Initial Draft for CMPUT 466 Mini-Project

Jinzhu Li
1461696

Introduction:

One major environment concern nowadays is the occurrence of the forest fires (wildfires), which is a threat to the forest preservation, economical and ecological environment, and human activities. Such phenomenon is caused by multiple reasons such as lightings and human negligence, and each year millions of forest hectares are destroyed worldwide.

Therefore, this study is going to predict the burned area of forest fires using meteorological data. Three different machine learning algorithms, e.g. Neural Network (NN), Support Vector Machine (SVM), and Random Forest (RF) were tested on the real-world data collected from the northeast region of Portugal.

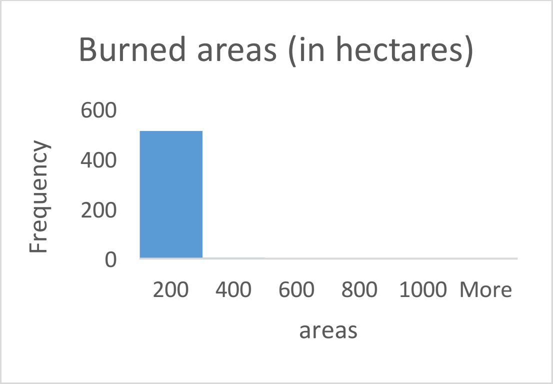
Finally, the result is going to answer the question that which algorithm works the best, that has the smallest error.

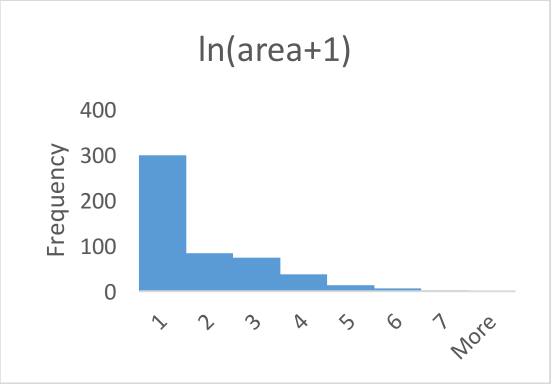
Description of the Dataset:

The data is the real-world data, collected from the Montesinho natural park, from the northeast region of Portugal, with the aim of predicting the burned area (or size) of the forest fires. The data used in the experiment were collected from January 2000 to December 2003.

There are total of 517 instances and 13 attributes: spatial location X, spatial location Y, Month of the year, Day of the week, Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), temperature, relative humidity, wind speed, rain, total burned area. The target value is the total burned areas of the forest fires. All these variables are numerical, and Mouth and Day have been converted to numbers accordingly.

The burned area is shown in the figure below, denoting a positive skew, with the majority of fires presenting in a small size. To reduce the skewness and improve the symmetry, the logarithm function $y = \ln(x+1)$, which is a common transformation that tends to improve regression results for right-skewed targets was applied to area attribute. The final transformed variable will be the output of this work.





Learning Methods:

Three learning methods are chosen for this project: Neural Network (NN), Support Vector Machines (SVM), and Random Forest (RF).

- Neural Network (NN): This study will consider multilayer perceptron with one hidden layer of 100 hidden nodes and the rectified linear unit function (f(x)=max(0,x)) as the activation function and one output node with a linear function.
- Support Vector Machines (SVM): In SVM regression, the input is transformed into a high dimensional feature space, by using a nonlinear mapping. The SVM finds the best linear separating hyperplane in the feature space. SVM presents theoretical advantages than NN, such as the absence of local minima in the model optimization phase, that's why it was chosen.
- Random Forest (RF): Random Forest is a method based on tree search. It is a resemble of T unpruned Decision Tree (DT), using random feature selections from bootstrap training samples. The RF predictor is built by averaging the outputs of the T tree. This method was chosen because it solves the problem that multiple regression can only learn linear mappings, and it has a substantial improvement over a single DT.

Design of experiment:

All the experiment reported in this study were conducted using the scikit-learn, an open source library for Python that facilitates data mining and data analysis. Within the total of 517 entries, 300 were used as training set, and 217 were used for predictions.

Before fitting the model, some preprocessing was required by NN, SVM and RF models. First, month and day were transformed into numerical attributes. Also, for NN and SVM methods, all attributes were standardized to a zero mean and one standard deviation. Next, the regression model was fitted. The default parameters were adopted for RF (e.g. T=500), the NN was adjusted using E=100 epochs until the tolerance less than 10e-4, and the Sequential Minimal Optimization was used for SVM. After fitting the models, the outputs were post-processed using the inverse of the algorithm transform. In few cases, this transformation may lead to negative numbers, and such negative outputs were set to zero. All the methods are using the squared loss as a cost function.

A paired t-test will be used for the statistical significance of the result.

Final Results:

# tables and graphs remain to be added

Conclusion:

# To be filled