SOLAR FLARE PREDICTION WITH ARTIFICIAL INTELLIGENCE

January 12, 2023

ABSTRACT

Design a program that predicts solar flares with the given dataset and find the attributes that most affect the algorithm. Making inferences according to the sunspot classifications that affect the forecast most.

Alper Birinci

22034899

Contents

Introduction

Solar flares are the strongest magnetic activity in the solar system. According to information shared in a report (Fletcher et al., 2011), they can release more than 1032 ergs of energy in tens of minutes. Radiation is emitted during solar flares. Solar flares are closely related to coronal mass ejections (CME). Solar flares can also affect people. It can damage the climate, the Global Positioning System (GPS), satellites, astronauts, and even electricity. Knowledge of how harmful solar flares are and when they will happen is important to get the best possible recovery from such damage. Geomagnetic storms triggered by solar activity cause the formation of auroras seen at the poles. The formation of solar flares is related to sunspots. The magnetic field inside sunspots is so strong that heat cannot reach the surface, so sunspots are cooler than the surface of the sun. Magnetic events around sunspots cause sudden bursts of energy called solar flares (Fletcher et al., 2011).

Today, according to the information shared by NASA on its blogs (Interrante, 2022), solar flares are studied by looking at the light they emit. Solar activity has an 11-year cycle. At the beginning of every 11 years, there are fewer sunspots and fewer solar flares. Solar activities increase towards the middle of the cycle and then decrease towards the end of the cycle again. One of the ways to predict solar flares is sunspots. Different classification methods have been created for sunspots. One of the most famous classifications is the McIntosh classification. McIntosh's classification has 3 components. The prediction of solar flares becomes easier thanks to 60 different types of classification that are formed together with these 3 components. These 3 components will be examined in more detail in the background section. Artificial intelligence, whose usage areas are expanding, and its popularity is increasing day by day, can also be used for examining sunspots and predicting possible solar flares.

Background

The aim of this study is to create an artificial intelligence that predicts solar flares by using machine learning with a data set about sunspots and for only C-class solar flares, the most influential features will be found according to the algorithm. Python will be used for data processing and machine learning. RandomForest algorithm, which is a very popular algorithm at this time, will be used for machine learning. A dataset shared on the Machine Learning Repository site (Bradshaw, 1989) will be used as the dataset. This dataset has 1389 samples and has a total of 13 attributes, 10 for training purposes and 3 for target purposes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1389 entries, 0 to 1388
Data columns (total 13 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Zurich               1389 non-null   object
 1   LargestSpot          1389 non-null   object
 2   SunspotDistribution  1389 non-null   object
 3   Activity             1389 non-null   int64
 4   Evolution            1389 non-null   int64
 5   24hour               1389 non-null   int64
 6   Historicallycomplex  1389 non-null   int64
 7   sunsdisk             1389 non-null   int64
 8   Area                 1389 non-null   int64
 9   AreaLargestspot      1389 non-null   int64
 10  Cclass24             1389 non-null   int64
 11  Mclass24             1389 non-null   int64
 12  Xclass24             1389 non-null   int64
dtypes: int64(10), object(3)
```

Image 1: Information about the attributes in the dataset

The first 3 attributes in the data set are the 3 components of the McIntosh classification, which is the sunspot classification method described in the introduction. The name of the method in which sunspots are classified according to whether there is a penumbra, the length of the penumbra, and how the penumbra is distributed is the Modified Zurich Class. Penumbra: Largest Spot is the name of the method in which the classification is made according to the type of the largest spot in the sunspot group, penumbra type, penumbra size, and symmetry of penumbra combination. Sunspot Distribution is the third and final method used in classification. This method provides more information about the sunspot group, and more importantly, it provides information about whether there are strong points near the polar reversal line extending between the leader and follower points. A total of 60 classes emerged according to these 3 separate components. (McIntosh, 1990). Classes of methods and the number of combinations are shown in **Image 2**.

Class	Penumbra: largest spot	Distribution	Number of types
(A)	(*x*)	(*x*)	1
(B)	(*x*)	(*o i*)	2
(C)	(*r s a h k*)	(*o i*)	10
(D E F)	(*r*)	(*o i*)	6
(D E F)	(*s a h k*)	(*o i c*)	36
(H)	(*r s a h k*)	(*x*)	5
Total allowed types:			60

Image 2: Groups, and number of types in the McIntosh sunspot classification (McIntosh, 1990)

The other 7 attributes provide further information about sunspots. These attributes include information such as whether the sunspot grows (evolution), its area, and its historical complexity (Bradshaw, 1989).

The last three attributes in the dataset have solar flare classes. The shared values given here show how many times the solar flare occurred in the last 24 hours in which class. These classes are C, M, and X. Normally there are 5 classes in this classification. These are A, B, C, M, and X, in order from weak to strong, but classes A and B will not be examined in this report and more focus will be placed on the C class. Each class has levels from 1 to 9, and all classes are 10 times stronger than the classes below it, so M is 10 times stronger than C, and X is 100 times stronger than C. Solar flares in class C are too weak to affect the world much. Class M solar flares can cause radio interruptions at the poles (solar flares begin to affect the Earth from the poles) and radiation storms that can endanger astronauts. X classes, on the other hand, can go up to 9 as they are the last level. X classes can affect GPS, satellites, communication systems, and more (Zell, 2013).

Many python libraries will be used in the project. A ready-made class structure in the “sklearn” library will be used for the Random Forest algorithm. The “pandas” library will be used to read the data, and the “matplotlib.pyplot” library will be used to better explain the report with various visualizations. These libraries are the libraries that will be used throughout the report, and various other libraries will also be used.

Data and Methodology

In this section, data cleaning will be done first to achieve the best performance. Afterward, machine learning will be done with the Random Forest algorithm. After the program is designed, various investigations will be made to better understand C-class solar flares.

Data

First, the data will be imported into python with the "Pandas" library. After the data is transferred, the missing data and duplicate data will be checked first. It can be seen in **Image 3** and **Image 4**.

```
Zurich                0
LargestSpot           0
SunspotDistribution   0
Activity              0
Evolution             0
24hour                0
Historicallycomplex   0
sunsdisk              0
Area                  0
AreaLargestspot       0
Cclass24              0
Mclass24              0
Xclass24              0
dtype: int64
```

Image 3: Sum of missing data for each attribute

```
Number of duplicate data:
893
Number of duplicate data after dropping:
0
New Shape of data:
(496, 13)
```

Image 4: Number of duplicate data and how much data would remain if it dropped

There is no missing data in the data. There are 893 duplicate values, but there is no deletion of duplicate values here. Duplicate values in the dataset are based on measurements and would be correct to affect the algorithm.

In order to get more efficient results, the values in the target columns are converted to binary. So now the value will be 1 if there was a solar flare and 0 if it didn't. Also, the first 3 attributes are not suitable for machine learning because they are in the object structure. For this reason, each value of these 3 attributes is created as another attribute with the “get.dummies” function in the "pandas" library. These new attributes are stored as binary. You can see an example of dummy variables in **Image 5**.

CATEGORICAL VARIABLE

sex
male
female
female
male
male
male
male
female
male

DUMMY VARIABLES THAT ENCODE THE SAME INFORMATION

male_dummy	female_dummy
1	0
0	1
0	1
1	0
1	0
1	0
1	0
0	1
1	0

Image 5: An example of dummy variables (Ebner, 2022)

In the final part of the data editing, the imbalance in the target will be checked. SMOTE, an oversampling method, will be used to eliminate the imbalance. The purpose of the Smote algorithm is to multiply the samples in the minority group so that the two groups are equal. When sampling the minority group with SMOTE, the Euclidean distance from the nearest k-neighbor is multiplied by a random number between 1 and 0. For the implementation of the SMOTE algorithm, the ready-made function in the "imblearn" library will be used. The difference between the before and after use of the SMOTE algorithm is shared in **Figure 1** and **Figure 2**.

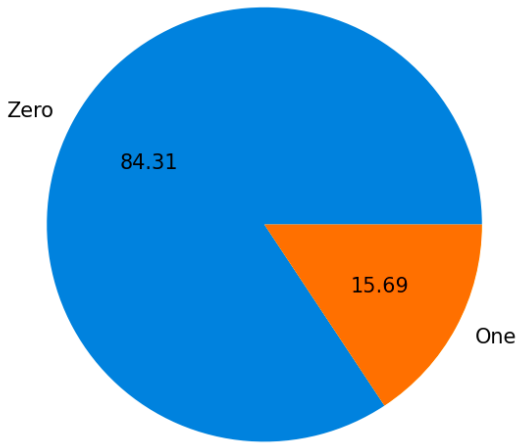

Figure 1: Difference before applying the Smote algorithm

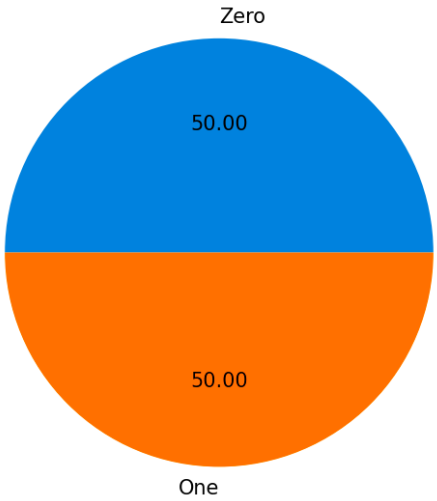

Figure 2: Difference after applying the Smote algorithm

Methodology

Random Forest algorithm will be used for machine learning. A python library "sklearn" will be used to implement the random forest algorithm on the data set. 5 Different test training rates will be tested. Then, thanks to the "Sklearn" library, when the algorithm is applied, it will be possible to show the attributes that affect the result the most as a table. For this, the "matplotlib.pyplot" library will be used and in the next section, various inferences will be made by sharing it with the user.

Random Forest

It is one of the first algorithms that comes to mind when considering machine learning algorithms in these times. The Random Forest algorithm was developed by L. Breiman in 2001 (Breiman, 2001). L. Breiman was inspired by many previous studies while developing the random forest algorithm. The Random Forest algorithm uses the divide and conquers technique. According to this technique, a problem is divided into many small parts, and after recursively solving these small parts, the solutions are combined to solve the main problem. Some of the reasons why the Random Forest algorithm is so popular are that it has very few parameters to adjust, it can be applied to a wide variety of estimation problems, it can handle both small and large-sized problems, and it is easy to implement (Biau and Scornet, 2016).

Analysis and Discussions

In this section, the scores of the Random Forest algorithm will be shared. Afterward, the factors that affect C-class solar flares the most according to artificial intelligence will be examined. To obtain the best result, 5 different ratios were tried while applying the Random Forest algorithm. It is observed that the best ratio out of 5 different ratios is 75:25. The accuracy result of 5 different test training ratios can be seen in **Table 1**.

RATES	85:15	80:20	75:25	70:30	65:35
ACCURACY	0.65	0.72	0.73	0.70	0.72

Table 1: The accuracy results of 5 different test training ratios

The most critical features that affect the result while applying the Random Forest algorithm can be seen in **Table 2**.

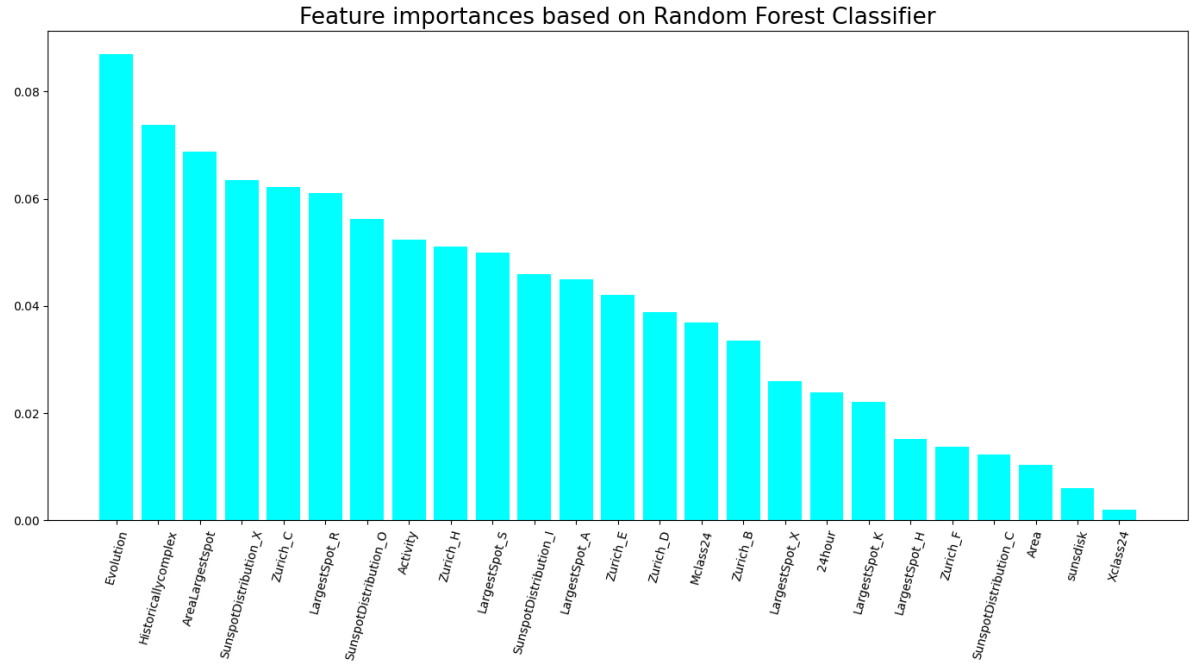

Table 2: Feature importance based on Random Forest Classifier

When the random forest algorithm was run on the data set, the top 3 most influential attributes were "Evolution", "HistoricallyComplex", and "AreaLargestspot". Then, 3 different values of the classifications used for sunspots appear. Then, 3 different values of the classifications used for sunspots appear. The most influential class on the distribution of sunspots was X. According to the description of this class, the sunspot group is unipolar but unidentified, in which case it could be inferred that there is a class A or H in the Zurich classification. The fact that the second most influential class H in the Zurich classification supports this situation. Class C was the most influential class in the Zurich classification. According to the characteristics of this class, there is a bipolar sunspot group, and especially a spot of this group has penumbra (McIntosh, 1990). Class C of the Zurich classification is shown in **Image 6** as an example.

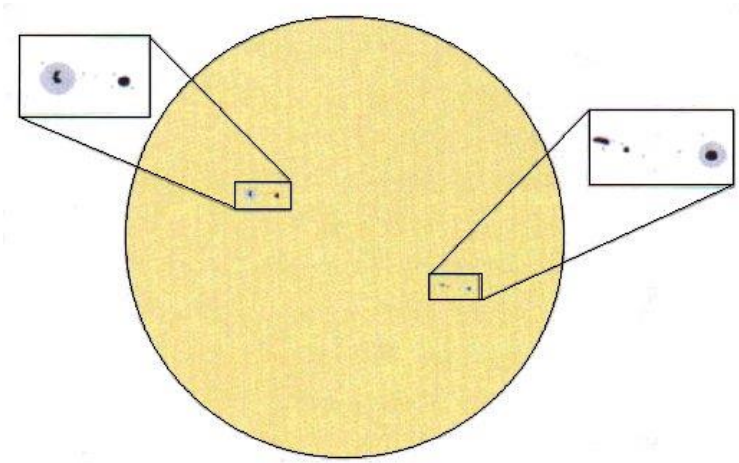

Image 6: Condition of sunspots according to class C of the Zurich Classification (Fleming, n.d.)

Finally, it has been observed that the biggest factor in the Largest Spot classification is the r class. According to this class, primitive penumbra partially surrounds the largest sunspot. This penumbra is grainy and incomplete. Also, this penumbra is brighter than the mature penumbra. In this case, it is thought that the primitive penumbra may be in the formation or dissolution stage (McIntosh, 1990).

Conclusions and Future Work

In conclusion, there are 5 different solar flare classes, of which C, M, and X classes affect us. To predict these solar flares, artificial intelligence can examine sunspots and more. After the machine learning application with the Random Forest algorithm, C-class solar flares were predicted. Also, some of the attributes used for machine learning were related to sunspots. While the random forest algorithm is working, the features that affect the result the most are examined thanks to the python library, and some inferences have been made about sunspots.

If I had more time, I would have studied the other solar flare classes. While doing this project, there was another issue that caught my attention. Class X solar flares didn't seem to have enough examples for machine learning, as solar flares occur more rarely than other solar flare classes. Because in this case, the algorithm would conclude that there would be no X solar flares except in a few specific cases. This may indeed be true. However, there may be an incorrect application as there are insufficient examples. I would like to do another research to solve the question that was stuck in my head.

Bibliography

Biau, G. and Scornet, E. (2016). Rejoinder on: A random forest guided tour. *TEST*, 25(2), pp.264–268. doi:10.1007/s11749-016-0488-0.

Bradshaw, G. (1989). *UCI Machine Learning Repository: Solar Flare Data Set*. [online] archive.ics.uci.edu. Available at: http://archive.ics.uci.edu/ml/datasets/solar+flare.

Breiman, L. (2001). Random Forests. *Machine Learning*, [online] 45(1), pp.5–32. doi:10.1023/a:1010933404324.

Ebner, J. (2022). *How to Use Pandas Get Dummies in Python - Sharp Sight*. [online] www.sharpsightlabs.com. Available at: https://www.sharpsightlabs.com/blog/pandas-get-dummies/.

Fleming, T. (n.d.). *The Zurich Classification System of Sunspot Groups | aavso*. [online] www.aavso.org. Available at: https://www.aavso.org/zurich-classification-system-sunspot-groups.

Fletcher, L., Dennis, B.R., Hudson, H.S., Krucker, S., Phillips, K., Veronig, A., Battaglia, M., Bone, L., Caspi, A., Chen, Q., Gallagher, P., Grigis, P.T., Ji, H., Liu, W., Milligan, R.O. and Temmer, M. (2011). An Observational Overview of Solar Flares. *Space Science Reviews*, [online] 159(1), p.19. doi:10.1007/s11214-010-9701-8.

Interrante, A. (2022). *Solar Flares FAQs – Solar Cycle 25*. [online] blogs.nasa.gov. Available at: https://blogs.nasa.gov/solarcycle25/2022/06/10/solar-flares-faqs/.

McIntosh, P.S. (1990). The classification of sunspot groups. *Solar Physics*, 125(2), pp.251–267. doi:10.1007/bf00158405.

Zell, H. (2013). *Solar Flares: What Does It Take to Be X-Class?* [online] NASA. Available at: https://www.nasa.gov/mission_pages/sunearth/news/X-class-flares.html.