RISE RESEARCH INNOVATION SCHOLARSHIP ENTREPRENEURSHIP

Abstract ID#: 3350
Category: Computer and Information Sciences
Undergraduate/Graduate: Undergraduate

Northeastern University

Comparison of Machine Learning Algorithms to Predict the Occurrence of Forest Fires

Tejas Sathyamurthi

OPPORTUNITY

Wildfires are catastrophic phenomena that occur quite frequently in the West Coast of the United States. In a study in San Diego in 2003, the total economic costs of wildfire were estimated to be $2.45 billion with around 376,000 acres burned. Due to significant loss of life, property and environmental damage occurring every year the problem needs to be approached early and predicted beforehand.

OBJECTIVE

➢This research focuses on understanding the effects of the physical conditions such as temperature, wind, humidity and relative humidity on predicting the occurrence of a fire in a given area. Understanding this could be the first step towards preventing forest fires.
➢This research also aims to study the feasibility of predicting forest fires using the four machine learning methods namely
- ➢ Logistic regression
- ➢Decision Trees and Random Forests
- ➢Support Vector Machines
- ➢Neural Networks

DATA COLLECTION

➢The training dataset for this project is taken from Portugal fire dataset
➢ Test dataset was self-created from the forest fires in the west coast of United States since 2000

Handling Missing Values – Dataset consists of certain observations with missing features. This could be traced back to the logging of data. During the occurrence of certain fires, due to the absence of the right measuring equipment, certain features could have gone unlogged. However, the presence of missing values impacts model building and can cause unforeseen errors while interpreting the results. Therefore, it is very important to handle missing values effectively and strategically. Three common ways of handling missing values are as follows

➢ Remove entire row in case of a missing feature
➢ Impute the missing feature with either the mean or median
➢ Create a new categorical variable for the missing values in a given column

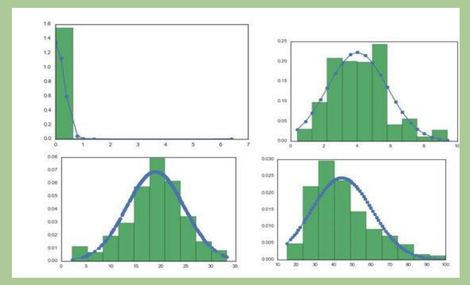
For this project, the missing values were too high to go with removing the entire row and too low to go with creating a new category. The missing values for this dataset were imputed to the mean.

$$\sum_{i=1}^{i=n} Xi/\text{n}$$

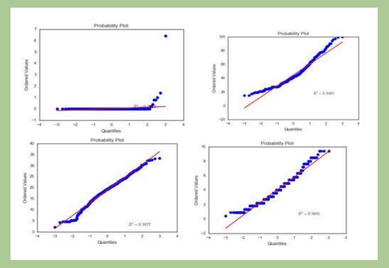
Xi = Features
i = 1, 2, 3, 4 … 517

Handling Normality- Many statistical tests rely on the assumption that data comes from a normal distribution. This implies that the data is distributed around the mean. That is, 50% of the data is higher and 50% of the data is lower than the mean.

However, most real-life datasets are not normally distributed. This can cause errors while interpreting results. The first step in the data preparation process is to check for normality using Bell Curve Test and Q-Q Plot Test.



Bell Curve Test – This involves plotting the histogram of the data and if it closely resembles a bell curve, then the distribution is likely to be normal. The graph above represent features, rain, wind, temperature and relative humidity respectively from top left to bottom right



Q-Q Plot Test -The non- normality of the feature rain is confirmed using Q-Q Plot Test (figure above)

APPROACH

METHOD 1 – LOGISTIC REGRESSION

Logistic regression is a type of regression model where the dependent variable is categorical. Logistic regression estimates the log odds of an event occurring by estimating a multiple linear regression function as below

$$log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

β – Coefficients of features, Xi – Features, p - Probability

METHOD 2 – DECISION TREES AND RANDOM FORESTS

Decision tree is a tree like model that contains root nodes, branches and leaf nodes. Based on the feature with the most classifying power, the node is split into two and that node is split further till a decision can be made about the observation.

Entropy is the measure that is used to determine where the leaf node will be split and lies between 0 and 1. The direction of maximum increase in entropy is called information gain and that is the feature that is used for the split.

Entropy is given by:

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

E – Entropy, c – Number of classes
p – Probability of occurrence of class *i*

METHOD 3 – SUPPORT VECTOR MACHINES (SVM)

Support vector machines are a discriminative classifier that uses a hyper plane to separate labeled classes. A hyper-plane with the largest minimum distance from the training data points is deemed optimal. Therefore, the optimal separating hyper-plane is the one that maximizes the margin of the data.

Hyper-plane can be represented below.

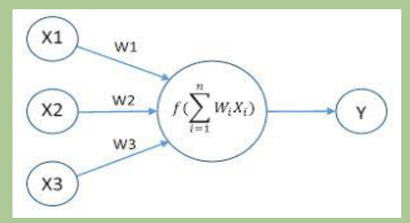
$$f(x) = \beta_0 + \beta^t x$$

$$distance_{support\ \ vectors} = \frac{|\beta_0 + \beta^t x|}{||\beta||} = \frac{1}{||\beta||}$$

$$\beta - coefficients\ of\ hyperplane$$

METHOD 4 – NEURAL NETWORKS

Neural Networks are a machine learning framework that mimics the learning process of biological neurons. Neurons receives inputs and based on these inputs produce a given output. A perceptron is a computer model consists of one or more inputs, an activation function, a bias and a single output. The perceptron multiples the inputs it receives with a certain weight, adds it to the bias and passes it through the activation function before giving out a single output. After receiving the output, the accuracy of the classification is calculated and the weights are re-trained to improve classification. This happens till a certain error rate is achieved or the end of the acceptable number of iterations has been reached.



X – Input Data , W – weights associated with each input
WiXi – Activation Function , Y = output

DATA/RESULTS

The dataset was divided into 10 folds in a technique called cross validation. The models were trained on 9 folds and tested on the last fold. This was performed for all the folds. The mean accuracy from all the folds has been plotted in the table below:

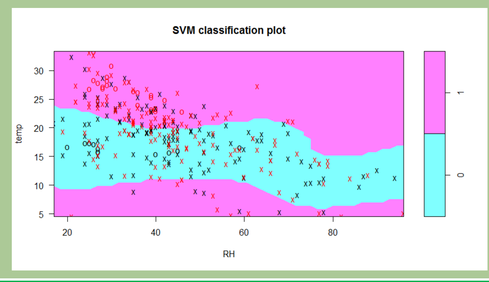
METHOD	PRECISION SCORE	RECALL SCORE
Logistic Regression	0.51	0.66
Decision Trees	0.61	0.73
Support Vector Machines	0.64	0.76
Neural Networks	0.58	0.64

As can be seen from the above table, SVM performs the best as compared to the other models. The table below shows the sum of False Positive and False negatives for the test data

Total Test Data Points	Sum of False Positive and False Negative	Percent (%)
217	65	30

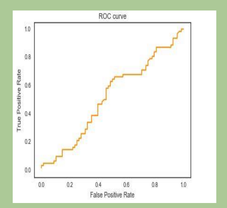
REASONS FOR SUPERIOR PERFORMANCE OF SVM

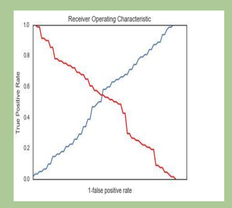
➢The relationship between the features and the response variable in this dataset is non-linear. Thus, logistic regression fails to separate the classes effectively
➢ Although trees tend to perform very closely to SVM, trees have the tendency to draw too complex decision boundaries thus leading to over-fitting. Thus, may not be suitable for predictions
➢ SVM's on the other hand, due to the presence of the regularization parameter can be controlled for over-fitting and due to its kernel, can be engineered well to suit the application. In this case, the radial kernel best captures the decision boundary



RECEIVER OPERATOR CHARACTERISTIC CURVES

Receiver Operator Characteristic curves is optimally used to choose the cut-off point in the SVM model. The model assumes a cut off probability of 0.5. This implies, anything above 0.5 would be classified as 1 and anything below it would be classified as 0.





From the above graph, it can be seen that the threshold is 0.54.

IMPACT

➢With this model, fires could potentially be predicted before-hand and can save lives, property and habitats, and it is beneficial on a global scale.
➢Emergency services can be alerted ahead of time, and this project opens multiple possibilities for proper evacuation methods when risk of a fire is high.
➢An application that could be run on smart phones to alert users and warn users of potential fires can be developed.
➢ Using machine learning methods showed that forest fires can be predicted with some degree of accuracy, and SVM performs the best in the prediction of forest fires. The most important factors/highest predictive powers are Temperature and Relative Humidity. With fires, the risk of a false negative is much higher than that of a false positive. It would be dangerous to classify an incident as 'Non-Threatening', when it is threatening. Therefore, the focus was on reducing false negatives where, SVM performs well with a score of 76%.