**Predicting Solar Flare Events with Linear Regression**

**Project One**

**COSC 480ML FA 2017**

John DeMey jtdemey@smcm.edu

Adam Van Stone ajvanstone@smcm.edu

Deonarine Sahdeo dsahdeo@smcm.edu

**Abstract**

*This paper presents our investigation/implementation of linear regression for predicting the number of solar flares that occur every night based on input variables. We hypothesized that we could achieve solid results with only categorical variables. Although we were unable to obtain those results through linear regression, we learned how to transform the data into coding that an algorithm could use to predict an output.*

**INTRODUCTION**

This paper explains the properties of the solar flare data set from the UCI Machine Learning Repository. The dataset contains variables that explain different classifications of recorded solar flares including, but not limited to spot distribution, area, historical complexity, and evolution. We will be using linear regression to predict the number of solar flares that occur on any given day with the standard linear equation : y =mx +b, where y is the target, m is the coefficient of x and b is the intercept term.

In Section II we provide details of our exploration of the data set. Section III outlines our data preparation and experimentation methods and Section IV provides the results and discussion of our experiments. In Section V we provide reflection on our project to consider limitations and future work. Finally, Section VI concludes the paper.

**II. EXPLORATORY ANALYSIS**

1. *Data Set Description*

The data set used for this study came from the UCI Machine Learning Repository and was uploaded on March 1st, 1989. The dataset contains ten independent variables and three dependent variables, of C-class, M-class, and X-class flares, which are common, moderate and severe flares respectively. The data types for these values are explained more in Table 1. The dataset contains 1389 instances.

No values were recorded missing from this dataset.

**Table 1: Data Types**

|  |  |
| --- | --- |
| Variable Name | Data Type |
| Zurich Classification | Multi-valued Discrete |
| Largest Spot Size | Multi-valued Discrete |
| Spot Distribution Code | Multi-valued Discrete |
| Activity | Binary |
| Evolution | Multi-valued Discrete |
| Previous 24-hour flare activity code | Multi-valued Discrete |
| Historically-complex | Binary |
| Historically-complex on current pass | Binary |
| Area | Binary |
| Area of the Largest Spot | Binary |
| Number of Flares | Continuous |

*B. Data Set Summary Statistics*

This section provides summaries for each variable in the Solar Flare Data Set.

**Table 2: Proportions for Zurich Class (n = 1066)**

|  |  |  |
| --- | --- | --- |
| Zurich Class | Frequency | Proportion |
| A | 0 | 0% |
| B | 147 | 13.8% |
| C | 211 | 19.8% |
| D | 239 | 22.4% |
| E | 95 | 8.9% |
| F | 43 | 4.0% |
| H | 331 | 31.1% |

**Table 3: Proportions for Spot Size (n = 1066)**

|  |  |  |
| --- | --- | --- |
| Spot Size Code | Frequency | Proportion |
| X | 145 | 13.6% |
| R | 218 | 20.5% |
| S | 414 | 38.8% |
| A | 216 | 20.3% |
| H | 27 | 2.5% |
| K | 46 | 4.3% |

**Table 4: Proportions for Spot Distribution Code (n = 1066)**

|  |  |  |
| --- | --- | --- |
| Spot Distribution Code | Frequency | Proportion |
| X | 331 | 31.1% |
| O | 477 | 44.7% |
| I | 223 | 20.9% |
| C | 35 | 3.3% |

**Table 5: Proportions for Activity (n = 1066)**

|  |  |  |
| --- | --- | --- |
| Activity | Frequncy | Proportion |
| 0 | 902 | 84.6% |
| 1 | 164 | 15.4% |

**Table 6: Proportions for Evolution (n = 1066)**

|  |  |  |
| --- | --- | --- |
| Evolution | Frequency | Proportion |
| 1 | 77 | 7.2% |
| 2 | 484 | 45.4% |
| 3 | 505 | 47.3% |

**Table 7: Proportions for previous 24-hour flare activity code (n = 1066)**

|  |  |  |
| --- | --- | --- |
| Previous 24-hour activity code | Frequency | Proportion |
| 1 | 1028 | 96.4% |
| 2 | 13 | 1.2% |
| 3 | 25 | 2.5% |

**Table 8: Proportions for historically-complex (n = 1066)**

|  |  |  |
| --- | --- | --- |
| Historically Complex | Frequency | Proportion |
| 0 | 635 | 59..6% |
| 1 | 431 | 40.4% |

**Table 9: Proportions for historically complex on current pass (n = 1066)**

|  |  |  |
| --- | --- | --- |
| Historical complex current pass | Frequency | Proportion |
| 0 | 133 | 12.5% |
| 1 | 933 | 87.5% |

**Table 10: Proportions for area (n = 1066)**

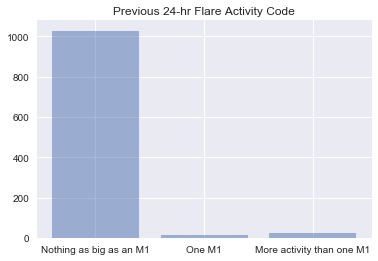
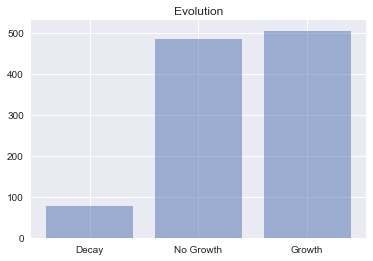
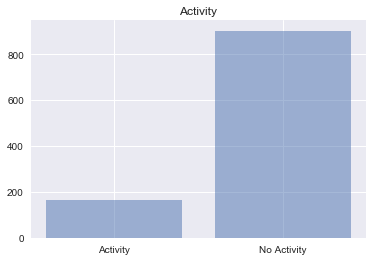
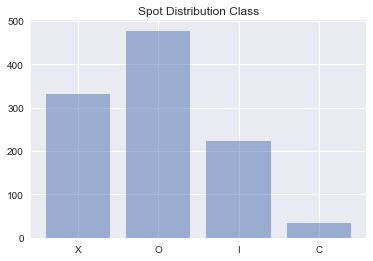
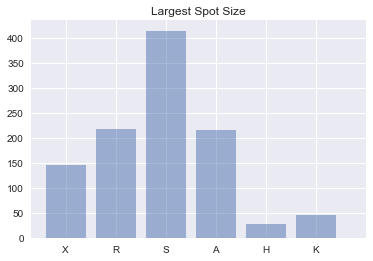
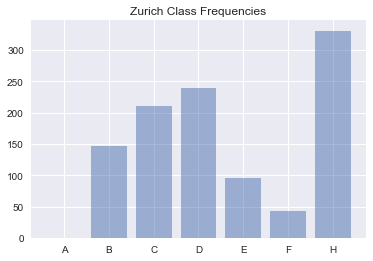
|  |  |  |
| --- | --- | --- |
| Area | Frequency | Proportion |
| 0 | 1039 | 97.5% |
| 1 | 27 | 2.5% |

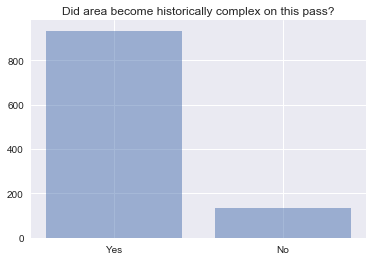
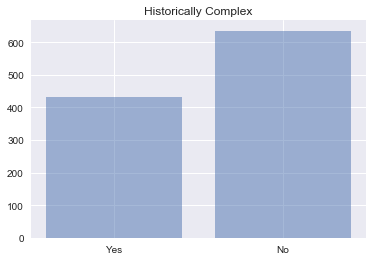
**Table 11: Proportions for area of the largest spot (n = 1066)**

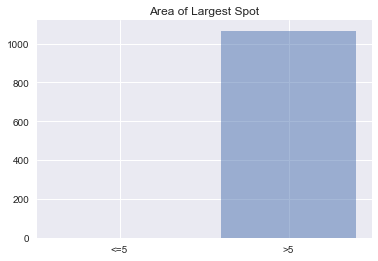
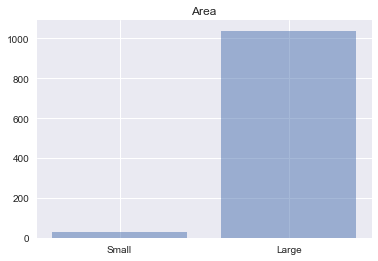
|  |  |  |
| --- | --- | --- |
| Area of largest spot | Frequency | Proportion |
| 0 | 1066 | 100.0% |
| 1 | 0 | 0.0% |

*C. Data Set Graphical Investigation*

This section provides specific visualizations of this data set generated during exploratory analysis. Since there are no continuous variables all of the variables will be shown using bar charts.





**

**Figure 1: Bar graphs for each categorical variable**

*D. Additional Exploration*

None of our variables are continuous, so there is nothing to place here

*E. Exploratory Findings*

**III. METHODS**

1. *Data Preparation*

The solar flare data set was completely intact, but required some modifications for linear regression. No instances of the data had to be dropped, but since all of the variables were categorical, it was much harder to create the equation to compute the linear regression. While we were familiar with most of these libraries,we were not familiar with all of the features that we needed to use. For example, we knew how to transform categorical variables into data that the regression could read and use, but had no idea on how to implement said algorithm. We had to test many different versions of the linear regression algorithm before we could find one that worked with our data set.

1. *Experimental Design*
2. *Tools Used*

Python v3.6.1 running Anaconda v 4.4.0 for Apple Macintosh and Microsoft Windows computer was used for all analysis and implementation. In addition to base Python, the following libraries were also implemented: Pandas 0.20.1, Numpy 1.12.1, Matplotlib 1.5.3, Seaborn 0.7.1, SKLearn 0.18.1, Patsy 0.4.1

**IV. RESULTS**

1. *Classification Measures*

*\*

1. *Discussion of Results*

**V. REFLECTION**

1. *Problems Encountered*

One problem that we encountered was the fact that our data set only used categorical variables, thereby making it more difficult to implement linear regression. The inability to use a generic version of linear regression also made it harder to implement since we had to not only design the equation from scratch,, but also account for the fact that there were no continuous variables.

1. *Limitations of Implementation*

Since we could not use SKLearn to implement the actual linear regression equation, creating an equation of linear regression that works with only categorical variables was difficult. Using one-hot encoding, we were able to transform all of the variables into binary information. Unfortunately, we had a difficult time getting the program to take in these variables to perform a linear regression, and had very little experience

1. *Improvement/Future Work*

No data was determined missing from our data set, however it would preferably be better to use data that has continuous independent variables. In future experiments, we may consider using data sets that have continuous and categorical variables, or only continuous variables.

**VI. CONCLUSION**

To summarize, the application of linear regression to predict the number of solar flares from solely categorical variables was not fully achieved. This is mostly due to categorical variables being harder to implement into a linear regression..

**REFERENCES**

**APPENDIX**