Analysis of the NASA OMNI2 Dataset

Illias Dadashov

June 13, 2024

Abstract

Solar activity significantly influences Earth's magnetosphere, ionosphere, and thermosphere, impacting satellite operations, communication systems, and power grids. Comprehending the interactions between solar wind and geomagnetic activity is essential for forecasting space weather events and reducing their negative effects. This report presents the results of the analysis and modeling of the NASA OMNI2 dataset, which contains solar wind and geomagnetic activity data, conducted using the Python programming language and relevant packages. During the project, the exploration of the dataset was conducted, followed by the analysis of missing/fill values and the creation of histograms for key metrics distributions. Additionally, an autocorrelation analysis was performed for these metrics. In the modeling phase, an OLS Regression model was developed to predict the strength of geomagnetic storms, with an R-squared parameter of 0.178 for the test sample. Principal Component Analysis was also conducted, resulting in the creation of a Scree plot and a histogram of the weights. In the final part, Gaussian Naive Bayes, Linear Discriminant Analysis, and Logistic Regression models were obtained, with corresponding ROC-AUC values of 0.768, 0.761, and 0.761, respectively. The results of this work can be used as a methodology for further analysis of solar wind activity in the future. Additionally, these results can serve as a teaching aid for university students.

1 Introduction

The solar wind is a continuous stream of charged particles, primarily electrons and protons, that flow outward from the Sun's corona into space. This stream varies in speed, typically ranging from 300 to 800 kilometers per second, and carries the Sun's magnetic field with it. The solar wind significantly influences the dynamics of our solar system, affecting planetary atmospheres, magnetospheres, and space weather phenomena. When the solar wind interacts with Earth's magnetic field, it can cause geomagnetic storms, auroras, and disruptions in satellite communications and power grids.

NASA's OMNI dataset compiles near-Earth solar wind and magnetic field data from multiple spacecraft missions, along with other parameters. This dataset offers valuable insights into the behavior of the solar wind and its interactions with Earth's magnetosphere. Researchers and scientists use the OMNI dataset for various studies related to space weather, magnetospheric physics, and solar-terrestrial interactions. It is a crucial resource for understanding and predicting space weather events and their potential impacts on technological systems and human activities in space and on Earth.

The Geomagnetic Disturbance Storm-Time (Dst) index measures the strength of geomagnetic storms. It is derived from the horizontal component measurements of Earth's magnetic field at mid-latitude magnetometer stations. The Dst index quantifies the disturbance in Earth's magnetic field during geomagnetic storms, with negative values indicating magnetic storms and positive values indicating quieter conditions. Researchers and space weather forecasters commonly use the Dst index to monitor and assess the impact of solar activity on Earth's magnetosphere and ionosphere.

The objectives of this analysis are to explore the OMNI2 dataset, preprocess the data, and apply linear regression, principal component analysis (PCA), and classification models to investigate the relationships between solar wind parameters and geomagnetic disturbances. By identifying the key factors influencing the Dst index, we aim to enhance our understanding of solar wind-magnetosphere interactions.

2 Exploration of OMNI2 Data

2.1 Data Format and Missing/Fill Values

The OMNI2 dataset consists of 289,296 records and 7 variables, including 'time', 'Bx', 'By', 'Bz', 'density', 'V', and 'Dst', that were made 01/01/1988 to 31/12/2020.

- This variable records the date and time of each measurement each hour.
- Magnetic Field Components (Bx, By, Bz): These components measure the strength and direction of the interplanetary magnetic field in nanoTeslas (nT).
- Plasma Density (density): Measured in particles per cubic centimeter (particles/cm^3), this variable indicates the density of the solar wind plasma.
- Solar Wind Speed (V): This variable, measured in kilometers per second (km/s), shows the velocity of solar wind particles
- Geomagnetic Index (Dst): Given in nanoTeslas (nT), the Dst index measures the intensity of geomagnetic storms, with negative values indicating stronger storms

The dataset contains no missing values, but some variables have fill values as seen on Figure 1, which are given in Table 1. As it can be observed from Figure 1 the data set contains fill values mainly in the data collected before 1995. Moreover, in each year from 1988 to 1995 the fill values are from approximately 55% up to 70% of the yearly collected data.

Table 1: Table of the default fill values

Quantity	Description	Units	Default fill value
Bx, By, Bz	Components of magnetic field	nT	999.9
density	Plasma density	N/cm3	999.9
V	Plasma bulk velocity	Km/s	9999
Dst	Geomagnetic index	nT	99999

Such unequal data distribution can lead to inaccuracies in calculations and analysis, including a distortion of the data distribution. In addition, the standard values for filling values are slightly higher than the maximum measured value for the corresponding features (Figure 2). For example, the maximum measured value for density is around 55 N/cm3, while the filling value for this quantity is 9999 N/cm3. This can lead to an expectation that if the obtained data is expected to follow a normal distribution, then such an imbalance caused by filling values will disrupt the expected data distribution. Also, there are not any fill values for "Dst" feature, therefore it is not presented in Figure 1.

It was decided to drop the records with fill values, resulting in a cleaned dataset with 246,486 records. A significant reduction in fill values is observed after 1995. Post-1995, the percentage of fill values drops sharply to near zero, with occasional increases in density fill values around the early 2000s. This trend might indicate improvements in data collection and processing over time.

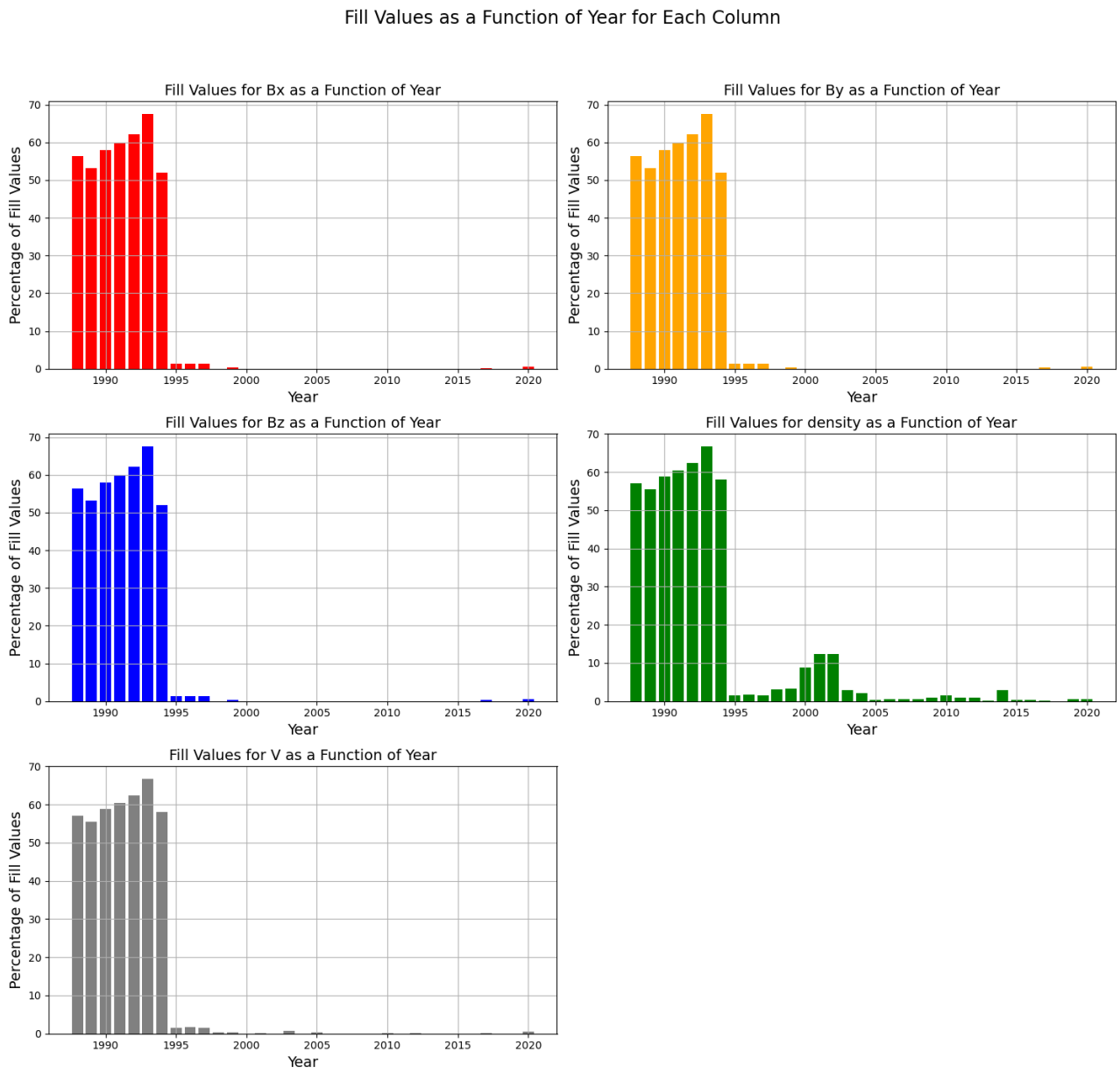

Figure 1: Percentage of Fill Values by Year: The plot illustrates the percentage of fill values for various solar wind parameters (Bx, By, Bz, density, V) in the OMNI2 dataset from 1988 to 2020.

2.2 Data Distribution and Main Metrics

Figure 2 presents the numerical quantities recorded in the dataset, after cleaning the fill values. Also, the table of the main numerical metrics (Table 2) was produced for the cleaned dataset. The appropriate number of bins for each histogram was obtained with Sturge's rule, given by $bin\,number = 1 + ceil(log_2(n))$, where n is a number of observations.

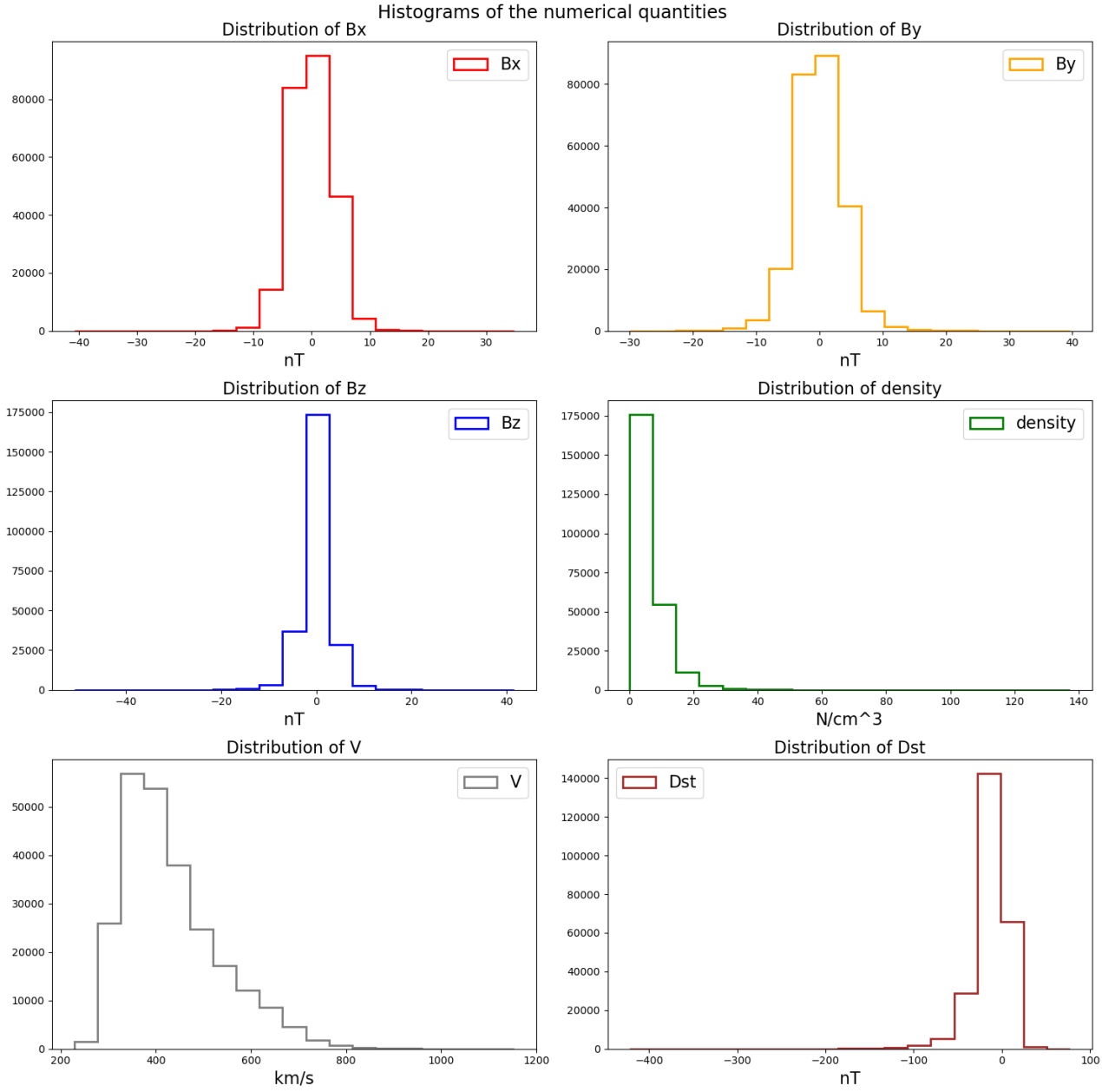

Figure 2: Histograms of distribution of magnetic field components (Bx, By, Bz), solar wind density, velocity (V), and the Disturbance Magnetic index (Dst).

Table 2: Main Metrics of the Clean Dataset

Metric	Bx	By	Bz	density	V	Dst
Mean	0.008	0.003	-0.034	6.442	431.556	-12.990
Median	0.000	0.000	0.000	5.000	408.000	-9.000
STD	3.523	3.829	3.015	5.191	102.112	20.274
IQR	5.100	4.800	2.900	4.800	130.000	20.000

Figure 2 shows that in general, the distribution of numerical quantities from the dataset is close to a normal distribution, especially for magnetic field components (Bx, By, Bz). Distributions for solar wind density and velocity have right tails, while for the magnetic index distribution, there is a short left tail. However, the observed tails do not represent significant deviations, suggesting that all the presented distributions are normal, which will be utilized further in applying the Gaussian Naive Bayes classification. Notably (from Table 2), the solar wind velocity (V) has a high standard deviation, suggesting significant variability, which is typical of dynamic solar wind conditions.

Also, the Pearson Correlation Matrix was obtained and presented as a heat map (Figure 3) below with a 2-dimensional pair plot for each pair of numerical features.

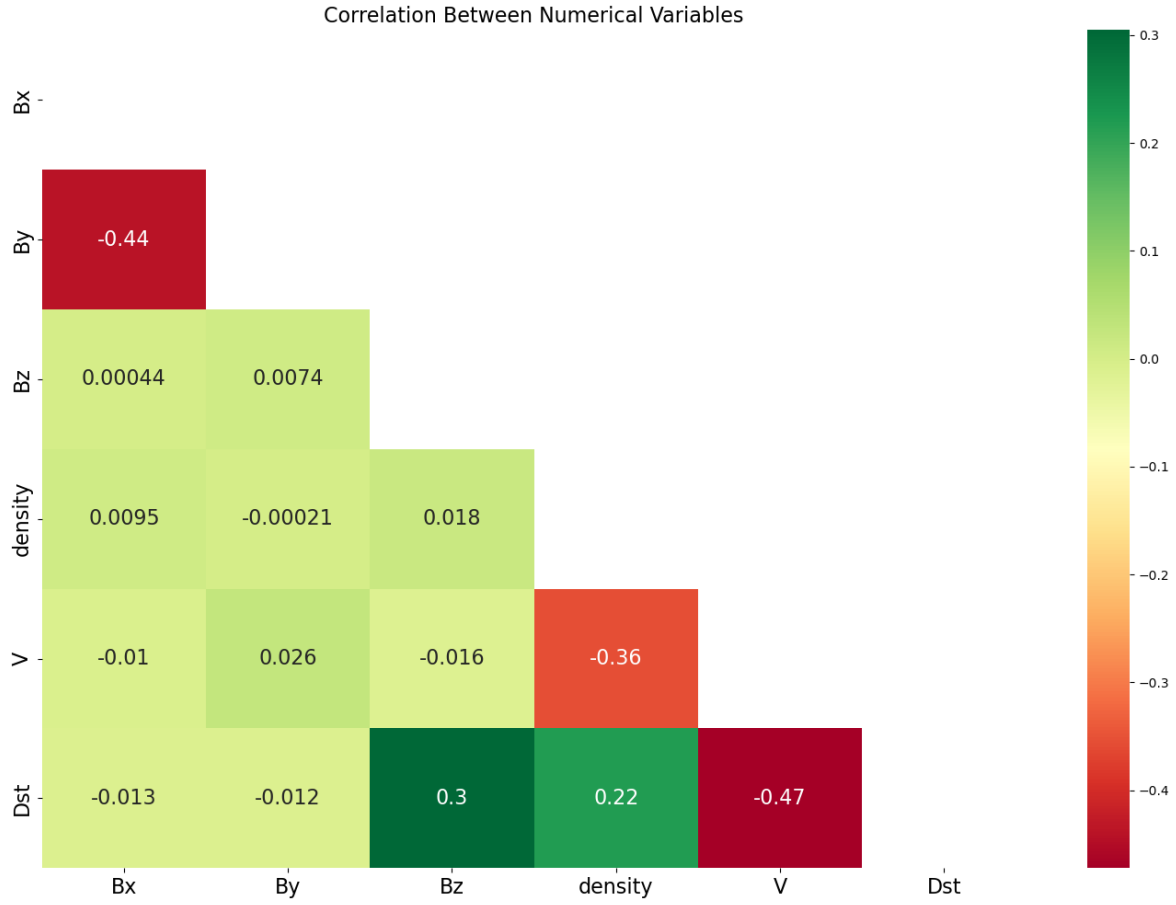

Figure 3: Heatmap of Correlation Coefficients: This heatmap displays the correlation coefficients between solar wind parameters, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

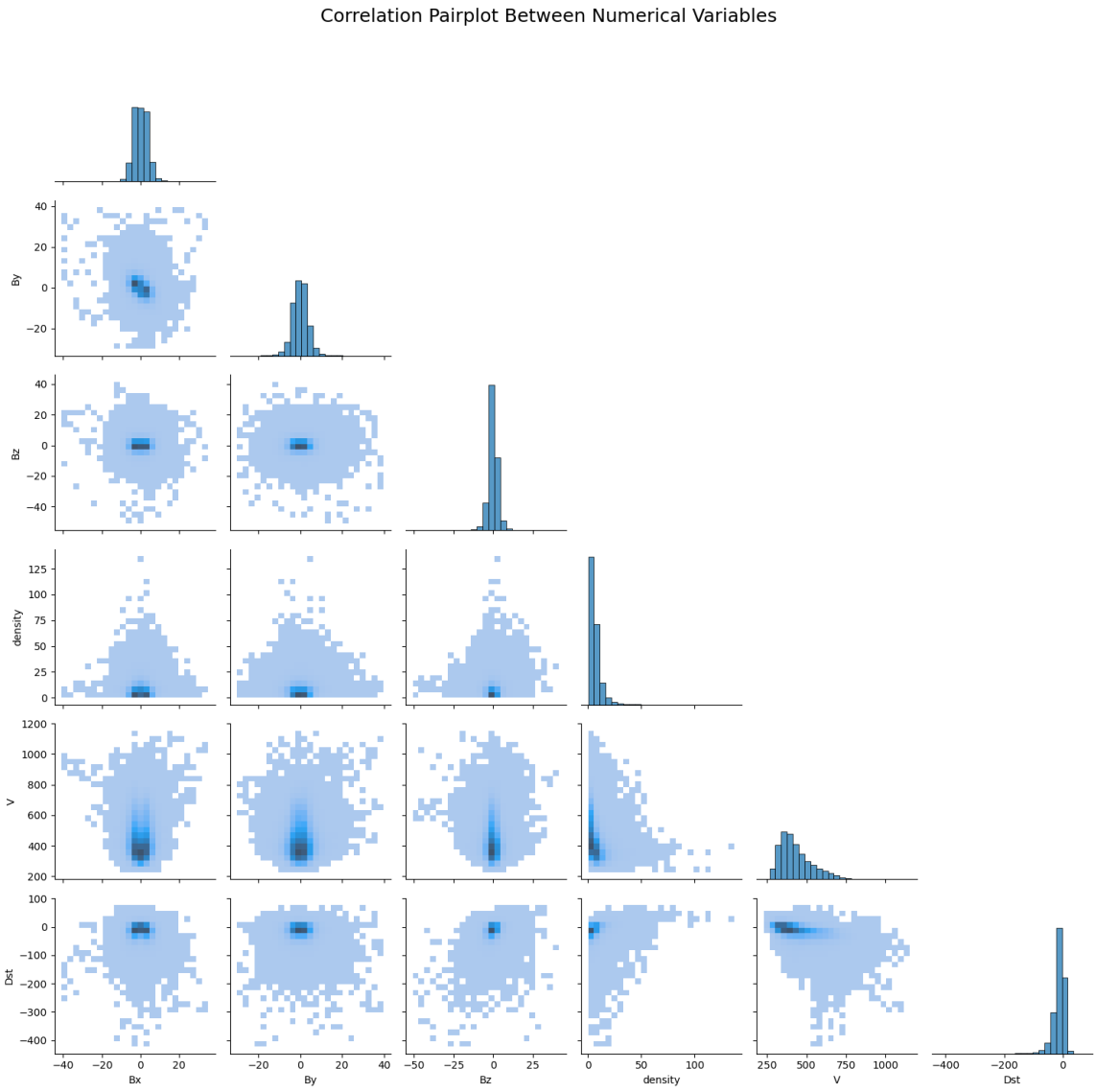

Figure 4: Scatterplot Plot of Solar Wind Parameters. In general we can see, that all pairs do not demonstrate a linear relationship

From Figure 3 we can observe, that in general, the pairs of numerical features do not demonstrate a very strong correlation. The strongest positive correlation from Dst is given by the Dst-Bz pair (+0.3) and the strongest negative correlation is given by the Dst-V pair (-0.47). As will be observed later, the Z-component of the magnetic field and its speed have the most influence on predicting Dst values with Linear regression techniques.

The existence of correlation between the variables means that a model could potentially be constructed to predict future geomagnetic storms or identify their source. However, Figure 4 demonstrates, that relationships for all of the feature pairs are far away from linear, therefore it may lead to low performance of the Linear Regression Models, which will be proved later.

2.3 Autocorrelation Analysis

An examination of autocorrelation was conducted (refer to Figure 5), indicating a prominent peak at approximately 650 hours. This peak might be associated with the solar rotation cycle, which lasts around 27 days (equivalent to 648 hours), indicating a repeating pattern in solar wind characteristics driven by the Sun's rotation [2].

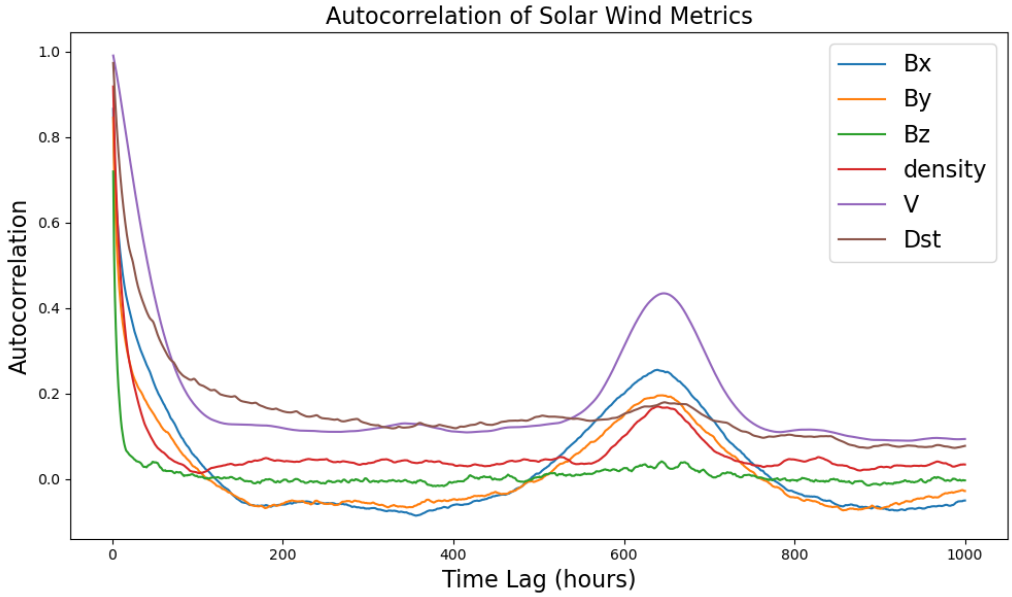

Figure 5: This plot displays the autocorrelation of solar wind features (Bx, By, Bz, density, V) and the geomagnetic Dst index as a function of time lag, measured in hours.

Figure 5 shows that Bx and By demonstrate significant autocorrelation at various time lags, suggesting robust periodic behavior. This periodicity is attributed to the formation of large-scale solar wind structures like coronal holes and streamers, which are influenced by the Sun's rotation. As these structures rotate with the Sun, they induce periodic fluctuations in solar wind parameters, which leads to increased values of autocorrelation. However, Bz demonstrates lower autocorrelation, suggesting that its fluctuations are either more stochastic or influenced by other factors compared to Bx and By. So, it requires deeper research, which is not the aim of this report. Density and velocity (V) also show pronounced peaks around 650 h, which align with the solar rotation period [1].

In general, the variance in autocorrelation values among different features and the lower autocorrelation in Bz and Dst underscores the complexity of forecasting geomagnetic storms, which will proved below with Linear Regression modeling. This suggests the importance of integrating additional information, such as solar wind-magnetosphere coupling and the state of the Earth's magnetosphere, to improve space weather prediction models.

The Dst index, a measure of geomagnetic activity, does not demonstrate a high autocorrelation, indicating that it is influenced by more complex and less predictable factors, such as the interaction between the solar wind and the Earth's magnetosphere. The lower autocorrelation of Dst means that it may not be sufficient to rely only on past Dst values to predict future geomagnetic storms, since these events often arise as a result of sudden changes in solar wind conditions.

3 Modeling of the OMNI2 Data

This part of the work involves separating the numerical features from the dataset into target values, which is the Geomagnetic "Dst" index, and a list of predictors [Bx, By, Bz, density, V, log density]. The log of the magnetic field density was added to the data set to be used as an additional predictor. The dataset was split into a training sample, which includes all observations made before 2020, and a test sample, with observations made after 2020. The ratio for the test/train samples is presented below. After, the backward elimination method was applied to the training set to rank the features in order of importance.

- **Size of the training sample:** 237758.
- **size of the test sample:** 8728.
- **the test/train ration:** 3.67%.

3.1 Multiple Linear Regression

The OLS Regression model was used to predict the value of the Dst index using the solar wind parameters as predictors. The regression coefficients and main statistics were obtained, indicating the significance of each predictor (Table 3). The histogram of the residuals and a plot of the absolute value of the residuals were obtained (Figure 6) as well as a plot of the partial residuals for the test sample (Figure 7).

Table 3: Summary of Regression Analysis

Variable	Coefficient	Std. Error	t-value	P-value
Intercept	24.646	0.240	102.501	<0.001
Bx	-0.137	0.011	-12.510	<0.001
By	-0.061	0.010	-6.072	<0.001
Bz	2.013	0.011	175.894	<0.001
Density	0.272	0.014	20.009	<0.001
Log Density	-1.175	0.249	-4.716	<0.001
V	-0.089	0.000	-231.877	<0.001

Model Fit Statistic	Value
Test R^2	0.177
Test RMSE	10.091
Test MAE	7.655
Training R^2	0.313
Training RMSE	16.981
Training MAE	11.162
Prob (F-Statistic)	<0.001

Based on the data presented in Figure 6, we can conclude that the obtained model has low accuracy, as the values on the Absolute Residuals against the Predicted Values plot are scattered far from zero on the y-axis. Additionally, the distribution plot of residuals indicates relatively large obtained values, where a multitude of values is concentrated.

Additionally, based on the data presented in the table of key metrics for the obtained model (Table 3), we can confirm our conclusion that this model is not suitable for forecasting Geomagnetic "Dst" index values based on the given numerical features. The obtained R-Squared values for the Training and Test samples are 31.3% and 17.7%, respectively, which is quite low.

The analysis of the Partial Residuals Plot (Figure 7) confirms our assumption that the presented numerical features from the dataset (Bx, By, Bz, density, V, and log density) are not suitable for this model. We can observe that the partial residuals values are significantly distant from the regression line. Among all the presented variables, the closest values to the regression line were obtained for Solar Wind Speed (V) (bottom right plot). Later, in the results of backward elimination, it will be demonstrated that this parameter has one of the most significant influences on the Geomagnetic index.

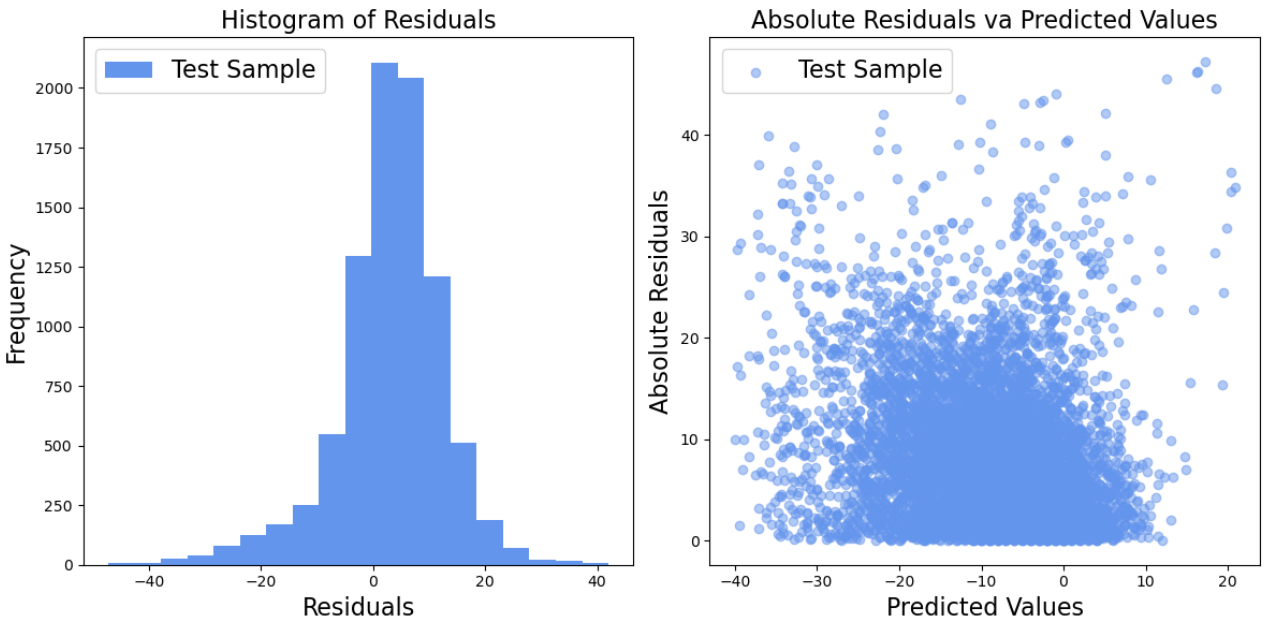

Figure 6: Plotof the residuals: This plot shows the histogram of the distribution of the residuals from the obtained model (on the left). and Absolute residuals against Predicted values (on the right)

3.2 Importance of the Numerical Variables

The backward elimination method on the training set to rank the features in order of importance. As may be observed from Table 3 the obtained main regression metrics fro all predictors (features) have p-values around zero. Therefore we need to use other metrics such as t-value and std error to choose the predictor that will be dropped. So, it was decided to remove log density, By, Bx, and density one by one as those predictors have relatively large std error and small t-values. The results of the obtained model are presented in Table 4.

Table 4: Model Obtained by Backward Elimination

Variable	Coefficient	Std. Error	t-value	P-value
Intercept	27.0159	0.152	177.931	<0.001
Bz	2.0171	0.011	175.939	<0.001
V	-0.0928	0.000	-237.192	<0.001

Model Fit Statistic	Value
Training R^2	0.309
F-statistic	5.327e+04
Prob (F-Statistic)	<0.001

Hence, the latest model with only two predictors (Bz and V) exhibits nearly as strong a fit (R2: 0.309) as the original model with six predictors (R2: 0.313) based on the training sample. Nonetheless, both models display subpar performance, indicated by their relatively low R-squared values (around 30

Threfore, based on the data presented in Table 4, we can conclude that the speed of solar wind and the Bz magnetic field component has the greatest influence on the Geomagnetic "Dst" index, while the remaining variables (Bx, By, density, log density) have the least impact, consistent with previously obtained correlation dependencies.

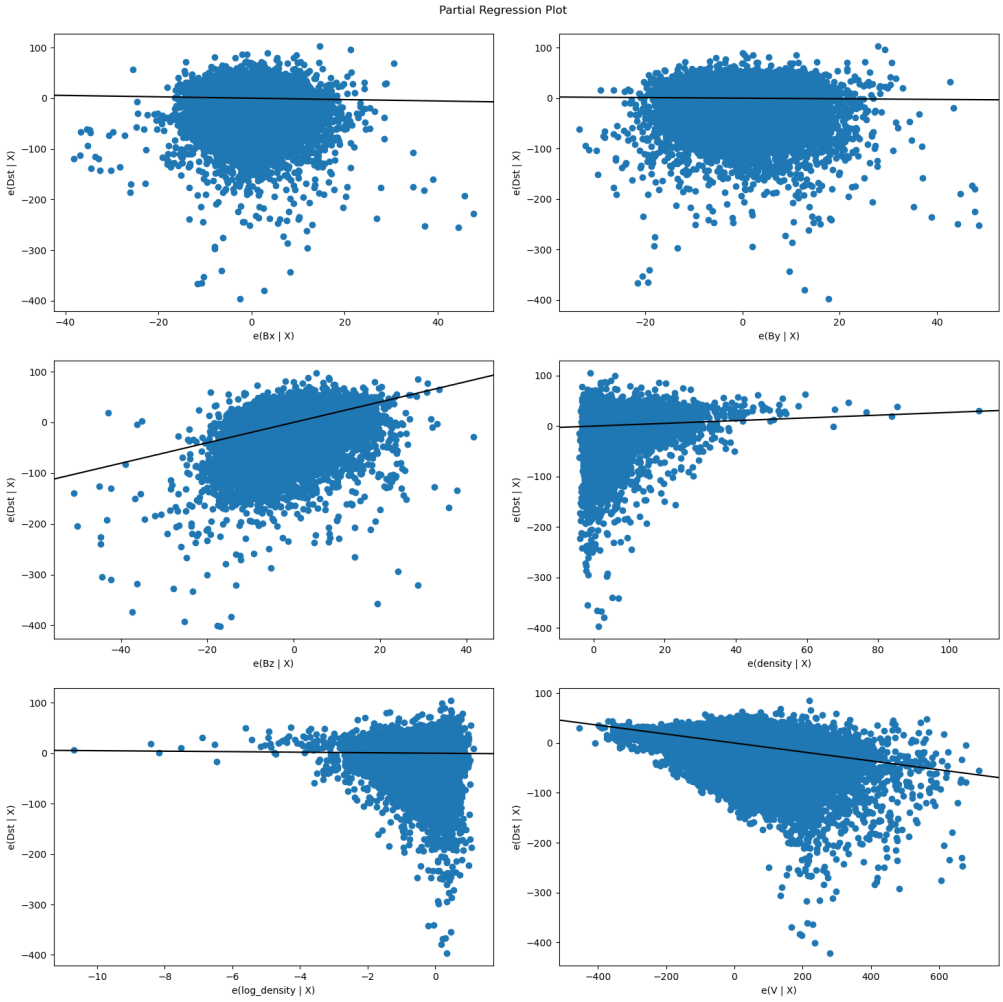

Figure 7: Partial Residual Plots for Solar Wind Parameters: These plots depict the association between each predictor variable (Bx, By, Bz, density, log density, and V) and the Dst index, adjusting for the impact of other predictors in the linear regression model. The black lines represent the trend lines, illustrating the linear influence of each predictor on the Dst index.

3.3 Influence Plot

The influence plot (Figure 8) visually illustrates the effect of each observation on the regression model. It displays standardized residuals against predicted values, with the size and color of each point indicating Cook's Distance. Points with higher Cook's Distance exert more influence on the model. The annotated points correspond to notable geomagnetic storms, as corroborated by sources such as SpaceWeatherLive, which documents the largest geomagnetic storms of the past 50 years. This correlation underscores the significance of these events in the dataset and their influence on the regression analysis.

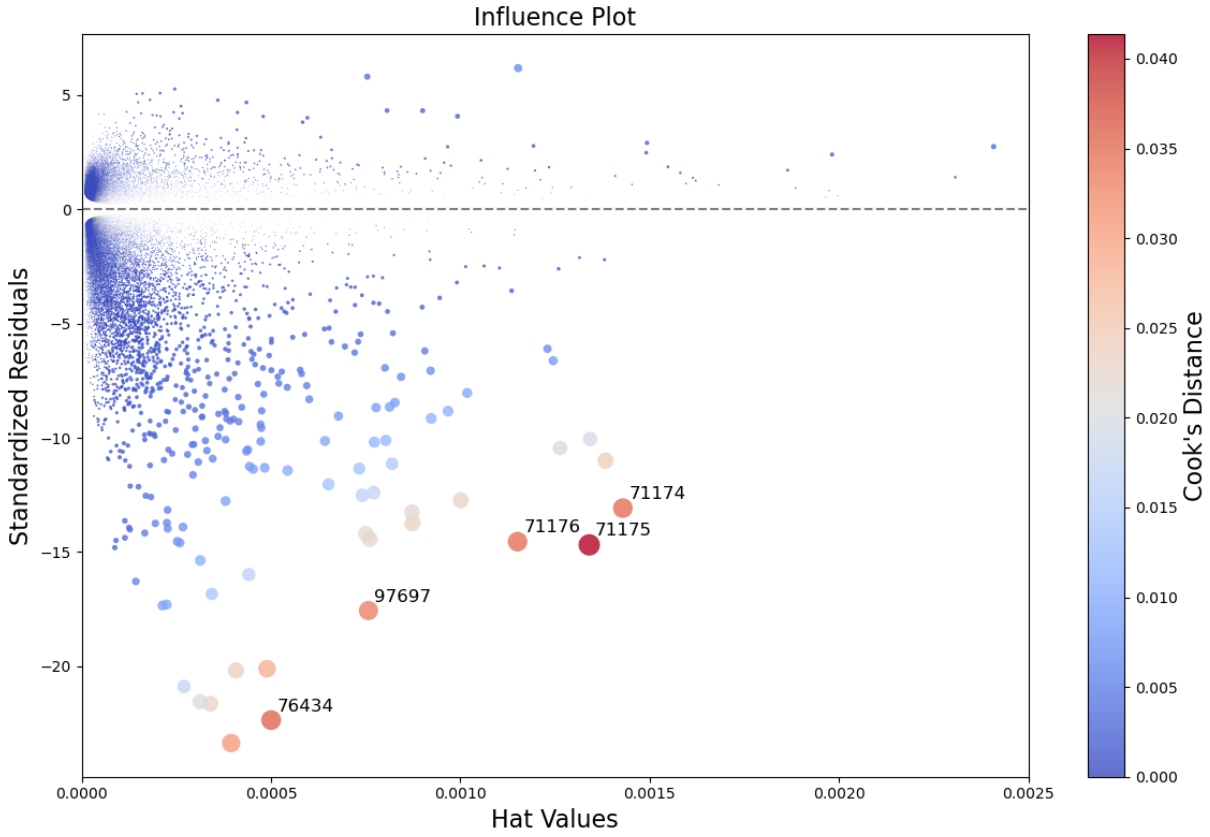

Figure 8: This plot displays the standardized residuals against predicted values for each observation in the dataset. The size and color of the points correspond to Cook's Distance, indicating the influence of each observation on the regression model. Larger points indicate more influential observations. To prevent label overlap, only the most prominent point for any given day is shown.

In the figure, we can observe that the five largest bubbles are indexed. These indices correspond to the row indices in the dataset. Thus, obtaining these indices allowed us to extract the corresponding rows from the dataset. It was found that these rows are dated with information collected on 16/07/2000, 31/03/2001, and 20/11/2003. These dates correspond to some of the most intense geomagnetic storms recorded on Earth [3]:

- Geomagnetic storm in June 14-16, 2001; (71174, 71175, 71176)
- Halloween storm (October-November 2003); (97697)
- Geomagnetic Storm in March-April 2001; (76434)

The data collected during these storm periods have the greatest impact in the given dataset.

3.4 Principal Component Analysis (PCA)

PCA was utilised on the predictors to unveil the underlying patterns and relationships. The scree plot (Figure 9) was obtained to assess the relative significance of principal components and the plot of the weights of each component (Figure 10) was produced to provide insights into the contribution of each variable to the principal components.

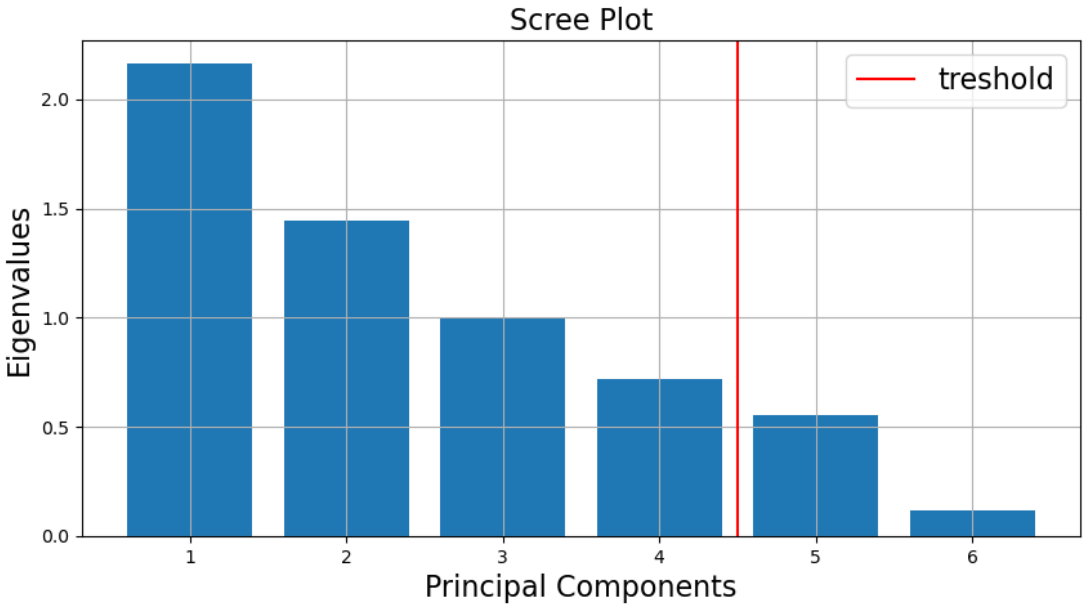

Figure 9: This chart illustrates the eigenvalues associated with each principal component in the PCA. The plot demonstrates the diminishing importance of each component, with the initial components elucidating the majority of the variance in the dataset.

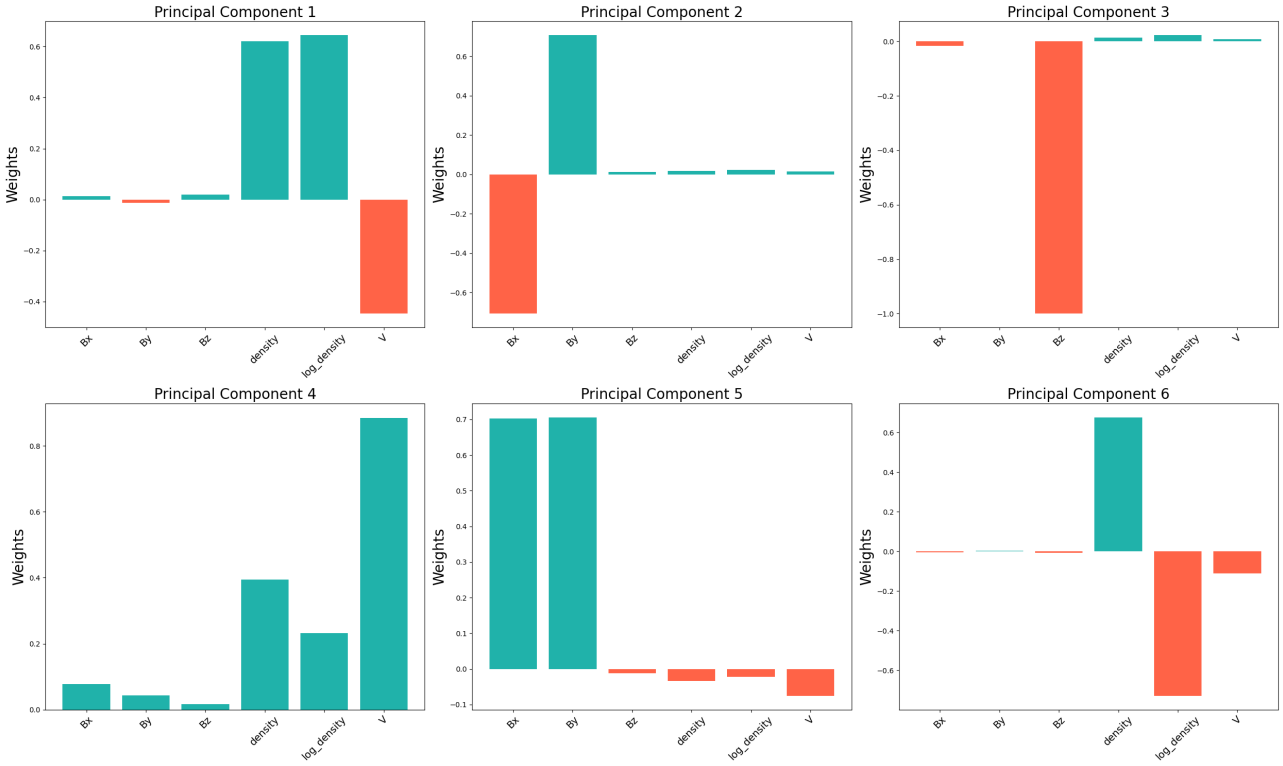

Figure 10: This bar chart illustrates the weights of each feature for the first six principal components obtained from the PCA. The loadings represent the contribution of each original feature (Bx, By, Bz, density, log_density, V) to the principal components (PC1 to PC6). High absolute values of the loadings indicate that the corresponding features exert a strong influence on the respective principal component.

In Figure 9 we can see the cumulative explained variance plot illustrates that the first five principal components explain approximately 90% of the total variance. This suggests that these five components effectively represent the data's underlying structure, reducing dimensionality while retaining most of the information.

Interpreting the loadings highlights the importance of Bz and V in the first principal component, capturing magnetic field intensity and solar wind velocity aspects. The second and third components capture directional magnetic field components and density variations, with Bz showing a strong negative loading in the second component, indicating its significance in explaining north-south magnetic field variance.

Additionally, density and log_density notably contribute to the third and fourth components, reflecting their impact on solar wind density variations. Velocity (V) displays significant loadings across multiple components, emphasizing its role in capturing solar wind speed dynamics.

However, analysis of the linear regression model suggests non-linear patterns in the data, indicating the need for non-linear models to more accurately represent and predict geomagnetic activity.

4 Classification

In order to explore the non-linear nature of the data, the following models were deployed: Gaussian Naive Bayes, Linear Discrimination, and Logistic Regression. The "Dst" variable, which measured the strength of geomagnetic storms, was utilized to classify geomagnetic storms. A categorical variable, "Dst label," was added to the dataset, where it was True when the Geomagnetic "Dst" index had values below -20 nT and False when the Geomagnetic "Dst" index had values equal to or above -20 nT.

To complete the classification analysis, the dataset was balanced due to a significant imbalance in favour of Dst_label = False values, indicating more days without geomagnetic storms than with storms, aligning with reality. To perform the balancing, the dataset will be divided into two subsets: the first will contain data with Dst_label = True, and the second with Dst_label = False. Then, after comparing the sizes of the subsets, one of them will be undersampled or upsampled to match the size of the other. Afterwards, both subsets will be combined back into a single dataset.

4.1 Modelling Results

Below is a comparative table with the obtained data for each classification model (Table 5), as well as a graph of the ROC curves with the corresponding AUC values (Figure 11).

Based on the data presented in Figure 11 and Table 5, it can be observed that the LDA and Logistic Regression models have a reasonably good accuracy of 79%, along with a relatively good AUC value of 0.766. However, the Gaussian Naive Bayes Model demonstrates the highest accuracy at around 81% with a corresponding AUC value of 0.775. This suggests that the Gaussian Naive Bayes is the most suitable classification model for analyzing this dataset, further confirming our earlier approximation that the distribution of numerical features from the dataset can be considered normal.

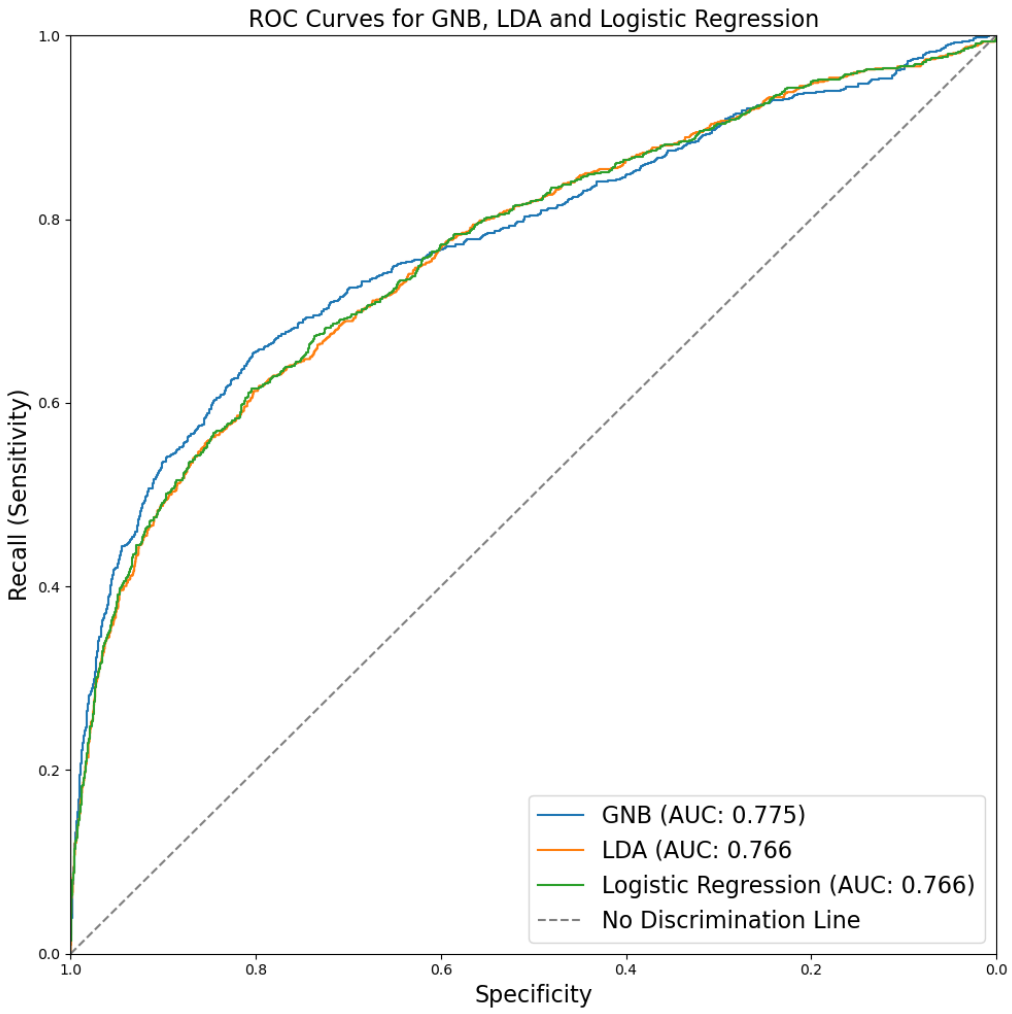

Figure 11: The plots show the Receiver Operating Characteristic (ROC) curves for various classification models trained on a balanced dataset and the obtained Area Under the Curve (AUC) parameter.

Table 5: Performance Metrics for Various Models on Balanced Data (Post-2020 Test Split)

Model	**Class**	**Precision**	**Recall**	**F1-score**	**Support**	**ROC-AUC Score**
Naive Bayes	False	0.84	0.93	0.88	2606	0.7631
	True	0.69	0.48	0.56	894	
	Accuracy		0.81 (3500 samples)			
	Macro avg	0.76	0.70	0.72	3500	
	Weighted avg	0.80	0.81	0.80	3500	
LDA	False	0.84	0.89	0.86	2606	0.7552
	True	0.61	0.50	0.55	894	
	Accuracy		0.79 (3500 samples)			
	Macro avg	0.73	0.70	0.71	3500	
	Weighted avg	0.78	0.79	0.78	3500	
Logistic Regression	False	0.84	0.88	0.86	2606	0.7569
	True	0.59	0.52	0.55	894	
	Accuracy		0.79 (3500 samples)			
	Macro avg	0.72	0.70	0.71	3500	
	Weighted avg	0.78	0.79	0.78	3500	

5 Conclusion

This report offers a thorough examination of the NASA OMNI2 dataset, concentrating on the connections between solar wind parameters and geomagnetic disturbances. Initial data exploration unveiled the presence of fill values, notably before 1995, which were subsequently eliminated for analysis. Investigations into the distributions of magnetic field components (Bx, By, Bz), solar wind density, velocity (V), and the Disturbance Storm Time (Dst) index emphasized their variability and occasional extreme occurrences.

Correlation analyses revealed a strong negative correlation between solar wind velocity and the Dst index, indicating that higher velocities correlate with stronger geomagnetic storms. The Bz component displayed a moderate positive correlation with Dst, underscoring its significance in geomagnetic activity. Autocorrelation analysis unveiled a periodic pattern in solar wind features, likely linked to the solar rotation cycle.

The linear regression model underscored the substantial impact of the Bz and V variables on the Dst index, aligning with theoretical expectations. Diagnostic plots, such as influence plots and partial residual plots, provided insights into the model's performance and the relationships between predictors and the response variable. PCA results further supported the significance of Bz and V in elucidating variability in the dataset.

Various machine learning models, including Gaussian Naive Bayes, Linear Discriminant Analysis, and Logistic Regression were applied to the data. The Gaussian Naive Bayes achieved the highest performance with an AUC of 0.775 on the balanced dataset.

Future work could entail integrating additional geomagnetic indices, exploring the temporal evolution of solar wind features, and implementing more advanced techniques to address imbalances and noise in the dataset. Enhancing predictive models and deepening our understanding of the underlying physical processes will bolster our capacity to forecast space weather events and mitigate their impacts.

Overall, this analysis of the OMNI2 dataset offers valuable insights into the relationships between solar wind parameters and geomagnetic disturbances. The results from linear regression and PCA underscore the importance of the Bz and V variables in predicting the Dst index. Machine learning models showcase potential for enhanced geomagnetic storm prediction while also pinpointing areas for further refinement and exploration. These findings set the stage for future research and the development of more precise space weather forecasting tools.

References

[1] Pevtsov, A.A. and Canfield, R.C. (2001). 'Solar magnetic fields and geomagnetic events', Journal of Geophysical Research: Space Physics, 106(A11), pp. 25191–25197. doi: 10.1029/2000ja004018.

[2] Mouradian, Z., Bocchia, R. and Botton, C. (2002). 'Solar activity cycle and rotation of the Corona', Astronomy & Astrophysics, 394(3), pp. 1103–1109. doi: 10.1051/0004-6361:20021244.

[3] Zhang, J. et al. (2007). 'Correction to "solar and interplanetary sources of major geomagnetic storms (Dst $\leq$ -100 NT) during 1996–2005"', Journal of Geophysical Research: Space Physics, 112(A12). doi: 10.1029/2007ja012891.