using AUC scores. Logistic Regression was the best-performing model, achieving an AUC score of 0.839, outperforming both Naïve Bayes (0.784) and Linear Discriminant Analysis (0.758).

In conclusion, future research should enhance solar wind data analysis by improving methods to handle missing data, employing nonlinear models such as Random Forests or Neural Networks to capture complex relationships, and refining feature selection to gain deeper insights. Incorporating time-series models like LSTM networks, clustering analysis to identify distinct solar wind

patterns, ensemble modeling for better classification, and domain-specific data augmentation could improve predictive performance, providing more accurate geomagnetic storm forecasts and insights into space weather.

Dataset

OMNI2 dataset: https://spdf.gsfc.nasa.gov/pub/data/omni/low_res_omni/

References

[1] Mack, C., Su, Z., & Westreich, D. (2018, February 1). *Types of missing data*. Managing Missing Data in Patient Registries - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK493614/

[2] *Geomagnetic Storms | NOAA / NWS Space Weather Prediction Center*. (n.d.). https://www.swpc.noaa.gov/phenomena/geomagnetic-storms#:~:text=Another%20solar%20wind%20disturbance%20that,rotating%20interaction%20regions%2C%20or%20CIRs.

[3] Osmane, A., Dimmock, A. P., Naderpour, R., Pulkkinen, T. I., & Nykyri, K. (2015). The impact of solar wind ULF Bz fluctuations on geomagnetic activity for viscous timescales during strongly northward and southward IMF. *Journal of Geophysical Research. Space Physics*, *120*(11), 9307–9322. https://doi.org/10.1002/2015ja021505

[4] Taylor, S. (2023, November 22). *R-Squared*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/r-squared/#:~:text=The%20most%20common%20interpretation%20of,explained%20by%20the%20regression%20model.

[5] *The Great Geomagnetic Storm of 9 November 1991: Origin in a disappearing solar filament*. (n.d.). NASA/ADS. https://ui.adsabs.harvard.edu/abs/2006AGUSMSH43A..06C/abstract#:~:text=The%20great%20storm%20of%209,1982%20and%2016%20July%202000.

[6] Jaadi, Z. (2024, February 23). *A Step-by-Step Explanation of Principal Component Analysis (PCA)*. Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[7] *APA PsycNet*. (n.d.). https://psycnet.apa.org/fulltext/2016-15750-001.html

[8] *Data analysis in the geosciences*. (n.d.). http://stratigrafia.org/8370/lecturenotes/principalComponents.html#:~:text=The%20elements%20of%20an%20eigenvector,to%20a%20particular%20principal%20component.

[9] Brownlee, J. (2020, January 14). *A gentle introduction to imbalanced classification*. MachineLearningMastery.com. https://machinelearningmastery.com/what-is-imbalanced-classification/#:~:text=Imbalanced%20Classification%20Problems,-The%20number%20of&text=Imbalanced%20classification%20refers%20to%20a,is%20instead%20biased%20or%20skewed.

[10] Bhandari, A. (2024, April 23). *Guide to AUC ROC Curve in Machine Learning : What is specificity?* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/