Regression Analysis of Solar Flares: A Multilayer Perceptron Approach with Feature Selection Techniques

Mohamad Tamer Rabie
Department of Computer Engineering
University of Sharjah
Sharjah, UAE
U16103165@sharjah.ac.ae

Ali Bou Nassif
Department of Computer Engineering
University of Sharjah
Sharjah, UAE
anassif@sharjah.ac.ae

Maha AlaaEddin
Department of Computer Engineering
University of Sharjah
Sharjah, UAE
malaaeddin@sharjah.ac.ae

***Abstract*— In this paper, we are going to analyze and test the solar flare dataset from the UCI Machine Learning Repository [10], by improving it using feature selection techniques such as stepwise regression, detecting the most effective attributes, importance ranker using k-fold and leave-one-out cross validation methods. We are going test the model by evaluating the dataset using Multi-linear regression, by looking at the P-values and the VIF to show the effectivness of the dataset attributes. Multilayer perceptron model will be created using the holdout regression by partitioning the dataset into training and testing to model the testing dataset into an MLP model with 5 hidden layers. The model will show the mean absolute error and variance of the model to test its accuracy.**

I. Introduction

A solar flare is essentially a large combustion on the external surface of the Sun that happens when magnetic field lines from sunspots coil and flare up. A solar flare is outlined as an rapid, intense, and abrupt alteration in brightness. A solar flare happens once magnetic energy that has designed up within the star atmosphere is suddenly discharged. In mere minutes, matter is heated to several various degrees and radiation is emitted across the whole electromagnetic spectrum, from the long wavelength end at radio waves, to the short wavelength end at X-rays and gamma rays. the quantity of energy discharged is corresponding to various nuclear bombs exploding all at identical time.

Solar flares are associate typically prevalence once the Sun is active within the years around solar most. several solar flares will occur on only one day throughout this period. Around solar minimum, solar flares may occur once per week. Huge flares are less periodic than lesser ones. Coronal mass ejection is when some solar flares will launch vast clouds of solar plasma into space. A geomagnetic storm and intense auroral displays is caused once a coronal mass ejection arrives at Earth.

Solar flares are sorted as A, B, C, M or X in line with the peak flux (in watts per square meter, W/m2) of one to eight Ångströms X-rays close to Earth, as measured by XRS instrument on-board the GOES-15 satellite that is in a geosynchronous orbit over the Pacific Ocean. [1]

The benefit of regression analysis on the solar flare data set, it refers to a way of mathematically searching for which variables might have an effect. it helps verify which factors is the most significant and effective, and the way those attributes cooperate with one another. The significance of regression analysis is that it provides a robust method that examines the link between each variables of interest.

II. Technical Background

A. Solar flares

Solar flares directly have an effect on the ionosphere and radio communications at the earth, and additionally unleash energetic particles into space. Therefore, to understand and predict 'space weather' and the impact of solar activity on the earth, therefore, an understanding of the Machine learning datasets helps us learn more about the effects of solar flares on the earth.

This has necessary implications for understanding and predicting the consequences of solar activity on the earth and in space. If a coronal mass ejection collides with the earth, it will excite a geomagnetic storm [2].

B. Multi-Linear Regression

Multiple linear regression (MLR), additionally better-known as multiple regression, is a statistical technique that uses many informative variables to predict the result of a response variable [23-25]. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and the predicted (dependent) variable [3].

The formula for Multiple linear regression is represented in the following formula:

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}+\epsilon$$

Where the variable y_i is the dependent variable, x_i is the explanatory value, β_0 is the y-intercept, β_p is the slope coefficients for each explanatory variable, ϵ is the model's error term (also known as the residuals)

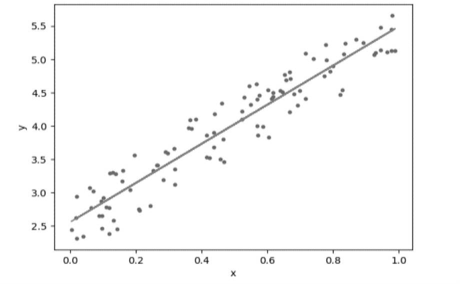


Figure 1

C. Holdout Regression

The first step in developing a machine learning model is training and validation. so as to train and validate a model, you want to initially partition your dataset, that involves selecting what partition of your data to use for the training, validation, and holdout datasets using linear regression [4]. The function approximator fits a function by applying the training set solely. Then the function approximator is asked to predict the output values for the data within the testing set. The errors it makes are accumulated as before to provide the mean absolute test set error, that being used to check the model.

The advantage of this technique is that it's sometimes preferred to the residual method and does not take long to compute. However, its analysis will have a high variance. The analysis could rely heavily on that data points end up within the training set and which end up in the test set, and therefore the analysis could also be considerably completely different looking on however the division is formed [5].

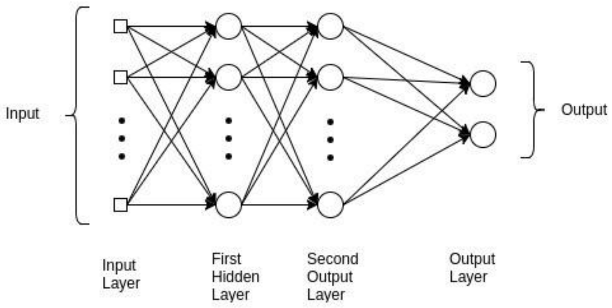
D. Multilayer Perceptron

A perceptron is a single neuron model that was a precursor to larger neural networks. straightforward models of biological brains is accustomed solve troublesome computational tasks just like the predictive modeling tasks we tend to see in machine learning. The goal is to develop robust algorithms and data structures which will be accustomed model tough problems [18].

In MLP, there is more than one combinations of neurons. for instance, for a three-layer network, first layer is the input layer and last will be output layer and middle layer will be known as hidden layer. The input data is fed into the input layer and the output from the output layer. the number of the hidden layer can be increased according to the task. [19].

For the weight of the neuron to be updated, it uses this equation:

*weight = weight + learning_rate * (expected - predicted) * x*



E. Leave-one-out cross validation

Leave-one-out cross-validation is a special case of cross-validation wherever the amount of folds equals the number of instances within the data set. Thus, the training algorithm is applied once for every instance, applying all alternative instances as a training set and using the chosen instance as a single-item test set [6].

With K equal to N, the quantity of data points within the set. meaning that N separate times, the function approximator is trained on all the data aside from one point and a prediction is created for that time. As before the average error is computed and accustomed to assessing the model. [5]

F. K-fold cross validation

The data set is split into k subsets, and therefore the holdout technique is repeats k times. Each time, one of the k subsets is used as the test set and also the other k-1 subsets are combined together to create a training set. Then the average error across all k trials is computed. The advantage of this technique is that it does not matter how the data gets divided. Each data point gets to be in a test set precisely once and gets to be in a training set k-1 times. The variance of the product estimate is reduced as k is incremented. The disadvantage of this technique is that the training algorithm must be rerun from scratch k times, which implies it takes k times as much computation to form an analysis. A variant of this method is to arbitrarily divide the data into a test and training set k different times. The advantage of doing this is that you just can independently select how large every test set is and how many trials you average over [5].

G. Stepwise Regression

Stepwise regression has the ability to add and remove predictors iteratively, within the predictive model, in order to seek out the subset of variables within the dataset leading to the best model, that's a desired model that lowers prediction error.

Stepwise regression is a modification of the forward selection in order that when every step within which a variable is included, all candidate variables within the model are checked to visualize if their significance has been reduced below the stated tolerance level. If a nonsignificant variable is found, it's discarded from the model. Stepwise regression needs two significance levels: one for adding variables and one for removing variables. The cutoff probability for adding variables ought to be below the cutoff probability for removing variables in order that the procedure doesn't get into an infinite loop [7].

H. Feature selection

Feature selection is additionally known as variable selection or attribute selection. it's the automated selection of attributes in your dataset that are most relevant to the predictive modelling drawback you're performing on. Its goal is to reduce the quantity of attributes within the dataset. [8]

Benefits of feature selection on the dataset: [9]

- Reduces Overfitting: Fewer redundant data means that less chance to form selections based on noise.
- Improves Accuracy: Fewer misleading data means that modeling accuracy improves.
- Reduces Training Time: Less data points scale back algorithm complexness and algorithms train quicker.

III. Dataset Preprocessing and feature selections

The Solar flare Dataset which is provided in the in the "UCI Machine Learning Repository" [10], discuses each class attribute, it counts the number of solar flares of a certain class that occur in a 24-hour period. The dataset contains three potential categories, one for the amount of times an explicit type of solar flare occurred in a twenty-four-hour interval. Every instance represents captured features for one active region on the sun. the data are divided into 2 sections. From all the 13 predictors inside the dataset, three classes of flares are predicted. [11][12][13][14][15][16][17]

The number of instances (rows) of the dataset are 1066 rows, with 13 attributes (columns). There are no missing values. The data are divided into 2 sections. The second section has had rather more error correction applied to the it and has therefore been treated as a lot more reliable.

A. *Attribute information:*

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: (1 = yes, 2 = no) on this pass across the sun's disk
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 outliers of the Solar flare dataset has been identified using the 'R' programming Language using Rstudio. The Boxplot method was used since it is suitable for large datasets**.** It can represent a wide range of information in an elegant way, showing the IQR and it can identify the outlier of the dependent variable

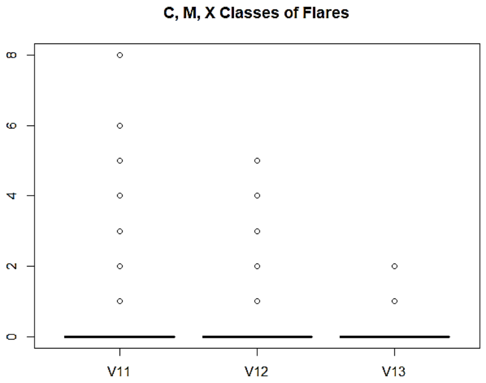


Figure 2

The outliers with values: 4, 5, and 6 have been dealt by removing the whole row from the dataset using Rstudio, since the inputs are causing the output to have an outlier, that means the inputs are bad and need to be removed

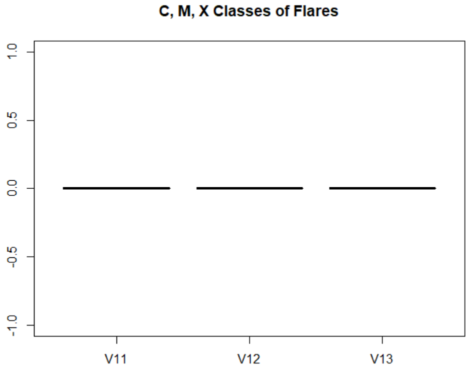


Figure 3

B. *Featrue selection methods*

It's important to get rid of any necessary variables, which will evidently avoid overfitting and reduce training time. Feature selection was done through the R programming language, on the solar flare dataset based on their importance

The results for the Rank Important features are shown below in Fig. 4 and also the Regression subset results, using 10- K fold cross-validation.

Using the “caret” and “mlbench” packages in R, the importance of the features in the dataset to rule out the ineffective attributes. The method used for the K-fold cross validation was linear regression “lm” to split the dataset into training and testing partitions. The plot in figure 4 below shows the rank according to the ROC curve.

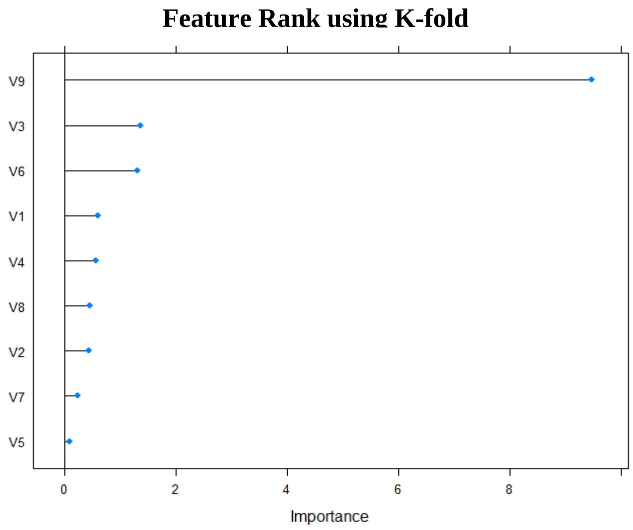


Figure 4

The results show that V10 (Area of the largest spot) is ineffective in that dataset.

The method used for the Leave-one-out cross validation “LOOCV” was also linear regression “lm”. The plot in figure 5 shows the rank according to the ROC curve.

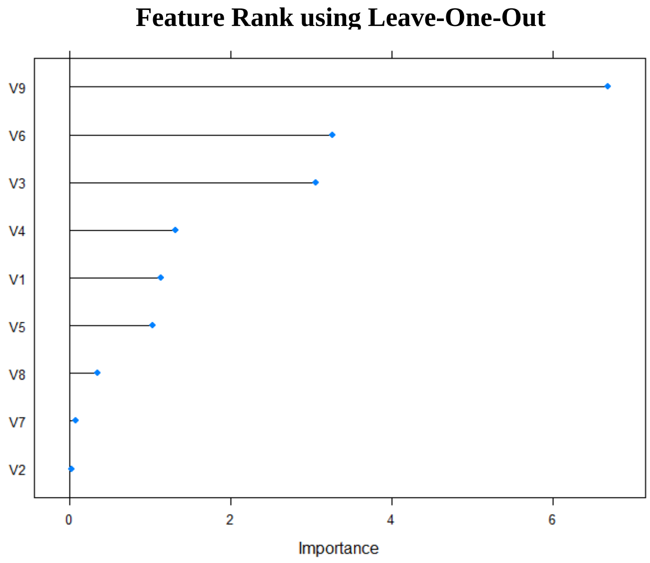


Figure 5

The results for leave one out show a similar effect on the effective features for Solar flare effect on the ionosphere and radio communications at the earth. V10 (Area of the largest spot) is also ineffective in that Feature rank method.

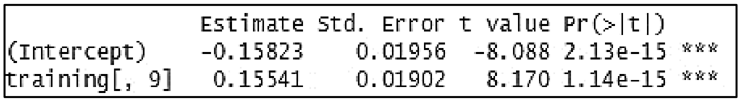
According to this test. The effective features are:

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: (1 = yes, 2 = no) on this pass across the sun's disk
9. Area: (1 = small, 2 = large)

Stepwise Regression

Stepwise regression uses the sequential replacement algorithm “Leapseq” to determine the lowest prediction errors to find out the subset of variables in the data set resulting in the best performing model.

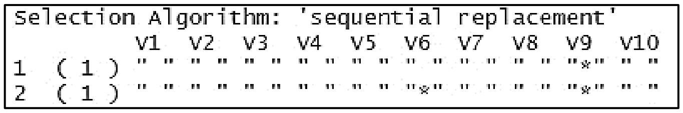
The Holdout regression 80% training and 20% testing split is going to be utilized for finding the step model of the training model to discover the analysis made by stepwise regression.



```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.15823    0.01956  -8.088 2.13e-15 ***
training[, 9]   0.15541    0.01902   8.170 1.14e-15 ***
```

Figure 6

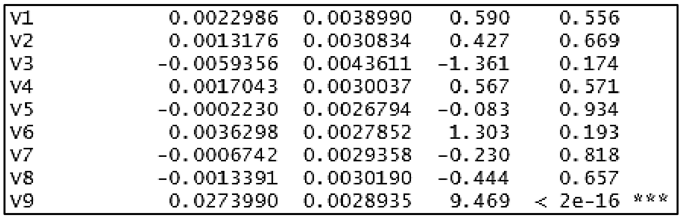
As Figure 6 shows, V9 (Area) is considered to be the most effective attribute, which goes in line with feature selection rank shown using the importance function.



```
Selection Algorithm: 'sequential replacement'
         V1  V2  V3  V4  V5  V6  V7  V8  V9  V10
1  ( 1 ) " " " " " " " " " " " " " " " " "*" " "
2  ( 1 ) " " " " " " " " " " "*" " " " " "*" " "
```

Figure 7

According to Stepwise regression shown in Figure 7, V6 (Previous 24-hour flare activity code) and V9 (Area) are the most effective attributes in the dataset.



```
V1          0.0022986  0.0038990   0.590    0.556
V2          0.0013176  0.0030834   0.427    0.669
V3         -0.0059356  0.0043611  -1.361    0.174
V4          0.0017043  0.0030037   0.567    0.571
V5         -0.0002230  0.0026794  -0.083    0.934
V6          0.0036298  0.0027852   1.303    0.193
V7         -0.0006742  0.0029358  -0.230    0.818
V8         -0.0013391  0.0030190  -0.444    0.657
V9          0.0273990  0.0028935   9.469  < 2e-16 ***
```

Figure 8

Showing the summary of the K-fold cross validation also confirms that V9 (Area) is in fact the most effective attribute in the dataset.

IV. Related work

T. Colak and R. Qahwaji in [20] proposed a way to predict significant solar flare activity using SOHO/Michelson Doppler Imager images, which uses image processing and machine learning to analyze sunspots based on the McIntosh classification method and compared it with NOAA weather predictions.

Other work done by L.E. Boucheron *et. al*. [21] have studied the prediction of solar flare size and time-to-flare by using 38 features that shows the magnetic complexity of the photospheric magnetic field. They presented their results using SVM regression and showed the confusion matrix of the Predicted Flare Size.

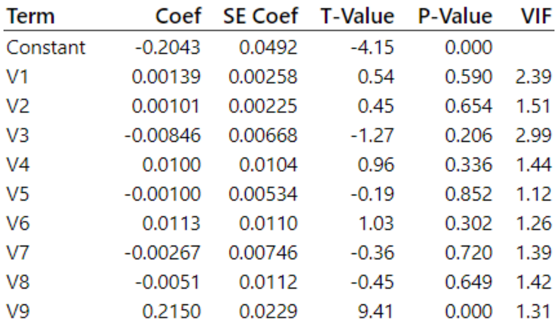
Maria Jakimiec and Anna Bartkowaik [22] made a multivariate regression based predictions using two datasets, one for the prediction algorithm and the other for which the prediction is preformed, to characterize sunspot groups, and the results have shown insignificant differences between both datasets which shows the quality of the predictions.

Despite the extensive findings of the respected previous works, they did not use feature selection techniques such as step wise regression or leave-one-out cross validation method, nor did they present a comparison that shows the difference between the predictors' accuracy and effectiveness in the dataset using the several regression methods mentioned in this paper.

V. Model design

A. Multi-Linear Regression

The Minitab software was used to perform Multi-linear regression, by using the Fit regression tool on the Holdout partitioned training dataset exported from Rstudio.



Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.2043	0.0492	-4.15	0.000	
V1	0.00139	0.00258	0.54	0.590	2.39
V2	0.00101	0.00225	0.45	0.654	1.51
V3	-0.00846	0.00668	-1.27	0.206	2.99
V4	0.0100	0.0104	0.96	0.336	1.44
V5	-0.00100	0.00534	-0.19	0.852	1.12
V6	0.0113	0.0110	1.03	0.302	1.26
V7	-0.00267	0.00746	-0.36	0.720	1.39
V8	-0.0051	0.0112	-0.45	0.649	1.42
V9	0.2150	0.0229	9.41	0.000	1.31

Figure 9

Using the information provided by the fit regression analysis shown in Figure 9, the P-values and the VIF can help us determine further the effectivness of the dataset attributes. The variance inflation factois a way to measure the aftermath of multicollinearity in a normal least squares regression analysis, between the predictors that is used to specify when two or more predictors are highly correlated. Because the VIF is less than 5, it's in good shape, therefore, the dataset does not have a Multicollinearity problem.

The p-value is the probability of getting test results a minimum of as extreme because the results truly discovered throughout the test, assuming that the null hypothesis is correct.
The P values for V9 (Area) is less than 0.05, making it statistically significant and effective to the dataset compared to the other attribues. Since the P is less than 0.05, we should reject the null hypothesis and therefore, we say that dataset is not normally distributed.

The Histrogram of the dependent variable in figure 10, it also shows that it is positively skewed, therefore it is not normally distributed.

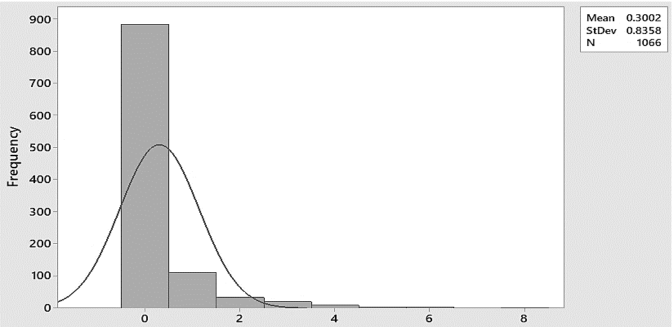


Figure 10

The MLR model testing using Holdout method using the R programming language, was done by summing the testing with the MLR coefficients. The mean absolute error was **1.017987**, under the variable name "newYLR".

B. Multilayer Perceptron

MLP is initially performed by converting both the independent and dependent attributes to be numeric. Holdout partitioning has been used in the process of converting the variables from named lists to matrices. The Multilayer perceptron was created using the R programming language library "monmlp.fit" with 5 hidden layers, using the training and testing matrices that were created.

The model created will show the compute model outputs using the "monmlp.predict" function in "monmlp" library. After creating the MLP model, the mean absolute error came out to be **0.0122019**, under the variable name newYMLP. The MAR is great given it has a low error rate.

VI. Results and Discussion

After evaluating the dataset and testing the attributes for their best features using the Cross Validation methods of K-fold and Leave-one-out, we found the best features of the dataset, that has been confirmed with the feature selection methods, giving the same results. The MAR for the MLR was also calculated using the Holdout method. Calculating the mean absolute error for the Multilayer perceptron and Multi linear regression is shown and proves that the model is accurate with a low error rate as shown in table 1.

Model		
	MAR	***Variance***
newYMLP	0.0122019	0.02108655
newYLR	1.017987	0.0005330106

Table 1

The outcomes of the feature rank are summarized in figure 11 below, showing the effectiveness of each attribute according to its importance. K-fold and Leave-one-out cross validation methods were used.

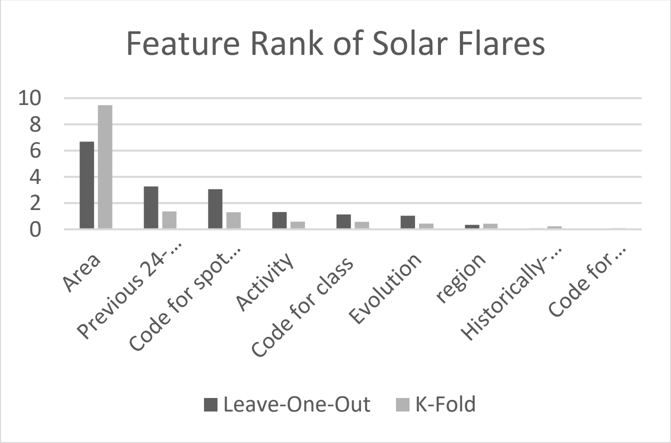


Figure 11

VII. Conclusion

In this paper, we analyzed and tested the solar flare dataset, using feature selection techniques such as stepwise regression, importance ranker using k-fold and leave-one-out cross validation methods, which all showed similar results that confirm that these tests are accurate. We tested the model by evaluating the dataset using Multi-linear regression, by looking at the P-values and the VIF that showed the effectivness of the dataset attributes. Multilayer perceptron model was created by using the holdout regression by partitioning the dataset into 80% training and 20% testing to model the testing dataset into an MLP model with 5 hidden layers. This showed the mean absolute error and variance of the model, deeming it accurate due to its low error rate.

References

[1] "Space Science Esa.int, 2020. [Online]. Available: https://www.esa.int/Science_Exploration/Space_Science/What_are_solar_flares. [Accessed: 01-May-2020].

[2] Will Kneton "How Multiple Linear Regression Works," *Investopedia*, 2020.[Online].Available:https://www.investopedia.com/terms/m/mlr.asp. [Accessed: 03-May-2020].

[3] "DataRobot Automated Machine Learning," DataRobot, 2019. [Online]. Available: https://www.datarobot.com/wiki/training-validation-holdout/. [Accessed: 01-May-2020].

[4] "Cross Validation," Cmu.edu, 2020. [Online]. Available: https://www.cs.cmu.edu/~schneide/tut5/node42.html. [Accessed: 01-May-2020].

[5] G. I. Webb et al., "Leave-One-Out Cross-Validation," Encyclopedia of Machine Learning, pp. 600–601, 2011..

[6] L. NCSS, "NCSS Statistical Software Stepwise Regression," 2016.

[7] Jason Brownlee PhD., "An Introduction to Feature Selection," Machine Learning Mastery, 05-Oct-2014. [Online]. Available: https://machinelearningmastery.com/an-introduction-to-feature-selection/. [Accessed: 01-May-2020].

[8] Raheel Shaikh, "Feature Selection Techniques in Machine Learning with Python," Medium, 28-Oct-2018. [Online]. Available: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e. [Accessed: 01-May-2020].

[9] "UCI Machine Learning Repository: Solar Flare Data Set," Uci.edu, 2020.[Online].Available:https://archive.ics.uci.edu/ml/datasets/Solar+Flare. [Accessed: 01-May-2020].

[10] Jinyan Li and Guozhu Dong and Kotagiri Ramamohanarao and Limsoon Wong. DeEPs: A New Instance-based Discovery and Classification System. Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases. 2001.

[11] Sally A. Goldman and Yan Zhou. Enhancing Supervised Learning with Unlabeled Data. ICML. 2000.

[12] Nir Friedman and Daphne Koller. Being Bayesian about Network Structure. UAI. 2000.

[13] Jinyan Li and Guozhu Dong and Kotagiri Ramamohanarao. Instance-Based Classification by Emerging Patterns. PKDD. 2000.

[14] Christophe G. Giraud-Carrier and Tony R. Martinez. An Integrated Framework for Learning and Reasoning. J. Artif. Intell. Res. (JAIR, 3. 1995.

[15] Nir Friedman and Daphne Koller (koller@cs. stanford. edu. A Bayesian Approach to Structure Discovery in Bayesian Networks. School of Computer Science & Engineering Hebrew University.

[16] C. Titus Brown and Harry W. Bullen and Sean P. Kelly and Robert K. Xiao and Steven G. Satterfield and John G. Hagedorn and Judith E. Devaney. Visualization and Data Mining in an 3D Immersive Environment: Summer Project 2003.

[17] Jason Brownlee, "Crash Course On Multi-Layer Perceptron Neural Networks," Machine Learning Mastery, 16-May-2016. [Online]. Available: https://machinelearningmastery.com/neural-networks-crash-course/. [Accessed: 03-May-2020].

[18] Nitin Kumar Kain, "Understanding of Multilayer perceptron (MLP)" Medium, 21-Nov-2018. [Online]. Available: https://medium.com/@AI_with_Kain/understanding-of-multilayer-perceptron-mlp-8f179c4a135f. [Accessed: 03-May-2020].

[19] Brian Beers, "What P-Value Tells Us," Investopedia, 2020. [Online]. Available: https://www.investopedia.com/terms/p/p-value.asp. [Accessed: 03-May-2020].

[20] T. Colak and R. Qahwaji, "Automated Solar Activity Prediction: A hybrid computer platform using machine learning and solar

imaging for automated prediction of solar flares," Space Weather, vol. 7, no. 6, p. n/a-n/a, Jun. 2009.

[21] L. E. Boucheron, A. Al-Ghraibah, and R. T. J. McAteer, "PREDICTION OF SOLAR FLARE SIZE AND TIME-TO-FLARE USING SUPPORT VECTOR MACHINE REGRESSION," The Astrophysical Journal, vol. 812, no. 1, p. 51, Oct. 2015.

[22] Maria Jakimiec, Anna Bartkowaik, "Distance-based Regression in prediction of solar flare activity", Harvard.edu, 2020. [Online]. Available: http://adsabs.harvard.edu/full/1994AcA....44..115J. [Accessed: 18-Jul-2020]

[23] A. B. Nassif, M. Azzeh, A. Idri, and A. Abran, "Software development effort estimation using regression fuzzy models," Comput. Intell. Neurosci., vol. 2019, 2019.

[24] A. B. Nassif, "Software Size and Effort Estimation from Use Case Diagrams Using Regression and Soft Computing Models," University of Western Ontario, 2012.

[25] A. B. Nassif, L. F. Capretz, and D. Ho, "Estimating software effort using an ANN model based on use case points," in Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012, 2012, vol. 2, pp. 42–47

Authors Contributions: Mohamad Rabie wrote the paper. Ali Bou Nassif and Maha AlaaEddin revised the experiments and the whole paper