SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

SAS®: Cataclysmic Damage

Power Grids and Telecommunications

#SASGF

THE UNIVERSITY OF ALABAMA® | Culverhouse College of Commerce

SAS®: Cataclysmic Damage to Power Grids and Telecommunications

Taylor K. Larkin, Denise J. McManus

The University of Alabama

Abstract

Coronal mass ejections (CMEs) are massive explosions of magnetic field and plasma from the Sun. While responsible for the northern lights, these eruptions can cause geomagnetic storms and cataclysmic damage to Earth's telecommunications systems and power grid infrastructures. Hence, it is imperative to construct highly accurate predictive processes to determine whether an incoming CME will produce devastating effects on Earth. One such process, called "stacked generalization," trains a variety of models, or base-learners, on a data set. Then, using the predictions from the base-learners, another model is trained to learn from the metadata. The goal of this meta-learner is to deduce information about the biases from the base-learners to make more accurate predictions. Studies have shown success in using linear methods, especially within regularization frameworks, at the meta-level to combine the base-level predictions. Here, SAS® Enterprise Miner™ 13.1 is used to reinforce the advantages of regularization via the Least Absolute Shrinkage and Selection Operator (LASSO) on this type of metadata. This work compares the LASSO model selection method to other regression approaches when predicting the occurrence of strong geomagnetic storms caused by CMEs.

Motivation

Typically, Earth's magnetic field is able to guard against the harmful components of a CME. However, when a CME contains a strong southward-directed magnetic field component, energy is transferred from the CME to Earth's magnetic field through a process called magnetic reconnection (Howard, 2011) (animated in figure 1). This compresses the Earth's magnetic field towards the equator, leaving greater proportions of Earth to be exposed. Given that these phenomena can contain 220 billion pounds of solar material expelled with a force equaled to a billion hydrogen bombs ("Coronal Mass Ejections", 2012), the resulting amassed power in the upper atmosphere can lead to over-saturation of power transformers and failures of telecommunications systems (Board, 2008). On September 1, 1859, Richard Carrington and Richard Hodgson observed a solar storm outside of the city of London which disrupted telegraph communications worldwide (Boteler, 2006). Noted as the "Carrington Event," this event precipitated the most powerful geomagnetic disturbance on record. It was estimated that if such an event were to occur in today's society, it would result in a financial impact of tens of billions of US dollars due to damages of commercial satellite structures (Odenwald et al., 2006). On July 23, 2012, Earth narrowly avoided a extraordinarily powerful CME (Bridgman, 2014)(shown in figure 5). If Earth had been in its direct path, the ensuing impact would have been far more detrimental than the Carrington Event of 1859 (Baker et al., 2013).

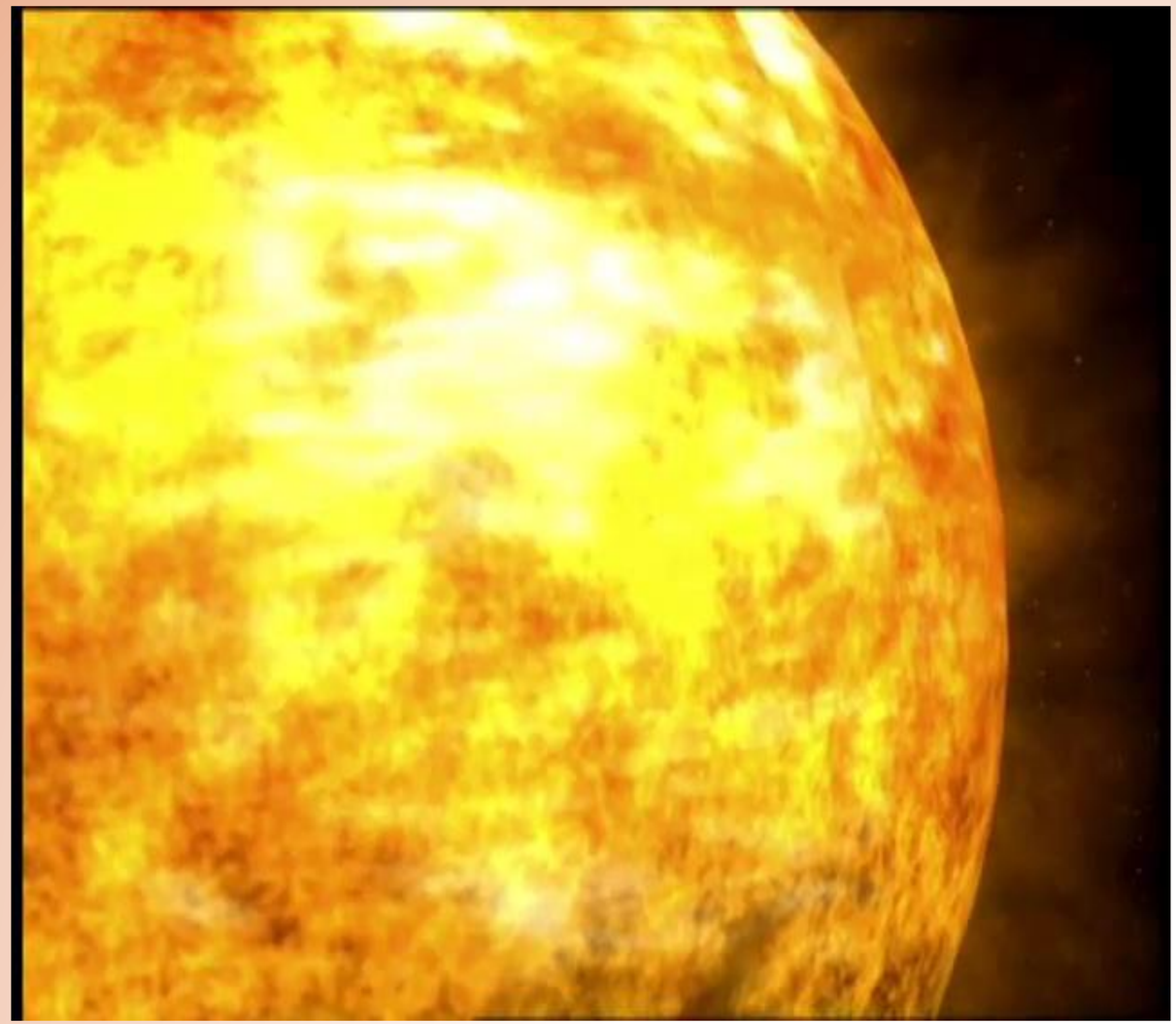

Figure 1: Animation of the magnetic reconnection. Credit listed in acknowledgments section.

Methodology and the Data

Taylor K. Larkin and Denise J. McManus

The University of Alabama

Stacked Generalization and the LASSO

A generalized stacking scheme (Wolpert, 1992) is utilized to construct an accurate framework for predicting the strength of impending CMEs. This process can be simplified into two parts:

- Construct a dataset consisting of class predictions from a set of level 0 (or base) learners using a training and a test set. Typically, this is done by k-fold cross-validation (CV).
- Train a level 1 (or meta) learner that utilizes the predictions made at the previous level as inputs

The level 1 learner's purpose is to gain information about the generalization behavior of each learner trained at the base-level. Popular choices for meta-learners have been linear models such as Ting & Witten (1999). While this ensemble strategy leverages the strengths and weaknesses of the base-learners, it can be prone to over-fitting (Caruana et al., 2004). Therefore, in order to combat this issue, employing regularized linear methods can perform better than their non-regularized counterparts (Reid & Grudic, 2009). One such regularization method called Least Absolute Shrinkage and Selection Operator (LASSO) seeks to correct for the traditional high variance problems of regression by sacrificing bias to greatly reduce variance through use of a penalty constraint (Tibshirani, 1996; Hastie et al. 2009). In other words, instead of the traditional, unconstrained parameterization of the ordinary least squares (OLS) solution,

$$\hat{\beta}^{ols} = \underset{\beta}{argmin} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

the following constraint is imposed on the betas.

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

Increasing the LASSO parameter *t* in discrete steps facilitates a set of variables to be defined at each value of *t* (Cohen, 2006) where a greater number of variables are allowed to enter and exit the model in a continuous fashion. Enabling this penalty allows some of the regression coefficients to be exactly zero, provided *t* is small enough. Thus, it is possible to obtain sparse solutions which encourages the idea of parsimony. Sparsity is especially important when *p* > *N*. Efficient methods have been developed to calculate the entire LASSO coefficient path such as with the least angle regression (LAR) algorithm.

The Datasets

I. Original Data – a set of 18 standardized variables characterizing 182 near-Earth CMEs. This is comprised of interplanetary measurements (Cane & Richardson, 2003) (Richardson & Cane, 2010), initial CME characteristics at the Sun given by the Large Angle and Spectrometric Coronagraph (LASCO) located on the Solar and Heliospheric Observatory (SOHO) satellite (Gopalswamy et al., 2009), and some solar phenomena recorded by the National Oceanic and Atmospheric Administration (NOAA) (Space Weather Prediction Center, September 2015). In addition, the NOAA database is used to create a binary variable signifying whether a CME produced a strong geomagnetic storm. This will serve as the response variable for the predictors.

II. Metadata - a set of 320 class probability predictions on the original data from 20 different models created via 10-fold CV from the **caret** package in R (Kuhn, 2008)(R Core Team, 2015). As with the StackingC approach (Seewald, 2002), only the probabilities of generating a strong geomagnetic storm are used. Each model is trained across 16 different tuning parameters. The models implemented as base-learners are listed in table 1.

Classification and Regression Trees (CART)	C5.0 Decision Trees and Rule-Based Models (C50)	Ripper Rule Learners (JRIP)	Generalized Additive Models using Splines (GAM)
Flexible Discriminant Analysis (FDA)	LASSO and Elastic Net Regularization Linear Models (GLMNET)	Random Forests (RF)	Conditional Inference Random Forests (CIRF)
Stochastic Gradient Boosting (GBM)	Penalized Multinomial Regression (PMR)	Neural Networks (NN)	Partial Least Squares Regression (PLS)
Nearest Shrunken Centroids (NSC)	Support Vector Machines with Radial Basis Function Kernel (SVM)	Neural Networks with Feature Extraction (PCANN)	Sparse Distance Weighted Discrimination (SDWD)
Tree Models from Genetic Algorithms (ET)	Boosted Logistic Regression (BL)	Boosted Classification Trees (ADA)	Rotation Forests (ROTF)

*Table 1: List of base-learner models generated from the **caret** package in R*

Leveraging SAS Enterprise Miner for Predicting Geomagnetic Storms

Taylor K. Larkin and Denise J. McManus

The University of Alabama

Objectives

- Examine whether LASSO will lead to better predictive performance compared to traditional regression methods on the metadata
- Investigate if exploiting the metadata yields better predictions compared to the original data

Experimental Procedure

- Partition the data into a training and a test set 3 times at 3 different percentages (60/40, 70/30, 80/20) with 3 different random seeds as demonstrated in figure 2.

For the metadata

- Using the LARS node, train a regression model using the LASSO variable selection method
- For comparison, also train other regression models with traditional variable selection methods (stepwise, backward, and forward) as well as a full regression model
- Aggregate the average area under the ROC curve (ROC) and misclassification rate across all 3 test sets for each model

For the original data

- Train a diverse set of models (gradient boosting, partial least squares, neural network, memory based reasoning, decision tree, rule induction) along with LASSO and stepwise regression.
- Aggregate the ROC and misclassification rate across all 3 test sets for each model

Notes

- Selecting the best *t* for this model is conducted using 10-fold CV as is usually done for LASSO. That is, CV is conducted for each step in the variable selection process as *t* increases. The step in which the set of regression coefficients delivers the lowest CV error is selected as the stopping point.
- For a fair comparison, model selection criteria for the regression models are set to cross-validation misclassification. For each step in the model selection process (based on p-values), the leave-one-out CV scheme is implemented (Sarma, 2013). The step which yields the lowest misclassification rate on the held out observations is chosen as the best model.
- All non-regression approaches are left at their default settings
- The SAS node is used to compute the mean across the 3 test sets for each model using PROC SQL code for the ROC and misclassification rate.

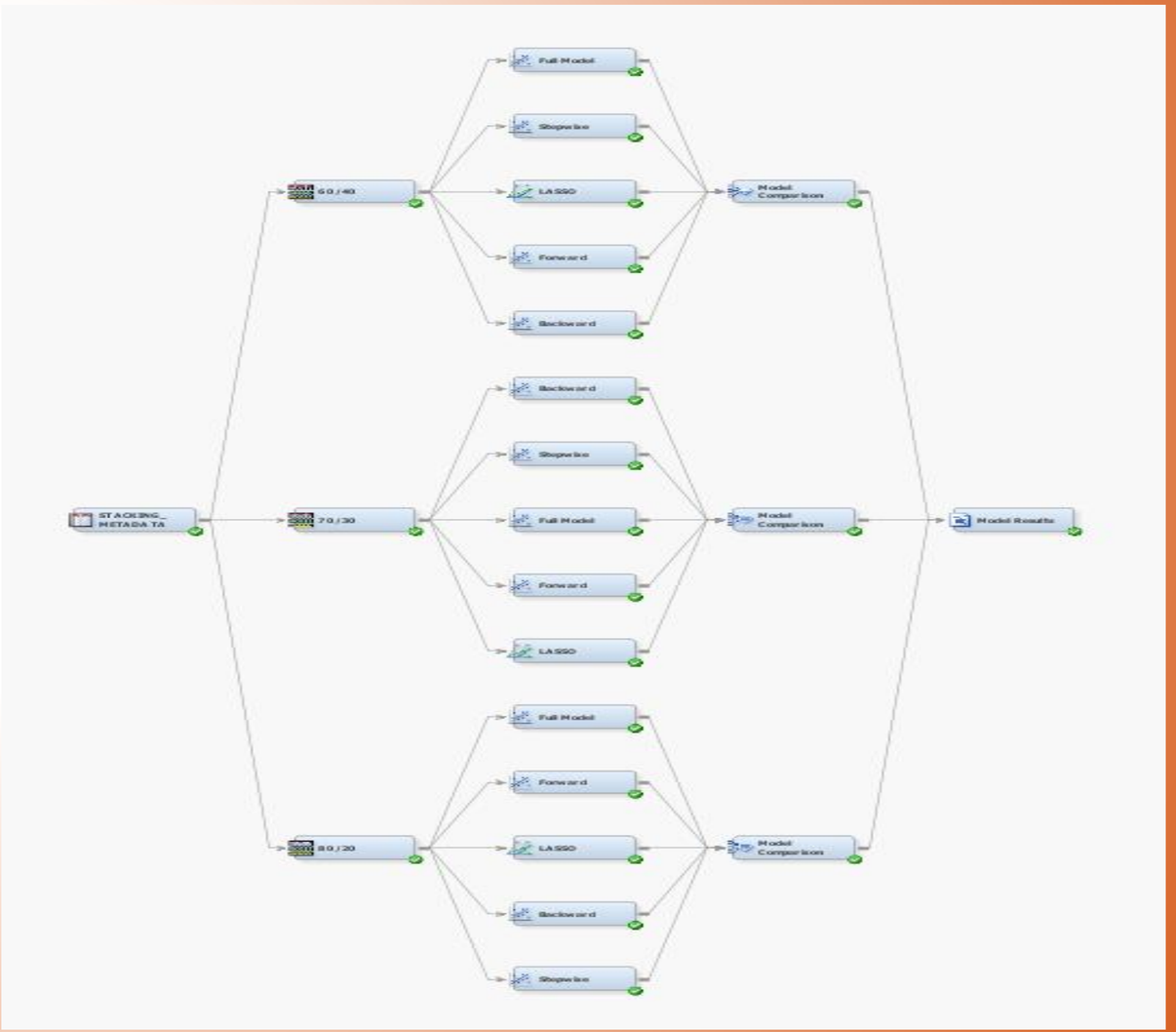

Figure 2: Diagram of metadata modeling in Enterprise Miner

Analyzing the Results

Taylor K. Larkin and Denise J. McManus

The University of Alabama

Visualizations

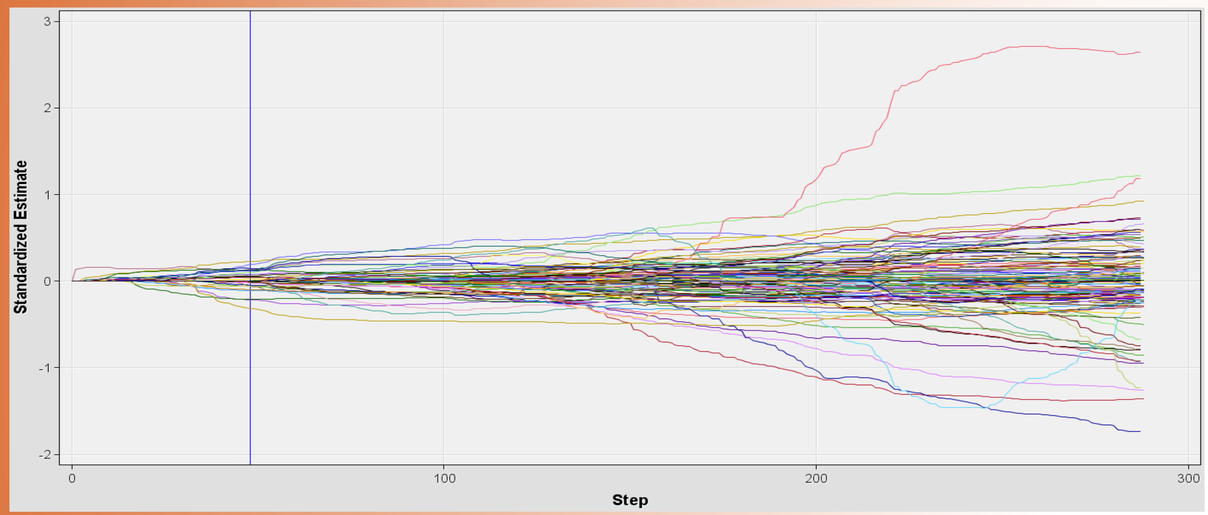

Figure 3: Coefficient paths for 80/20 data partition in metadata modeling

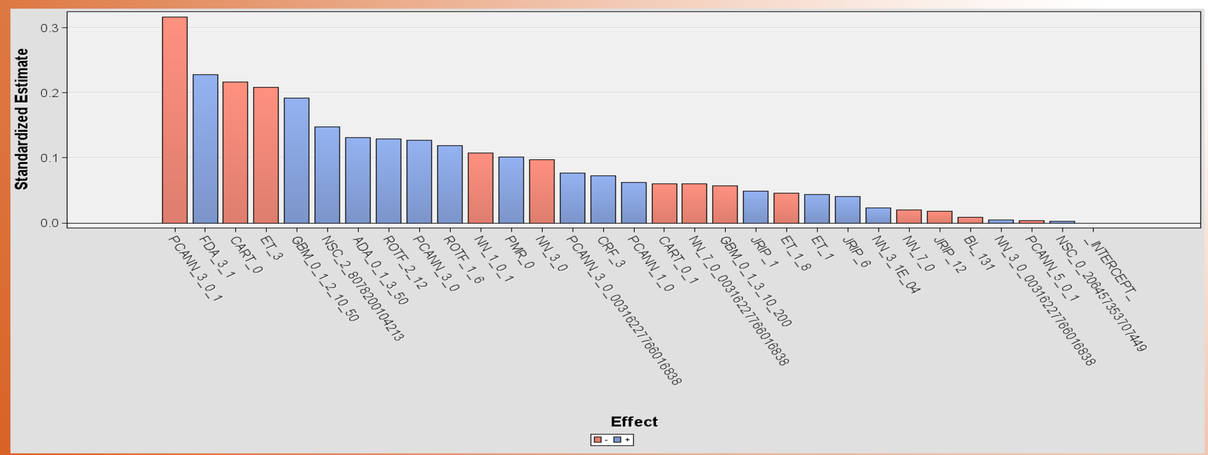

Figure 4: Selected variables for 80/20 data partition in metadata modeling

Performance Metrics

Obs	Model	Average ROC Index	Average Misclassification Rate
1	LASSO	0.86433	0.21108
2	Stepwise	0.81967	0.24798
3	Forward	0.77967	0.20997
4	Backward	0.66267	0.32583
5	Full Model	0.58367	0.42105

Table 2: Averaged results on the 3 test sets from the metadata analysis

Obs	Model	Average ROC Index	Average Misclassification Rate
1	Stepwise	0.84467	0.23605
2	LASSO	0.83333	0.25360
3	Gradient Boosting	0.83100	0.23155
4	Partial Least Squares	0.76167	0.25202
5	MBR	0.74367	0.25652
6	Rule Induction	0.70700	0.26102
7	Decision Tree	0.68800	0.27391
8	Neural Network	0.66500	0.28892

Table 3: Averaged results on the 3 test sets from the original data analysis

Partition	LASSO	Stepwise	Forward	Backward
60/40	47	3	3	26
70/30	30	10	11	27
80/20	30	8	19	44
Average	35.7	7.0	11.0	32.3

Table 4: Number of variables selected from the metadata

THE UNIVERSITY OF ALABAMA® | Culverhouse College of Commerce

SAS®: Cataclysmic Damage to Power Grids and Telecommunications

Taylor K. Larkin, Denise J. McManus

The University of Alabama

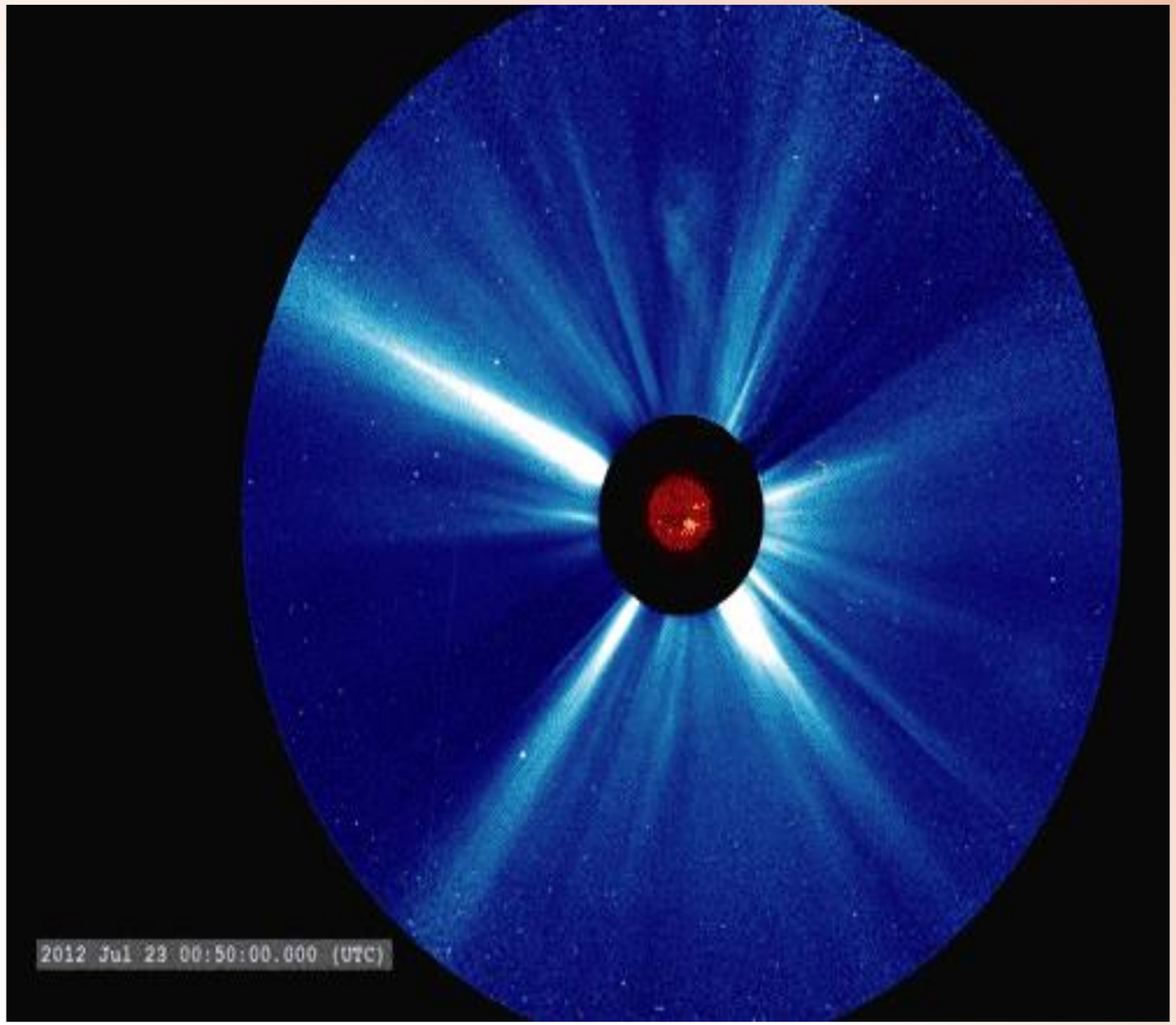

Figure 5: Animation of CME shown through a coronagraph. Credit listed in acknowledgments section

Discussion

Figure 3 shows an example of coefficient paths for the LASSO for one of the data partitions as parameter *t* is increased with each step with the stopping point denoted by the vertical line. Figure 4 demonstrates an example of the chosen base-learners given by LASSO from one of the data partitions. Tables 2 and 3 report the averaged performance metrics for the models for the metadata and the original data, respectively. Table 4 displays the average number of chosen base-learners used for prediction by each of the variable selection techniques. The results show that inducing regression with LASSO variable selection on the metadata yields a **2.33% increase in ROC and 10.58% decrease in misclassification rate** compared to the best model (via ROC) on the original data. This provides strong evidence that the stacked generalization delivers a more productive learning framework. In addition, while the LASSO chooses more base-learners for prediction than traditional variable selection techniques, it still performs favorably as a meta-learner, especially compared to when the LASSO is executed on the original data. Furthermore, it is no surprise that the full regression model produces the worst results, since it is unable to estimate all the parameters in a $p > N$ situation.

Conclusion

In this work, a regularized regression approach, the LASSO, is examined as a meta-learner against traditional regression methods for predicting dangerous CMEs using metadata from a stacked generalization framework. LASSO enjoys both being able to penalize the regression coefficients while also promoting sparse solutions, which is important in cases where $p > N$. Using the power of SAS Enterprise Miner, LASSO solutions are computed through the use of the SAS LARS node. Two main questions are addressed:

1) Does using LASSO variable selection perform better than traditional regression variable selection methodologies?
2) Does implementing stacked generalization produce better predictive outcomes?

The results exhibit evidence that the answer to both of these questions is yes. The results show that using LASSO variable selection not only yields the best averaged performance in terms of ROC and misclassification rate on the metadata, but also outperforms the best model executed on the original data. Specifically, the predictions made on the metadata increase ROC by 2.33% and decrease the misclassification rate by 10.58% when compared to those made on the original dataset, even against a diverse set of models. Given the potential cataclysmic damage that CMEs can wreak on telecommunication and power companies, advanced techniques for improving classification performance are an absolute necessity for saving these industries millions of dollars.

Acknowledgements and References

Taylor K. Larkin and Denise J. McManus

The University of Alabama

Acknowledgements

We would like to thank the National Aeronautic and Space Administration (NASA) for their creative visualizations. Credit for figure 1 goes to NASA/Goddard Space Flight Center Conceptual Image Lab. Credit for figure 5 goes to NASA's Scientific Visualization Studio. In addition, we would like to thank NASA for the LASCO observations. This CME catalog is generated and maintained at the CDAW Data Center by NASA and The Catholic University of America in cooperation with the Naval Research Laboratory. SOHO is a project of international cooperation between ESA and NASA. Finally, we would like to thank the NOAA for their public use database found at ftp://ftp.swpc.noaa.gov/pub/warehouse.

References

1. Baker, D. N., Li, X., Pulkkinen, A., Ngwira, C. M., Mays, M. L., Galvin, A. B., & Simunac, K. D. C. (2013). A major solar eruptive event in July 2012: Defining extreme space weather scenarios. *Space Weather*, *11*(10), 585-591.
2. Bridgman, T. (2014, July 23). As Seen by STEREO-A: The Carrington-Class CME of 2012. Retrieved September 8, 2015, from http://svs.gsfc.nasa.gov/vis/a000000/a004100/a004177/index.html
3. Board, S. S. (2008). *Severe Space Weather Events--Understanding Societal and Economic Impacts:: A Workshop Report*. National Academies Press.
4. Boteler, D. H. (2006). The super storms of August/September 1859 and their effects on the telegraph system. *Advances in Space Research*, *38*(2), 159-172.
5. Cane, H. V., & Richardson, I. G. (2003). Interplanetary coronal mass ejections in the near-Earth solar wind during 1996–2002. *Journal of Geophysical Research: Space Physics (1978–2012)*, *108*(A4).
6. Coronal Mass Ejections. (2012, March 5). Retrieved September 8, 2015, from http://helios.gsfc.nasa.gov/cme.html
7. Cohen, R. A. (2006, March). Introducing the GLMSELECT procedure for model selection. In *Proceedings of the Thirty-first Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc*.
8. Gopalswamy, N., Yashiro, S., Michalek, G., Stenborg, G., Vourlidas, A., Freeland, S., & Howard, R. (2009). The soho/lasco cme catalog. *Earth, Moon, and Planets*, *104*(1-4), 295-313.
9. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.
10. Howard, T. (2011). *Coronal mass ejections: An introduction* (Vol. 376). Springer Science & Business Media.
11. Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1-26.
12. Odenwald, S., Green, J., & Taylor, W. (2006). Forecasting the impact of an 1859-calibre superstorm on satellite resources. *Advances in Space Research*,*38*(2), 280-297.
13. R Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.
14. Reid, S., & Grudic, G. (2009). Regularized linear models in stacked generalization. In *Multiple Classifier Systems* (pp. 112-121). Springer Berlin Heidelberg.
15. Richardson, I. G., & Cane, H. V. (2010). Near-Earth interplanetary coronal mass ejections during solar cycle 23 (1996–2009): Catalog and summary of properties. *Solar Physics*, *264*(1), 189-237.
16. Sarma, K. S. (2013). *Predictive modeling with SAS Enterprise Miner: Practical solutions for business applications*. SAS Institute.
17. Seewald, A. K. (2002). How to Make Stacking Better and Faster While Also Taking Care of an Unknown Weakness (C. Sammut & A. G. Hoffmann, Eds.). In Machine learning: Proceedings of the Nineteenth International Conference (ICML 2002): University of New South Wales, Sydney, Australia, July 8-12, 2002 (pp. 554-561). San Francisco, CA: Morgan Kaufmann.
18. Space Weather Prediction Center. (n.d.). FTP directory /pub/warehouse at ftp.swpc.noaa.gov. Retrieved September 8, 2015, from ftp://ftp.swpc.noaa.gov/pub/warehouse
19. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.
20. Ting, K. M., & Witten, I. H. (1999). Issues in stacked generalization. *J. Artif. Intell. Res.(JAIR)*, *10*, 271-289.
21. Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241-259.

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

SAS®: Cataclysmic Damage

Power Grids and Telecommunications

#SASGF