Capstone Proposal

Machine Learning Engineer Nanodegree

Title: Air Quality

M.Ravindra Naik

February 26th, 2019

Proposal

Domain Background:

In early 2011, officials reported that pollution in Italy was reaching crisis levels. What's particularly troublesome is particle pollution that pervades Italy, and accounts for breathing and heart problems, causing a whopping 9% of deaths of Italians over the age of 30. New report finds that air pollution is the single biggest environmental health risk in Europe, causing hundreds of thousands of premature deaths. Particulate matter, ozone, nitrogen dioxide. Europe's air quality is significantly threatened by these pollutants, mostly in urban centres

Air quality is a significant concern for both healthy population and people suffering for different pathologies. Long or even short term exposure to significant pollution levels have been associated with the development or worsening of multiple pathologies ranging from Asthma to Lung Cancer . Air quality patterns may significantly vary in space and time due to complex fluid dynamic effects occurring in the city landscape or to the hourly, daily and seasonal variation of human activities. However, most of the national states rely on the operation of networks of certified air quality monitoring stations in order to detect and monitor air quality in cities. Unfortunately, the average low spatial

density of such networks do not permit to achieve the required resolution.

Reference link:
https://www.researchgate.net/publication/319338229_Cooperative_Air_Quality_Sensing_with_Crowdfunded_Mobile_Chemical_Multisensor_Devices

Problem Statement:

To check the quality of air using 'Air Quality Chemical Multisensor Device' by finding the R^2 score and coefficient of regression using different regression models and the best model is selected to evaluate the Air Quality.

Datasets and Inputs:

Dataset Link: https://archive.ics.uci.edu/ml/machine-learning-databases/00360/

In this project I have used 15 attributes and around 9350 trained and test data to evaluate the R^2 score. And this coefficient is found out by using different regression models

Data Set Information:

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth

hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyser. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value. This dataset can be used exclusively for research purposes. Commercial purposes are fully excluded.

Features Information:

1.Date - DD/MM/YY

2.Time - HH.MM.SS

3.CO(GT) - True hourly averaged concentration CO in mg/m^3

4.PT08.S1(CO) - PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)

5. NMHC(GT) - True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 6. C6H6(GT) - True hourly averaged Benzene concentration in microg/m^3

7. PT08.S2(NMHC) - PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)

8. NOx(GT) - True hourly averaged NOx concentration in ppb

9. PT08.S3(NOx) - PT08.S3 (tungsten oxide) hourly averaged sensor response

10. NO2(GT) - True hourly averaged NO2 concentration in microg/m^3

11. PT08.S4(NO2)- PT08.S4 (tungsten oxide) hourly averaged sensor response

12. PT08.S5(O3) - PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)

13.T - Temperature in Â°C

14.RH - Relative Humidity (%)

15.AH - Absolute Humidity

Solution Statement:

In this project, I am trying to predict the quality of air using regression score for a particular area. This can be achieved by evaluating all the features mentioned above and finding R^2 score using regression techniques like Linear regression, Lasso Regression, Decision Tree. I will explore the data set with matplotlib.py, seaborn libraries to plot. Visualization helps to understand the model more clearly.

Benchmark Model:

Here we compare the final model with the remaining models to see if it got better or same or worse. The R^2 score is compared among the models and the best model is selected. I think Linear Regression model can be set as the benchmark model and I'm sure that the final solution would outperform the Benchmark model.

Evaluation Metrics:

R^2 Score:

R-squared is a statistical measure that's used to assess the goodness of fit of our regression model. In R-squared we have a baseline model

which is the worst model. This baseline model doesn't make use of any independent variables to predict the value of dependent variable Y. Instead it uses the mean of the observed responses of dependent variable Y and always predicts this mean as the value of Y.

R-squared is given by Where SSE is the sum of squared errors of our regression model

And SST is the sum of squared errors of our baseline model.

Project Design:

Pre-processing: It is the first step to read the dataset and clean the data i.e. removing unwanted data or identifying null values. If any null values exist, we replace them with constant values or removing duplicates.

Exploration: Visualizing the dataset, detect outliers, replacing a missing value and cleaning the dataset, splitting training dataset into training and testing sets and checks for any correlation among the features using heatmap.

Prediction: Here we predict the quality of air by finding R^2 score using different Regression models

Finally, I declare that the model with the highest R^2 score on both training and testing datasets will be concluded as the best model for evaluating the Quality of Air.