Machine learning engineer nanodegree

Capstone proposal

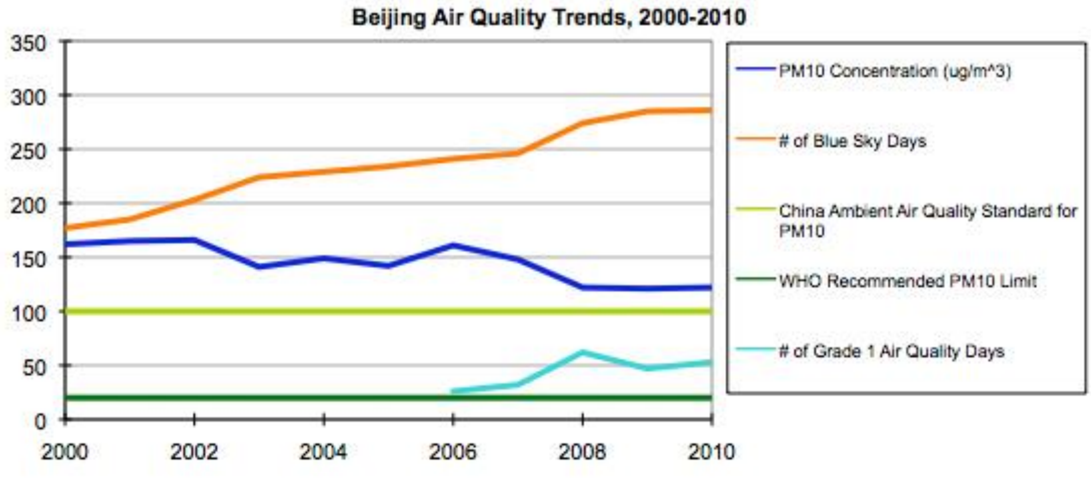
Time series analysis for Air Quality data

Ashish Kumar

February 18, 2020

Proposal

- **Domain Background**
  In recent years Air Quality has become a major health concern. The increase in amount of particulate matters like Pm-2.5, Pm-10, SO2, NO2 causes severe damage to lungs causing respiratory diseases and cancers. Peoples who suffer from seasonal allergies or have a weak immune system are more prone to the degradation in air quality. A time series analysis for the air quality can determine the increased risks associated with time so that necessary steps can be taken in advance like the availability of air masks, air purifiers etc.



The above picture**[2]** shows that the air quality trends of Beijing's air quality from 2000 to 2010 does not meet China's own air quality standard, and is six times worse than the recommended particulate matter target set by the WHO.

- **Problem statement**
  The problem is to analyse the concentrations of various pollutants in air like PM2.5, PM10, SO2, NO2 during the days, months, years and predict their concentrations in the upcoming years, thus forecasting the air quality index using time-series analysis. One approach to the problem is to use statistical models like Seasonal ARIMA. Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.

- ## Datasets and Inputs

  The dataset**[1]** was provided by Manu Siddhartha on Kaggle website. It contains the measurement of six major air pollutants PM2.5, PM10, SO2, NO2, CO and O3 at multiple sites in Beijing. The data set includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites. The air-quality data are from the Beijing Municipal Environmental Monitoring Centre. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017 containing enough information to analyse air quality and predict future pollution levels. I will be working on any one of the given sites.

- ## Solution statement:

  The solution to the above problem can be solved using Time series analysis. The first step would be data pre-processing and creating helper functions for generating time-series. As the dataset contains hourly data from year 2013 to 2017, various long and short time series can be made for analysing pollutants levels during daytime, during months and even during years. The time series can then be plotted using matplotlib library to analyse hourly and seasonal variations of pollutants. For prediction of air quality I will be using statistical methods like Seasonal ARIMA, Holt-Winters and FB Prophet. Later based on requirement, Long Short term memory or LSTM recurrent neural network can be trained to predict air quality. As an additional step, AQI Index can also be calculated from the predicted concentrations of different pollutants.

- ## Benchmark model:

  For this project, time series analysis of air quality, the benchmark model would be the Seasonal ARIMA or SARIMA. I will try to beat its performance with the LSTM and other statistical models like Holt-Winter and FB Prophet.

- ## Evaluation Matrix:

  The Time Series model can be evaluated based on various measures. As per the sites**[3] & [4]** the Time Series Models can be best evaluated using the following measures:

  1. **Mean Absolute Error(MAE)**

     The mean absolute error, or MAE, is calculated as the average of the forecast error values, where all of the forecast values are forced to be positive.

  2. **Mean Square Error(MSE)**

     The mean squared error, or MSE, is calculated as the average of the squared forecast error values. Squaring the forecast error values forces them to be positive; it also has the effect of putting more weight on large errors

3. **Root Mean Square Error(RMSE)**
   Root Mean Square Error (RMSE)**[4]** is the standard deviation of the prediction errors. RMSE is a measure of how spread out these residuals are. In other words, it tells us how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

- ## Project Design:

  The workflow for the project would be in the given order:

  1. Data loading
  2. Data cleaning and pre-processing
     a) Removing unwanted features
     b) Handling null values by replacing with mean, median etc. or by simply dropping.
     c) Handling Outliers
     d) Renaming and merging features based on requirement.
     e) Indexing dataframe.
  3. Visualising data for different pollutants hourly, monthly and yearly
  4. Defining necessary helper functions for Time series analysis.
  5. Creating training and test sets for time series.
  6. Finding parameters in case of SARIMA model based on lowest AIC value.
  7. Applying statistical models(SARIMA, Holt-Winters and FB Prophet), training model in case of LSTM
  8. Fitting and evaluating models.
  9. Predictions based on best model.
  10. (Optional) Calculating Air Quality Index based on predicted level of pollutants.

- ## References:

  1. https://www.kaggle.com/sid321axn/beijing-multisite-airquality-data-set
  2. http://www.livefrombeijing.com/2011/01/summary-of-beijings-2010-air-quality/
  3. https://machinelearningmastery.com/time-series-forecasting-performance-measures-with-python/
  4. https://www.statisticshowto.datasciencecentral.com/rmse/