Solar flare prediction over 24 hour period

R Tumber

07/09/2021

Introduction

The goal of this project was to determine the number of solar flares of classes C, M and X that occurred in a 24 hour period.
The data for this was obtained through the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Solar+Flare.

The dataset consisted of two separate sections, with the second having been deemed more reliable due to the application of more error correction than the first. This being the case, just this second set was downloaded for this project.

The downloaded data consisted of 1066 rows of 13 variables, all factor variables except three which were the target variables. There was no missing data.

Attribute	Explanation
1. Code for class (modified Zurich class)	(A,B,C,D,E,F,H)
2. Code for largest spot size	(X,R,S,A,H,K)
3. Code for spot distribution	(X,O,I,C)
4. Activity	(1 = reduced, 2 = unchanged)
5. Evolution	(1 = decay, 2 = no growth, 3 = growth)
6. Previous 24 hour flare activity code	(1 = nothing as big as an M1, 2 = one M1, 3 = more activity than one M1)
7. Historically-complex	(1 = Yes, 2 = No)
8. Did region become historically complex on this pass across the sun's disk	(1 = yes, 2 = no)
9. Area	(1 = small, 2 = large)
10. Area of the largest spot	(1 = <=5, 2 = >5)
11. C-class flares production by this region in the following 24 hours (common flares)	Number
12. M-class flares production by this region in the following 24 hours (moderate flares)	Number
13. X-class flares production by this region in the following 24 hours (severe flares)	Number

The distribution of the number of flare events of each class was examined against each prediction variable to determine any significance visually. It was found the dataset was highly imbalanced with respect to zero flare events, so the above was repeated for one flare event and over. Due to the nature and distribution of the dataset the scope of the project was reduced to prediction of C-class flare events of values 0-3, to avoid potential inaccuracy introduced by insufficient data.

Due to the discrete nature of flare events it was first decided to experiment and treat this as a classification problem. This proved unsuccessful so the experiment was shelved and the problem reverted to a regression task.

A variety of models and feature selection methods were used to determine the best approach in constructing the model however RMSE remained high in all cases, resulting in a model that proved inaccurate much of the time, a possible result of the frequency of zero flare events in the dataset.

Method and Analysis

Dataset Preparation and initial examination

The second data file, the one with the greater degree of error correction, was downloaded from the UCI Machine Learning Repository, imported and arranged into a useful format. Column headers & explanations (as above) were scraped from the hosting webpage and applied before the raw data were examined. The dataset consisted of 1066 rows of 13 variables, 1-10 were factor variables and 11-13 contained the number of C-class_flares_(common), M-class_flares_(moderate) and X-class_flares_(severe) that were observed. These existed as factor variables for the attempted classification and as numeric for the regression.

Data examination

To begin, the distribution of the different class flare events was examined.

Number of Flares	C_class	M_class	X_class
0	884	1030	1061
1	112	29	4
2	33	3	1
3	20	2	0
4	9	1	0
5	4	1	0
6	3	0	0
7	0	0	0
8	1	0	0

For C class events a little over 17% of observations flare activity, this dropped to just over 3% for M class events and to less than 1% for X class events.

The distribution of the ten factor variables in relation to the three target variables was then examined, before the process was repeated but this time only including those records where one or more flare event had taken place.

At this stage modelling for M and X class flares was discontinued due to the small sample size. From here observations and processes were concerned solely with C-class events.

Where observations with zero flares are removed, some relationships begin to emerge.

- Code for class (modified Zurich class) - Class D is most likely
- Code for largest spot size - S & A spots most likely
- Code for spot distribution - I & O distribution most likely
- Activity - Activity is typically reduced
- Evolution - Growth or no-growth usually observed, decay is rare
- Previous 24 hour flare activity code - Usually 1, nothing as big as an M1
- Historically-complex - Fairly even split, tending towards 2, not historically complex
- Did region become historically complex on this pass across the sun's disk - Mostly not
- Area - Area is mostly small
- Area of the largest spot - Always <= 5

Each variable was then plotted against the others and flare count to see if any relationship could be refined, however since the was nothing immediately telling the variables that were carried forward to modelling were:

- Code for class (modified Zurich class)
- Code for largest spot size
- Code for spot distribution
- Activity
- Evolution
- Previous 24 hour flare activity code
- Did region become historically complex on this pass across the sun's disk
- Area
- *Potentially Historically-complex variable, though the fairly even split could indicate it is of low relevance*

Data Modelling, Results & Discussion

The modelling can be split here into two separate sections, the first was to treat the number of flare events as factors and attempt a classification and if unsuccessful to revert to modelling as a regression. In both cases caret was used to train the model and predict results and in both cases the same 70:30 train/test split was used. It was decided to drop records for four flare events and over due to small sample size, to attempt to get a good baseline, with a mind to adding them back should a favourable outcome arise.

Classification

Initial modelling was performed using a variety of models and used over or under sampling to attempt to compensate for the dataset imbalance. The evaluation metric used in the classification was Macro-averaged F1 score, to account for the imbalanced proportions of the different classes.
This proved ineffective so the training dataset was artificially balanced using the ROSE package. Synthetic data was generated and used to even the class balance and the models then retrained without over and under sampling. This approach was successful in improving the F1 score on training and more so after other models were tested and tuned. When this model was used to attempt classification on the test set the results proved as inaccurate as the imbalanced set.
This suggests one or more of the following:

1. Insufficient data in the imbalanced training set to accurately synthesise new data points - There is little that can be done if this is the case.
2. weak relationships between dataset features and the number of C class solar flares - The data exploration does not appear to support this conclusion
3. Over trained models - In model tuning, the training/validation split does not suggest vast overtraining
4. A categorisation exercise is not the correct approach

Given the above, the model building shifted to a regression approach.

Regression

Before regression model construction, dummy variables were set up for the factor variables and the target variable class was set to numeric. Initially a basic linear regression was performed on the full dataset to determine a basic viability, following this A similar approach to the classification modelling was taken, using 10 fold cross validation, however the performance was determined by calculation of RMSE. The features to be modelled were first determined by recursive feature elimination.
The features to be modelled as a result of this were *Historically_complex, Mod_Zur_Class_Code, Lrgst_spot_size_code, Area, Evolution.* The model types used to begin with were ranger, xgbtree, pcaNNet, glm. The models were constructed using caret and the variables added individually so the effect on RMSE could be better monitored. In practice, only variables *Mod_Zur_Class_Code, Historically_complex* and *Lrgst_spot_size_code* reduced the RMSE, though after tuning the models, performance was not good. In an effort to improve the results, additional random forest and neural net models were applied as these performed best in the models created so far. These extra models used Rborist, rf, neuralnet and mlpML, the results of which were mixed and produced no real improvement.
At this stage a review of the basic linear regression revealed the significant features were not a match for those picked out in the recursive feature elimination. In attempting to clarify these differences, three methods were

used to determine the features that should be modelled: Variable importance from a random forest model (*Historically_complex, Mod_Zur_Class_Code, Lrgst_spot_size_code, Area, Evolution, Spot_dist_code*), variable importance from an earth model (*Mod_Zur_Class_Code, Activity, Lrgst_spot_size_code, Spot_dist_code, Area*) and stepwise regression (*Mod_Zur_Class_Code, Lrgst_spot_size_code, Activity, Area*). Each of these methods determined a different set of features so models were constructed using the features selected and the best performing six models, ranger, pcaNNet, Rborist, rf, neuralnet, mlpML. Initial evaluation revealed no major gains however tuning the models did improve accuracy a little.

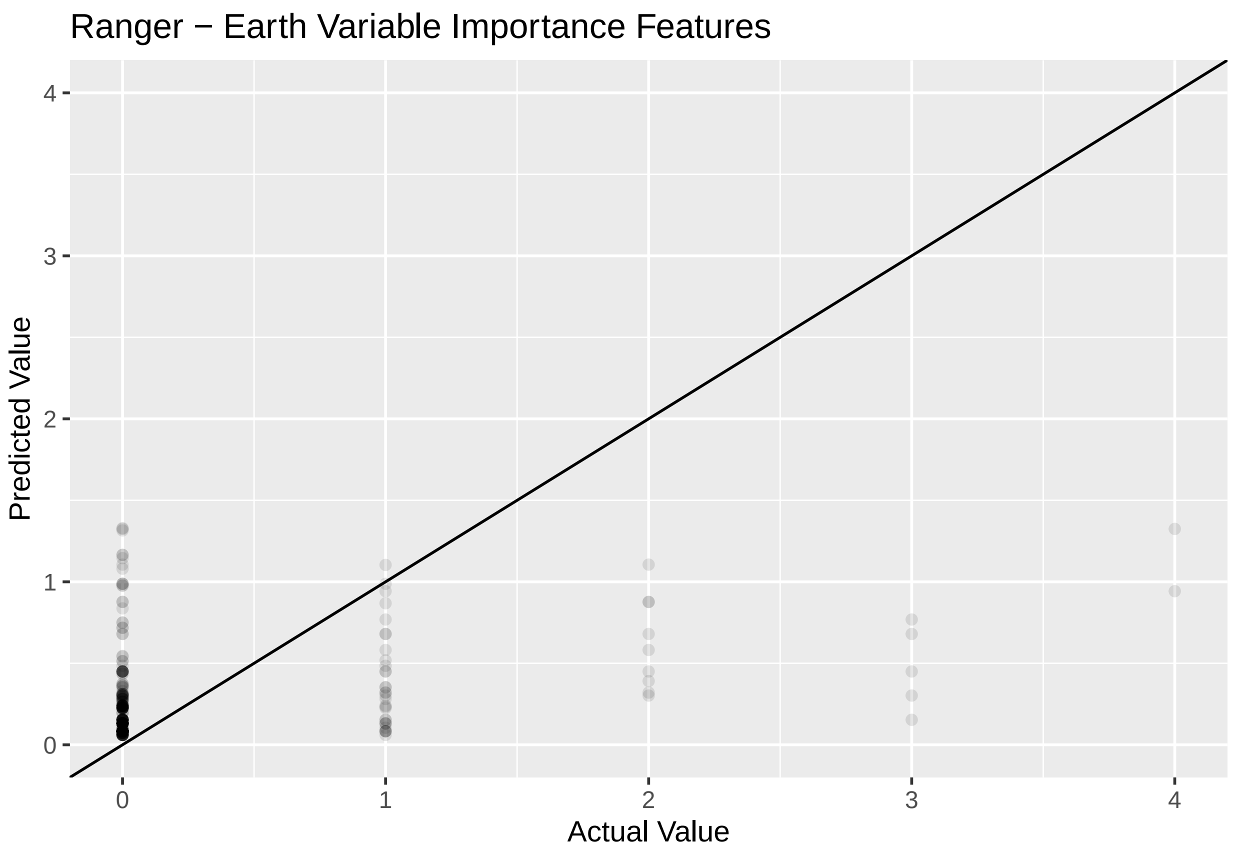
Model	Feature Selection	Model RMSE	Validation Predicted RMSE
pcaNNet	Earth Variable Importance	0.7656535	0.8739533
Rborist	Stepwise Regression	0.7665033	0.8708514
ranger	Stepwise Regression	0.7679465	0.8733137
pcaNNet	Variable Importance	0.7689311	0.8758313
pcaNNet	Stepwise Regression	0.7699001	0.8916498
rf	Stepwise Regression	0.7700051	0.8729178
Rborist	Earth Variable Importance	0.7761082	0.8686111
ranger	Earth Variable Importance	0.7767709	0.8715685
rf	Earth Variable Importance	0.7775257	0.8719014
neurlnet	Stepwise Regression	0.7796952	0.8976471
neurlnet	Variable Importance	0.7824911	0.9062493
mlpML	Earth Variable Importance	0.7835508	0.9034544
ranger	Variable Importance	0.7840376	0.8813679
rf	Variable Importance	0.7849812	0.8771754
Rborist	Variable Importance	0.7852066	0.8792416
mlpML	Stepwise Regression	0.7875038	0.8796425
neurlnet	Earth Variable Importance	0.7934999	0.9100477
mlpML	Variable Importance	0.8018005	0.9026255

The observed differences between the training and validation RMSEs, indicating a degree of overtraining. In terms of model performance, stepwise regression generally gave a more favourable RMSE, however when it came to RMSE on the validation set it was more split between Variable Importance calculated from an Earth model and Stepwise Regression. This result may have been a little misleading since it could be reasonably expected for this to change in dealing with the potential overtraining. As a result of this, the three best performing models were retrained using only five fold cross validation and the tuned hyperparameters previously determined. Predictions were made on the test set using these retrained models and the RMSEs showed a marked increase in accuracy however the overall performance remained poor.

Model	RMSE
Rborist Earth Features	0.6066096
Rborist Stepwise Reg. Features	0.6124566
Ranger Earth Features	0.6049570

Conclusion

The best performing model, the ranger model built using features selected by variable importance from an Earth model was only 0.6042541, which does not represent a good result. A brief examination of the predicted values and the actual values shows the model is highly likely to under predict for 1+ flare events and over predict for zero flare events. As one might expect, the high frequency of zero flare events in the training data has resulted in a model that is skewed in that direction.



While disappointing, this does raise two questions that can be examined in the future.
Firstly, could there be a value in first creating a model to make a binary choice between zero and 1+ flare events followed by a model built specifically to predict the number of flare events over 0?
Secondly, can the accuracy be improved by using random over and under sampling to balance the dataset for use with regression models?

References

Dataset URL https://archive.ics.uci.edu/ml/machine-learning-databases/solar-flare/