What Machine Learning Algorithms Teach us about Which Explanatory Variables Matter Most in Predicting Bz within Coronal Mass Ejections

Pete Riley[1], M. A. Reiss[2] and C. Moestl[2]

[1] Predictive Science Inc., San Diego, California, USA.
[2] Space Research Institute, Austrian Academy of Sciences, SchmiedIstrasse 6, 8042 Graz, Austria.

ABSTRACT

Accurately predicting the z-component of the interplanetary magnetic field, particular during the passage of an interplanetary coronal mass ejection (ICME) is a crucial objective for space weather predictions. Currently, only a handful of techniques have been proposed and they remain limited in scope and accuracy. Recently, a robust machine learning (ML) technique was developed for predicting the minimum value of Bz within ICMEs based on a set of 42 'features', that is, variables calculated from measured quantities upstream of the ICME and within its sheath region. In this study, we investigate these so-called explanatory variables in more detail, focusing on those that were (1) statistically significant; and (2) most important. We find that number density and magnetic field strength, as well as the minimum value of IMF By accounted for a large proportion of the variability. Taken together, these features capture the degree to which the ICME compresses the ambient solar wind ahead. Intuitively, this makes sense: Energy made available to CMEs as they erupt is partitioned into magnetic energy density and kinetic energy. Thus, more powerful CMEs are launched with larger flux-rope fields (larger Bz), at greater speeds, resulting in more sheath compression (increased number density, By, and total field strength).

KEY POINTS

- ML algorithms driven by features within the upstream sheath region, can accurately predict minimum values of Bz within the ejecta
- The most important and statistically significant features are sheath number density, total field strength, and IMF By.
- These features capture compression upstream of the ICME, which correlates with the overall magnetic strength of the ejecta.

CONTACT

Dr. Pete Riley
Predictive Science Incorporated
Email: pete@predsci.com
Phone: (858) 450-6494
Website: www.predsci.com

OVERVIEW

In a previous study (*Reiss et al.,* 2021), we applied machine learning techniques to assess whether upstream in-situ ICME sheath region measurements could provide estimates of the resulting (1) minimum value in the Bz component of magnetic field; or (2) maximum value of the total field (B), within the following ICME. We developed a predictive model based on 348 ICMEs that were observed by Wind, STEREO-A, and STEREO-B spacecraft. we found moderately high associations (Pearson Correlation Coefficient, PCC = 0.71 and 0.91, respectively) between the target variable and a set of 42 input (explanatory) variables or features. These explanatory variables were made up of various statistical properties of the magnetic field vector, plasma density, temperature, and velocity vector, specifically: the mean value, standard deviation, minimum and maximum values, the ratio between the maximum and minimum values, and the ratio between the mean value and standard deviation (i.e., the coefficient of variation). Thus, for each of the eight variables there were six statistical measures, resulting in 42 features or input variables that could, in principle, explain the observed variations in either the minimum value of Bz within the ejecta, or the maximum value of the field. Since the objective was to derive a predictive model, we separated the 348 events into a training (4/5) and evaluation (1/5) dataset and applied three models (linear regressor (LR), random forest regressor (RFR), and gradient boosting regressor (GBR)). The high resulting PCC values suggested that this might be a promising forecast tool for estimating the strength of the ICME's magnetic field many hours before its arrival at 1 AU. In important limitation of the study, however, was that we did not: (1) assess which of the 42 input variables were most important, that is, which variables were responsible for most variations in Bz within the ICME? (2) which of the variables were statistically significant? And (3) scientifically, what was the underlying mechanism for the strong association amongst at least some of the input variables and the target variables?

In this study, we aim build on this previous study by addressing these specific questions. Specifically, we seek to identify those variables that are statistically significant, those that contribute most to the variability in the targets and provide an explanation for why. In doing so, we believe that this addresses a serious concern with respect to the application of ML techniques within Heliophysics: What do we learn from their use? While the previous study directly addressed an operational space weather need and, thus, was not primarily concerned with these issues, the present study aims to use ML approaches to better understand why such high correlations are present.

DATA

Reiss et al. (2021) identified 364 ICMEs that produced an upstream sheath region from the HELCATS ICME catalogue. These were observed by Earth-based or STEREO-A/B spacecraft between January 2007 and March 2021, and thus, were all located at approximately 1 AU. The requirement that they drive a sheath ahead (or, technically, displayed a density enhancement), meant that these were traveling faster than the ambient wind within which they were embedded.

From the time-series of these events, six statistical measures (mean, standard deviation (std), minimum (min), maximum (max), ratio of maximum to minimum values (minmax), and coefficient of variation (cv), i.e., the ratio of the mean to standard deviation) were calculated for each of the following variables: Total magnetic field (Bt), three components of the field (Bx, By, Bz), bulk plasma speed (vt), number density (np), and proton temperature (Tp). The temporal boundaries used to compute these quantities were from the start of the sheath to the end of the sheath interval.

Table 1: List of parameters used in multiple-regression analysis, together with their basic statistical properties. See text for detailed explanation of each parameter.

Feature	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
mean.bx.	348	−0.292	2.737	−8.388	−2.093	1.383	11.411
max.bx.	348	7.114	4.736	−4.941	3.838	9.357	32.120
std.bx.	348	3.239	1.914	0.000	1.946	4.206	11.714
min.bx.	348	−7.522	4.609	−27.209	−10.101	−4.387	3.577
cv.bx.	348	12.911	130.146	0.000	0.843	3.735	2,416.322
minmax.bx.	348	−1.745	15.220	−265.063	−1.506	−0.692	78.281
mean.by.	348	0.472	4.909	−21.306	−2.847	3.688	16.800
max.by.	348	10.184	6.908	−10.915	6.109	13.074	43.805
std.by.	348	4.591	3.270	0.000	2.536	5.908	25.036
min.by.	348	−9.401	7.095	−50.495	−12.904	−4.934	5.030
cv.by.	348	3.519	8.915	0.000	0.567	2.633	107.937
minmax.by.	348	−1.049	9.981	−138.319	−1.277	−0.535	120.411
mean.bz.	348	0.229	3.109	−9.976	−1.591	1.966	13.673
max.bz.	348	10.256	6.910	−5.495	5.821	13.033	60.193
std.bz.	348	4.558	2.968	0.000	2.667	5.790	20.432
min.bz.	348	−9.799	6.736	−40.842	−12.236	−5.743	7.602
cv.bz.	348	19.078	213.734	0.000	1.251	5.463	3,962.469
minmax.bz.	348	−0.957	2.311	−22.105	−1.209	−0.710	32.596
mean.bt.	348	9.399	4.546	2.026	6.417	11.183	28.042
max.bt.	348	14.362	7.992	2.681	8.777	18.131	61.447
std.bt.	348	2.253	1.741	0.000	1.122	2.901	14.293
min.bt.	348	3.076	2.274	0.228	1.535	4.115	16.107
cv.bt.	348	0.230	0.105	0.000	0.155	0.284	0.734
minmax.bt.	348	0.255	0.188	0.014	0.105	0.355	1.000
mean.vt.	348	447.167	106.012	283.544	369.984	494.505	823.933
max.vt.	348	498.517	131.658	300.348	401.584	567.684	1,001.594
std.vt.	348	19.926	15.332	0.000	8.860	27.290	119.767
min.vt.	348	400.165	88.781	239.981	337.344	439.992	723.750
cv.vt.	348	0.043	0.028	0.000	0.023	0.055	0.208
minmax.vt.	348	0.816	0.092	0.408	0.756	0.885	1.000
mean.np.	348	14.102	9.264	0.451	7.685	18.491	84.160
max.np.	348	28.930	20.587	1.164	14.502	37.137	147.326
std.np.	348	4.704	4.021	0.000	1.970	6.105	26.247
min.np.	348	5.812	4.904	0.112	2.728	7.271	49.334
cv.np.	348	0.326	0.173	0.000	0.201	0.422	1.135
minmax.np.	348	0.248	0.166	0.002	0.127	0.343	1.000
mean.tp.	348	153,878.500	150,688.400	13,725.420	52,988.160	207,266.300	964,935.600
max.tp.	348	560,333.500	751,663.600	19,754.490	117,018.800	647,930.000	5,249,402.000
std.tp.	348	74,633.040	96,655.200	0.000	15,796.020	91,670.020	659,472.700
min.tp.	348	45,772.240	44,346.380	4,565.096	20,026.440	54,835.390	360,332.500
cv.tp.	348	0.414	0.207	0.000	0.270	0.522	1.548
minmax.tp.	348	0.167	0.136	0.004	0.066	0.248	1.000
Target	348	−9.456	6.681	−60.782	−11.760	−5.506	0.000

MODELS

In this study, multiple regression models were analyzed using several packages in R, primarily relying on 'lm'. To identify the explanatory variables that provided the best performing model (i.e., one that lowered the prediction error), we used the 'stepAIC' function from the MASS package. Both forward selection and backward selection (i.e., backward elimination) were performed. In the former, initially, no predictors are included in the model, and the algorithm iteratively adds the most contributory variables, stopping when the improvement is no longer statistically significant.

These regression techniques, however, only identify the statistical significance of the variables. To estimate the relative importance of each explanatory variable in describing the variability in deaths, we applied several traditional statistical and more modern machine learning (ML) approaches. Using multiple approaches is important for assessing the uncertainty that should be ascribed to a particular ordering of the variables, since different techniques rely on different metrics for importance. The techniques applied included: random forest, Xgboost, relative importance, earth, stepwise regression, and DALEX. It is important to underscore that these techniques use different definitions of what signifies “important”, and, thus, we do not expect agreement between the results. Nevertheless, where the results do agree is where we can be most confident that the explanatory variable importance is significant, and where they do not, we must remain more cautious.

MULTIPLE REGRESSION ANALYSIS

Table 2: Summary of the six most significant explanatory variables for predicting the minimum value of Bz within the ICME.

	Dependent variable:
	Target
mean.bt.	0.379*** (0.109)
std.bt.	−0.982*** (0.144)
cv.bt.	0.535*** (0.094)
mean.np.	−0.698*** (0.083)
max.np.	0.322*** (0.087)
max.tp.	−0.107* (0.057)
Constant	−0.000 (0.039)
Observations	348
R^2	0.479
Adjusted R^2	0.470
Residual Std. Error	0.728 (df = 341)
F Statistic	52.231*** (df = 6; 341)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: Stepwise regression using the MASS package.

	Dependent variable:
	Target
std.bx.	0.117* (0.065)
cv.bx.	−0.068* (0.038)
min.by.	0.247*** (0.076)
min.bz.	−0.195** (0.078)
minmax.bz.	0.058 (0.038)
mean.bt.	0.591*** (0.144)
max.bt.	−0.468** (0.198)
std.bt.	−0.777*** (0.170)
cv.bt.	0.596*** (0.098)
mean.vt.	0.452*** (0.172)
std.vt.	−0.436* (0.254)
cv.vt.	0.472** (0.221)
mean.np.	−0.642*** (0.147)
max.np.	0.471*** (0.127)
min.np.	−0.171** (0.081)
cv.np.	−0.154** (0.070)
max.tp.	0.257** (0.115)
std.tp.	−0.335*** (0.123)
min.tp.	−0.144** (0.058)
minmax.tp.	0.143** (0.056)
max.vt.	−0.255 (0.168)
Constant	−0.000 (0.036)
Observations	348
R^2	0.565
Adjusted R^2	0.537
Residual Std. Error	0.680 (df = 326)
F Statistic	20.179*** (df = 21; 326)
Note:	*p<0.1; **p<0.05; ***p<0.01

CORRELATION AMONGST THE MOST SIGNIFICANT EXPANATORY VARIABLES

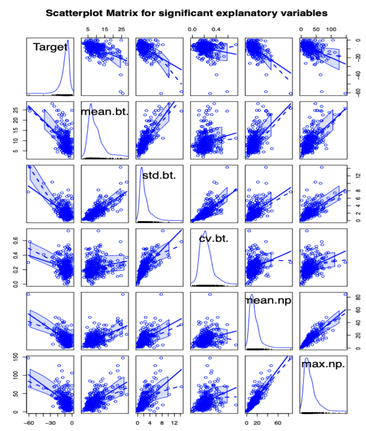


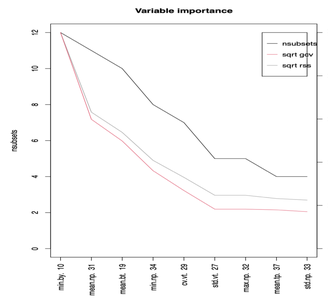
Figure 1: Scatterplot matrix of five significant explanatory variables for the minimum value of Bz within an ICME.} Panels show: Bz within the ICME (the output variable), mean.bt., std.bt., cv.bt., mean.np., and max.np. Data are shown by the circles, regression lines are solid, smoothed mean values are shown by the dashed line, and variances are shown by the dashed-dotted lines.0

MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

Table 4: MARS Method for relative importance of parameters.

Parameter	nsubsets	gcv	rss
min.by.	12	100.0	100.0
mean.np.	11	59.8	63.1
mean.bt.	10	49.8	53.8
min.np.	8	36.1	40.9
cv.vt.	7	26.9	33.0
std.vt.	5	18.3	24.7
max.np.	5	18.3	24.7
mean.tp.	4	17.9	23.2
std.np.	4	17.1	22.5

Figure 2: MARS Method for relative importance of parameters.

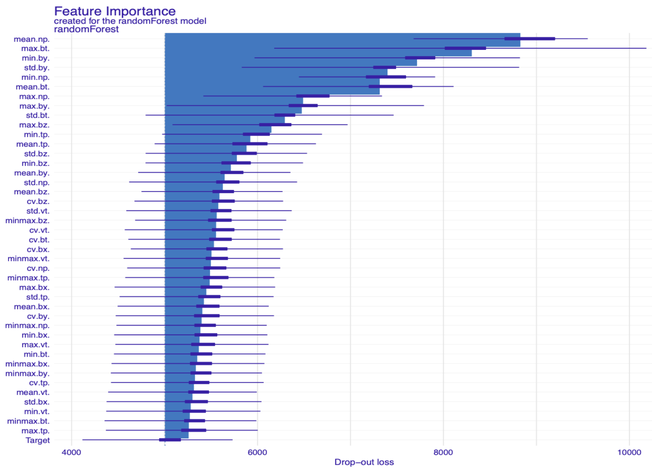


DALEX STEPWISE REGRESSION METHOD

Table 5: DALEX Stepwise Regression Method for relative importance of parameters.

No.	variable	mean dropout loss	label
1	full model	5005.384	randomForest
...	...	...	...
35	max.bz.	6149.186	randomForest
36	std.bt.	6293.088	randomForest
37	max.by.	6476.154	randomForest
38	max.np.	6489.984	randomForest
39	mean.bt.	7314.289	randomForest
40	min.np.	7314.622	randomForest
41	std.by.	7397.856	randomForest
42	min.by.	7715.715	randomForest
43	max.bt.	8304.951	randomForest
44	mean.np.	8827.488	randomForest
45	baseline	69085.770	randomForest

Figure 3: Relative importance of explanatory variables using the DALEX method.



SUMMARY

In this study, we found that statistical measures tracking number density and magnetic field strength accounted for a large proportion of the variability. Taken together, these features capture the degree to which the ICME compresses the ambient solar wind ahead. Intuitively, this makes sense: Energy made available to CMEs as they erupt is partitioned into magnetic energy density and kinetic energy. Thus, more powerful CMEs are launched with larger flux-rope fields (larger Bz), at greater speeds, resulting in more sheath compression (increased number density, By, and total field strength).