Machine Learning Engineer Nanodegree

Capstone Project Proposal

Insights Into Forest Fires

Vasantha Charitha Yadav

February 6th,2019

Domain Background:

History:

One major environment concern is the occurrence of forest fires (also called wildfires),which affect forest preservation, create economic and ecological damage and cause human suffering is due to multiple causes (e.g. human negligence and lightnings) and despite an increasing of state expenses to control this disaster. Fast detection is a key element for a successful firefighting. Since traditional human surveillance is expensive and affected by subjective factors, there has been an emphasis to develop automatic solutions. Weather conditions, such as temperature and air humidity, are known to affect fire occurrence.

In the past, meteorological data has been incorporated into numerical indices, which are used for prevention (e.g. warning the public of a fire danger) and to support fire management decisions.[1]

My main motivation to take this project is Forests cover 31 percent of the world's land surface, just over 4 billion hectares.As forest are main resource of environment and Forest fires are big problems in countries like Australia and America. I wanted to solve it by using some of the machine learning techniques.

Problem statement:

The main aim of my project is predicting the burned area due to small fires based on the features like Wind,Rain,Relative humidity etc., provided by my dataset collected from the Department of Information system,University of Minho,Portugal.This can be done by some machine learning techniques like regression algorithms.The best suitable regressor for my problem is predicted at last.

Datasets and Input:

This dataset is public available for research. The details are described in [Cortez and Morais, 2007].The data can be used to test regression (difficult task), feature selection or outlier detection methods.

Dataset: https://archive.ics.uci.edu/ml/datasets/forest+fires

Dataset information :

The attributes in the dataset include:

X - x-axis spatial coordinate within the Montesinho park map: 1 to 9

Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9

month - month of the year: "jan" to "dec"

day - day of the week: "mon" to "sun"

FFMC - FFMC index from the FWI system: 18.7 to 96.20

DMC - DMC index from the FWI system: 1.1 to 291.3

DC - DC index from the FWI system: 7.9 to 860.6

ISI - ISI index from the FWI system: 0.0 to 56.10

temp - temperature in Celsius degrees: 2.2 to 33.30

RH - relative humidity in %: 15.0 to 100

wind - wind speed in km/h: 0.40 to 9.40

rain - outside rain in mm/m2 : 0.0 to 6.4

area - the burned area of the forest (in ha): 0.00 to 1090.84

Solution statement:

I will be reaching every step involved in predicting the best regressor to obtain a solution.The algorithms namely linear regression,decision tree,SVM etc., are used for predicting my answer.I will explore the dataset with opencv and matplotlib libraries for better visualisation.

Benchmark model:

As benchmark is important to compare my results with some metric and decide which model suits the best . Here my bench mark is mean square error caluculated using the linear regression model and the bench mark is 1920.Then caluclating the mean squared error and variance for the selected models(SVM,Linear,Decision tree)then comparing the mean squared error and variance for the models which model has the least mean squared error and high variance is the best model for predicting the Burned Area

Evaluation metrics:

Here i am using the metrics for my models is correlation coefficient,mean squared error,variance

Here my models are regression models so i am taking the evalution metrics as the RMSE,variance

The square root of the mean/average of the square of all of the error.

The use of RMSE is very common and it makes an excellent general purpose error metric for numerical predictions.

Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors.

$$RMSE=sqrt(1n\sum_{i=1n}((y_i-y\hat{}_i)_2))$$

After calucating the RMSE value for all theselected models then comparing the RMSE value with each other.The model which got the least RMSE value that model is selected as the model for my project[3]

Project Design:

The project is started by following steps:

The first task is to read the dataset and perform visualisations to get some insights .After reading the data cleaning it i.e., removing unwanted data or replacing null values with some constant values or removing duplicates.

Next finding the corrrelation with each feature with burned area attribute.

After data exploration i want to split the data into training and testing data sets.Model selection step is performed in finding the metric values mentioned above .

Normalisation of data is done and learning models on the normalizied data.

Then converting the target value to binary classes and finding accuracy and visualising the results

Finally the one with least mean squared error is predicted as my best model.

REFERENCE LINK:

[1] www3.dsi.uminho.pt/pcortez/fires.pdf

[2]www.insightsonindia.com

[3]https://datascience.stackexchange.com/questions/15512/calculating-rmse-and-r-squared-from-the-confusion-matrix