National College of Ireland

Forecasting of Air Pollution in United Kingdom Using Deep Learning and Time series methods

MSc Research Project
Programme Name

Yash Vijaywargiya
Student ID: X18136842

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Yash Vijaywargiya

Student ID: X18136842

Programme: Data Analytics **Year:** 2019-2020

Module: Research Project

Supervisor: Dr Catherine Mulwa

Submission Due Date: 12 December 2019

Project Title: Forecasting of Air Pollution in United Kingdom Using deep learning and Time series methods

Word Count: 7915 **Page Count** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Yash Vijaywargiya

Date: 12-12-19

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Forecasting of Air Pollution in United Kingdom Using Deep Learning and Time series methods

Yash Vijaywargiya
X18136842

Abstract

Air pollution is found when particles in air exceed a particular concentration limit which makes it harmful for the ecosystem and human life. Air pollution is a huge factor for the cause of death because of its short term and long-term impact on health of the people. Many researchers have conducted a lot of analysis for forecasting the air quality. World Health Organisation (2019) has stated that 83 % areas of United Kingdom was found exceeding the air pollution level in 2019. This project evaluates the forecasting of No2 pollutant in air by comparing deep learning and time series models. However, In United Kingdom very less research has been performed for the air pollution. This project includes implementation of ARIMA, SARIMA, TBAT and neural networks. From the results it was clear that neural network with stacked LSTM has outperformed every other model. The results of the reviewed literature on air pollution in Europe are also presented.

1 Introduction

Air quality levels are growing dangerously in the world from last few years. World health organization stated that 9 out of 10 human beings are breathing polluted air. 7 million human beings have died in just one year because of pollution, out of which one third of deaths are anticipated from heart diseases, lung cancer & strokes. Air pollution in United Kingdom is been regarded as the serious health issue and an environmental threat. The top cities in United Kingdom like London, Scunthorpe, Gibraltar has spotted the pollution level to be above the limits assigned by the legal authority. Maheswaran et al., (2005) World health organisation int. (2019) marked 36 million deaths in a year are been attributed in United Kingdom from air pollution diseases like asthma, lung and heart disease. The air pollution is been classified by the pollutant such as NO2 Grineski et al. (2007 pp.535–554). From 90s, in the top cities the pollutant NO2is been marked as the serious issue and it's been increased nowadays Chan and Yao, (2008, pp. 1-42); NO2 come from the vehicle combustion and the power plants. NO2 pollutant causes asthma and also reason behind lung problem specially for the age of children below 20. No2 can also causes heart disease mortality Michael B, (2002). By Yanosky and Schwartz (2008 pp. 1593–1602), it has marked NO2 pollutant to be higher in United kingdom in high SES communities.

In this project, we have contemplated United Kingdom (country in Europe) for the research domain. The data is been downloaded from the Europe website which is European Environment Agency (EEA). This study involves the comparison and contrast of the models like TBATS, LSTM, ARIMA, stacked LSTM. The novelty behind the research is the model TBATS is never used for forecasting the pollutant no2 in United kingdom. The project is been divided in various section for better understanding and presentation. Firstly, it starts with the motivation and dataset where the reason behind taking this topic is covered. Second comes the section 2 which is Literature review which provides all the previous work related with this topic and the gaps we seen from the papers and what uniqueness will be seen in the paper. Literature is also divided on machine learning, the models used and for the countries in Europe. Section 3 is the methodology been written for the project. Section 4

which is the major part which covers the Implementation of the models used. The evaluation and the results are covered in section 5. The project ends with the conclusion and the references taken in Section 6.

1.1 Motivation and Background

Despite the progress in United Kingdom, the air pollution has begun to decline from past decades. UK court has presided the government for taking instant actions to retrench the air pollution in 2015. UK exceeding the number of deaths mainly from air pollution by 29,000 a year. The major air pollutant behind this cause was been NO2. NO2 is generally been induced from the diesel vehicles emission. Chiusolo et al. (2011 pp.1233-1238). It causes problem in lungs which further results in problem in breathing. UK was always marked breaking the NO2 damper since 2010. Stuart and Zeager et al. (2011). London being the capital of UK was been issued for high alert by monitoring station as the pollution level was higher than 100µg/m3 in 2016. This was an overview for my motivation behind this topic. My contribution by this project will furnishes a high standard of prediction mechanism for air pollution and will also help in reducing exposure to air pollution that will attenuate heath issues.

The purpose of this project was to predict the No2 air pollutant in United Kingdom, by utilizing advanced time series model. Today, cities of United Kingdom is facing a major issue of air pollution by fine particulates from smoke, dost and many other effecting the people (Mahmood, 2015). According to World health Organization et al. (2018), air pollution is been the prime cause of lung cancer.

1.2 Research Question

"To what extent can we improve the air pollution in United Kingdom which has been caused by (Pollution, population and transportation) using forecasting techniques (ARIMA, LSTM, SARIMA and TBATS) to enhance life span of people?".

1.3 Research Objective and Contributions

Table 1 presents the objective that were used to search the research question. Finding the research paper related to the air pollution in United Kingdom is the first objective. Other major objectives are discussed below.

Table 1 Objective and Implementation

Objective	Implementations
Objective 1	A review in literature for the Air pollution from 2013-2015
Objective 2	Implementation and evaluation of air quality forecasting models.
Objective 2 a)	Implementing, evaluating and acquiring results of ARIMA model.
Objective 2 b)	Implementing, evaluating and acquiring results of SARIMA model.

Objective 2 c)	Implementing, evaluating and acquiring results of LSTM model.
Objective 2 d)	Implementing, evaluating and acquiring results of Stacked LSTM model.
Objective 2 e)	Implementing, evaluating and acquiring results of TBATS model.
Objective 3	Comparison of developed model with its results

Contribution: The major contribution resulting from the project is air quality forecasting models to help the people living in the cities of UK as well as the government to make strict rules for reducing the amount of pollution contributed by the industries, vehicle and many other. For reducing the air pollution in United Kingdom it is now crucial to monitor the levels of pollution in air that we breathe. By forecasting the air pollution in United kingdom for upcoming years, it will favour people living in the country by deciding how to act prominently. The impacts on the increased population, ecosystem and other natural environment will be helpful by monitoring the pollution level. It will also avail people to dump any skin care ointment. Public should be get notified by the updated and reliable information regarding the pollution. Accurate Forecasting will help the person to think of taking proper actions for eliminating the pollution. People will also think of the effects they will concur from pollution on their health, cost they have to provide for the treatment so proper precautions they will take.
Minor contribution is the comparison between the developed models and to find out which model best predicts air quality index so that necessary actions could be taken to control it.
Basically, the forecasting will not only change a person behaviour but also change the policies of the people. This way the government will also make use of forecasting by limiting the air quality level in air and can change the consumers behaviour by marketing from advertisement.
The remaining report is been structured as, chapter 2 which introduce the existing literature of air quality, based on the air pollutant, learning methods and the countries in Europe. Chapter 3 will introduce the methodology used and the project will further go with chapter 4 which is the implementation of models, evaluation and results of everything. Chapter 5 is the conclusion and how the project will help in future work.

2 Literature Review on Air Quality (2009-2019)

2.1 Introduction

This section gives an overview of Air pollution in different countries in last 10 years, this is presented for giving an idea of what all work was done for pollution control. The prior studies helped in comparing and improving the gaps in this project. The section discussed below was an overview of the prior studies related to research topic in this project. The section was divided into 3 different parts, the first part consists of the various deep learning techniques in pollution domain, the second part consists of all the pollution related work done in Europe and the third part contains the research done on the air pollutant NO2.

2.2 A Critique of Deep Learning Techniques

In this study by Chaudhary et al. (2018) stacked LSTM model is been proposed for predicting the future air pollutant concentrations for O2, No2, SO2, PM10, PM2.5. The dataset is

different from this as the forecasting is done for Delhi city of India. The dataset taken creates the pollutant levels, traffic, and other meteorological features for revealing the forecasting of pollutant. The result determines that the pollution in Delhi with root mean square error equal or less than 5 for the next hour of forecast. The paper has only applied LSTM model which has a drawback that the result can't be compared with other time series models. In this research done by Lin et al. (2011) Support Vector regression (SVR) models have been applied for prophesy of the air pollutants with the SVRLIA model and the logarithm pre-processing procedure. The SVRLIA model can accurately forecast concentrations of air pollutants. For determining the workability of the developed SVRLIA model, particulate matter, nitrogen oxide/dioxide were gathered. From the results, it was found that SVRLIA model was found pretty accurate for forecasting concentrations of the pollution in air. The future perspective the model developed can be handy from the viewpoint of an application side by collecting more real time series data. Russo et al. (2013 pp.822-830) have done prediction of the air pollution in the urban city of India was constructed by performing Ensemble models. The fuel combustion and the vehicular emission was highlighted from principal component analysis as the major reason for air pollution during both summer and winter seasons. Various decision trees methods were constructed by taking several parameters and comparing the same with the standard machine learning, SVM. The Decision Tree Forest (DTF) and Decision Treeboost (DTB) models outperformed the SVM with 95% accuracy. For future work these models can act as the tools for predicting air quality index and for management purpose as well. When going through this research by Zhu et al. (2018) some unique and refined models are been proposed which was found restricted in most of the prior studies. When applying the prediction models the paper proposed regularization for Multi-task learning problem. The prediction was done hourly for air pollutants like sulphur dioxide, ozone and particle matter by using the machine learning approaches. This paper has taken quite a big data as the research was on an hourly basis, which makes the accuracy more convincing. The experiments result in achieving high performance for hour related regularizations in comparison with the regression models. A deep learning model is proposed in this paper by Zhang et al. (2014, pp. 3-5). for examining Internet of things(IoT) smart city data. The novelty in this paper was the long-short-term-memory networks for predicting the future values of air quality index in some of the cities. SVR based model gave 92.9% accuracy while LSTM based model brings out to give 95% accuracy and achieved very superior performance even in long historical data with a simple structure. The results were found promising in the future work and the same can be helpful for other cities problem as well.

The research by Delavar et al. (2019) for predicting air pollutions based on PM10 and PM2.5 pollutions concentrations in the city of Tehran is been done in this paper. The machine learning methods such as NARX, ANN, GWR and SVR are been included. The error percentage were also been improved by proposing prediction models. The NARX turn out to be the most optimum methods with high accuracy. Quixotic of the pollution coming from traffic in urban areas were the major cause of final air pollution. While considering the future research, first the stations which calculates the air pollution should be increased hence increasing the quality of the model proposed in the paper. Secondly, in the paper hourly data is been taken which will assist in the quality of the model suggested. Research by Li X et al.,(2016) support vector regression (SVR) models, spatiotemporal artificial neural network (STANN) and auto regression moving average (ARMA) is compared with the stacked autoencoder (SAE) model. The novelty in this paper was it proposed spatiotemporal deep learning (STDL) method for predicting air quality. The method which was been implemented in the paper has performed superior with 88% accuracy when compared with other time series models. Variety of studies have been done for monitoring the PM2.5 concentrations,

but for accurately calculating the pollutant, this paper has proposed implementing LSTM (Long Short-Term Memory) with RNN (Recurrent Neural network). For more accuracy they have taken high data sets and air pollutants can be learned more accurately even in atmospheric airflow of data. Finally, by the experiments performed in this paper for 60 stations in the city of Taiwan was able to detect PM2.5 concentrations. In this research by Zhao X et al. (2018 pp.346-354), deep learning is been availed for predicting air quality for time series data. Three models that were RNN, Random Forest and Support vector machine were considered out of which RNN model has performed the best. In the future more variables can be identified which will refine the analysis. The prediction can help in authorities making issue for important air quality data to the people in the city. Researcher Freeman et al. (2018). presents applications of deep learning techniques for predicting air pollution in time series. Like others, this research predicts the ozone(O3) pollutant, using recurrent neural network with LSTM. After training the LSTM model it was found that the accuracy has been improved when been measured from Mean Absolute Error. RNN doesn't performed well when compared to ARIMA and FFNN for other researchers with same datasets. Zepeng et al. (2019) proposed an air quality prediction model using Long Short-Term Memory (LSTM) and K-nearest neighbor (KNN) for the city of Beijing. From the predictions results high prediction accuracy was figure out nearer to the AQI value. The Mean absolute percentage error came to be around 6.592. Though the drawback came when other factors makes the prediction inaccurate.

2.3 A Comparative Review of Air Pollution in Europe

Europe has recognized the major air pollutant and found the burning of fossil fuels in industry or in agriculture field as the principal cause for the pollution. Europe has therefore set targets in upcoming 2020 year for all the air pollutants including No2. This will reduce the emission of the gases by more than twenty percent and also achieve an energy consumption by twenty percent in 2020. Chen et al. (2019) has studied and compared sixteen algorithms by regression and machine learning methods for the forecasting of PM2.5 and NO2 pollutant all over Europe. Out of 16 algorithm, machine learning algorithm was ANN, RF, SVR, GBM and KRLS and the other linear regression models were FLR, BLR, WLM, SLR, SVR. From the result it was stated that the boosted machine, random forest and bagging performed better than other Regression models and ANN for both the pollutant. The biasing was found to be low for every algorithm except Artificial neural network. Elangasinghe et al. (2017)also uses ANN model but with a combination of k-means clustering for determining PM10 and PM2.5 concentrations for time-series in the country New Zealand. In the result, the correlation coefficient was been enhanced from 0.771 to 0.791 for PM2.5. Similarly, for PM10 it is been increased from 0.63 to 0.69. Freeman, B. (2017), presents a unique approach for the evaluation of air quality methods by using Central limit theorem and Monte Carlo analysis. The author basically compares an ambient air classification approach over another. Results determines that the chronic daily intake has been differed by only 2.7% for certain period. Prybutok et al. (2000), studied by comparing neural network models with ARIMA and other regression models to predict maximum ozone concentrations. The comparison was done multiple times such that high accuracy and better model can be predicted and hence found Neural network was found very more efficient than regression and ARIMA models Researcher Stafoggia et al. (2019), has estimated PM10 and PM2.5 concentrations using machine learning methods from 2013-2015 in Italy. The researcher has merged different data for predicting satellite AOD by the ensemble models assist with machine learning models for obtaining and resolved PM predictions by taking more than 500 stations for both PM2.5 and PM10. The models were highly stable, and the predictions samples captured more than 94 %

of the variability in AOD. Russo et al. (2013) Researcher has applied artificial neural network models for forecasting the air pollution. The papers goal was slightly different from other papers as it trains the neural network for reducing the amount of input variables and keeping the predictive power of the proposed model constant.

2.4 A Review of No2 Pollutant for Forecasting

Osowski et al. (2007 pp.745-755) has tried to apply wavelet decomposition and support vector machine (SVM) for forecasting the daily air pollutant like SO2, CO, NO2 for the country Poland. Author break-up the dataset into a wavelet representation for obtaining high accuracy. It has trained No2 pollutant which was able to produce good prediction for other pollutant i.e. SO2, CO and dust. The mean absolute error was 4.88 [mg/m3] and standard deviation of error was 1.97[mg/m3]. Mayer et al.(1999) has done study in late 90s for the city Stuttgart in southern Germany. The author has taken time series trends from the air pollutants such as NO, NO2, O3 and Ox are. By this he compared the condition of air pollution with other cities. Results shows that the pollution in air is basically arriving from light industries and some domestic sources. In the research done by Neal et al. (2014 pp.385-393) automatic air quality forecasting for the current observations are been described by bias correction scheme. The implementation is been done by taking air quality forecasting for PM2.5, O3, NO, PM10 of United Kingdom, for enhancing the performance. SPPO technique has been used in the paper which is based on the short-term persistence for forecasting pollution. This technique gives benefits by inhomogeneity and for significant pollutant. Lloyd et al. (2004 pp.293-305) researcher has done mapping of No2 air pollutant for the United Kingdom. The techniques used for predicting point data were simple kriging, local linear regression (LR), ordinary kriging (OK), inverse distance weighting (IDW). Arhami et al. (2013, pp. 4777-4789) studied the use of machine learning algorithms for predicting air pollutant like NO, CO2, CO, and PM10 level using Artificial neural network and Monte Carlo simulations. The study proves that ANN performs significantly accurate for every pollutant except O3. The study concluded that for enhancing the performance of ANN models, deterministic prediction can be replaced by probabilistic PIs.

2.5 Identified Gaps and Conclusion

After reviewing all the literatures carefully it was found out that prior papers involved generally forecasting model with their time series, however the drawback with most forecasting model is there incapability to handle time series data with trend and seasonality in order to resolve the novel TBAT model has been implemented in this research which has capability of handling multiple seasonality and trends. All the models will be compared with the novel TBAT model based on motive evaluation matrix such as MAE and MSE. Since LSTM is a type of RNN it uses past values to forecast the future since the pollution data is dependent on its past values stacked LSTM model would be a good solution. The second part involves the study of pollution forecasting in Europe. Researchers forecast the data for upcoming years but was not that accurate, this project will try to produce better results. Table 2 presents comparison of reviewed techniques and methods. The major literatures which are closely related to our topic is been shown in this table.

Table 2 Objectives

Models	Evaluation parameters	Problem Addresed	Factors applied	Author name
ANN, RF, SVR, GBM, KRLS, SLR, KRLS	MAE and RMSE	Comparing models for developing spatial models and NO2	No2, CO2, PM 2.5	(Chen et al.,2019)
support vector machine (SVM)	The mean absolute error was 4.88 [mg/m3] and standard deviation of error was 1.97	Forecasting the daily pollution using support vector machine	maximum temperature mean wind speed	(Osowski and Garanty, 2007)
SVM, DTF and Decision Tree boost(DTB)	Accuracy was 95%	optimal neural networks for predicting pollution with variables	Hourly mean, temp, Pressure, humidity	(Russo, Raischel and Lind, 2013)
RNN,SVM,RF	Accuracy was RNN= 76.44% SVM=75.89% RF=75.06%	deep recurrent neural network for air quality classification	CO, No2, O3	(Zhao, Xiaosong, et al.)

3 Scientific Methodology Approach and Design Specification

3.1 Air Quality methodology approach

This section will give an overview of the methodology presented in this research project. Out of KDD and CRISP-DM, CRISP-DM was chosen which is the common methodology being adopted by many of the researchers. The methodology used in this forecasting of air quality pollution project is CRISP-DM (Figure 1). Cross Industry Standard Process provides a simple and clear model for analysing the data. In this project, CRISP-DM will come up with a strong and flexible methodology. Though this methodology has five steps, but for the specifications in our project (forecasting air pollution), we will make it in six hierarchical steps as shown in figure 1. Basically, an updated CRISP-DM model will be proposed according to the business requirement for this project.

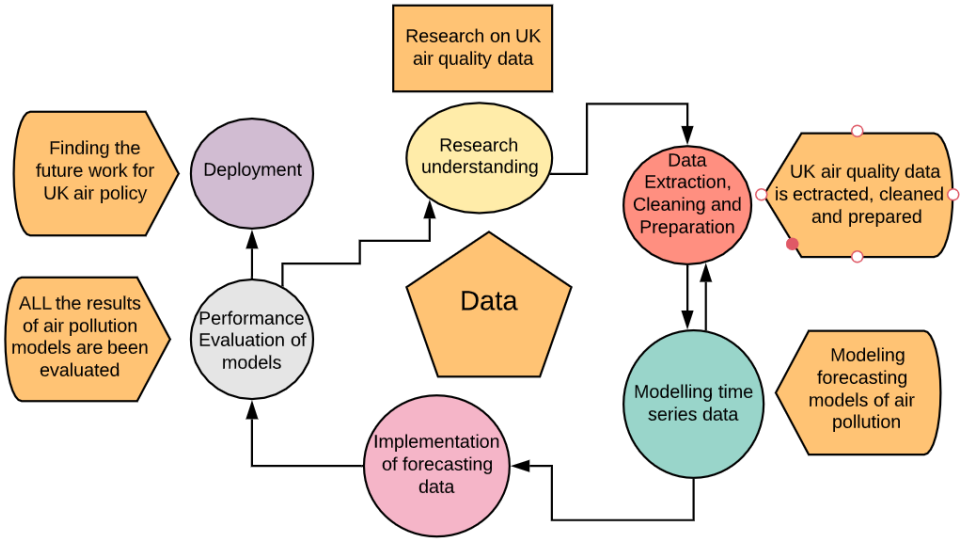

Figure 1 CRISP-DM methodology

The methodology of air quality methodology is modified for the needs and specification of this project. The 3tier architecture is used in this project architecture. The implementation, evaluations and results of the models is done for identifying the air pollution in Europe.

3.2 Project Process Flow Diagram

Business Understanding:

It starts with the business understanding which is the first step for the scope of the project. Business understanding determines the most part of the objective for the research. As we are forecasting the pollution in UK taking some pollutant into consideration, the overall understanding of the methodology which was utilized for the project was been searched and research has been done. In other word RTD i.e. research and technological development was undertaken in the first step.

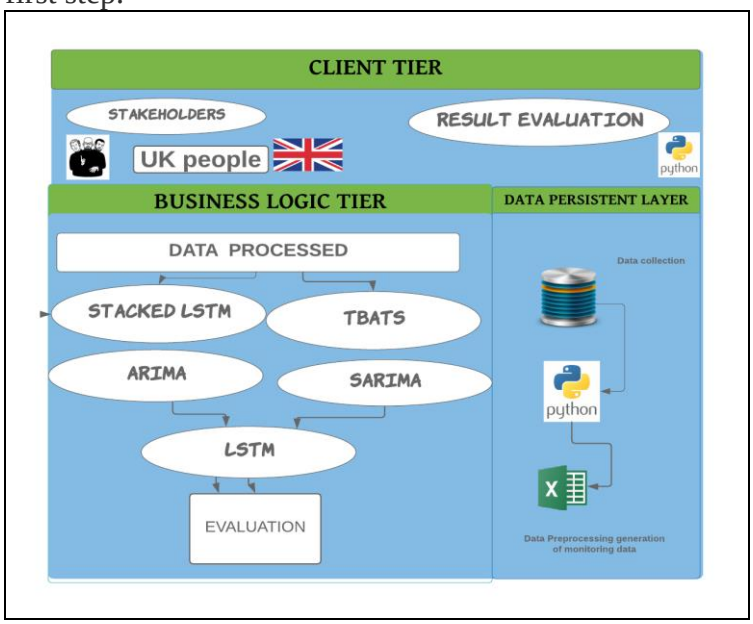

Figure 2 Design Specification

Data understanding:

In this phase of CRISP_DM, the data from the source is been obtained and test for the suitability of the data. This phase is important as by any flaws, it will result in revising the plan. For the research done in this project, data is been taken from The European Environment Agency. The website assists in extraction of NO2 air pollutant from the year 2015 till 2018. The website from which the data can be download anytime anywhere globally is https://www.eea.europa.eu/data-and-maps/data/aqereporting-2/gb/gb-aqereporting-2014.
The folder has two different files for all the pollutants specify by their pollutant number such as for PM10 pollutant its 5. The folder has 17 field names for every pollutant in United Kingdom. As the dataset is huge, missing values by various factors such as wrong entry of data or wrong reading can be anticipated. Hence it is been reduce in the next level.

Data preparation:

The most crucial phase in the implementation part is data preparation where all the messy data is cleansed from missing values and other special characters. The navigator name as Anaconda which enables a user to launch application such as Python inside it is been operated for the data preparation in this research. Anaconda navigator was chosen against other python terminal because of its package handling capacity. All the packages are run, install and updated in the environment and there is no need of doing the same again and again.

Data cleaning

Data cleaning is always the time taking and the major part of implemntation as a small mistake will lead to a bigger failure later with the models. As the dataset downloaded has 29,589 rows and 17 columns. Manually cleaning is not efficient, so the cleaning is done in Python. Cleaning requires the following steps:

- Removing missing value & Blank value: First the variables are been spotted and all the 6 variables were then removed from the dataset. Now all the missing values are been recognized and by imputation all the missing values are eliminated. Though there is a difference between missing values and some zeroes, as being zero value determines zero pollution in that area. So, zeroes were remained same.
- Eliminating same variable: Two variables name Air quality station and air quality EI code were found to be same, so one variable was been removed for preventing duplication. High correlation was found in one of the fields which can't be neglected and hence been eliminated
- Unit of Measurement: The unit in which pollution level was been measured was found to be different in one period and different in other period, resulting in faulty results hence all the units are been set to one default unit.
- Filtering out high data: In months of January, a large number of records on the 1st of Jan of 2014 and 31st of Dec of 2015 were found. These are the measurement done for the whole year and hence it can be filtered as the time period taken in the implementation lies within that range so for proper forecasting the data should be consistent.
- Removal of Sampling point & insertion: In some of the pollutant levels the measurement done is not complete, hence by taking the measurement of the previous days the pollutant can be predicted for the current day. Inserting a row with the next day data by taking minimum timestamp required. The sampling point is been cut off for days with less than 500 time stamp.

The methodology of CRISP-DM is modified for the needs and specification of this project. The 3tier architecture is used in this project architecture. The implementation, evaluations and results of the models is done for identifying the air pollution in Europe.

4 Implementation, Evaluation and Results of Air Quality Forecasting

4.1 Introduction:

The models in time series are evaluated by testing the performance based on the execution time, the root mean squared error and finally the mean absolute error. Unlike other models which used to test the percentage of accuracy for checking the model's throughput, time series model acts differently.

RMSE: RMSE which denotes root mean squared error. It is been specified as the gap between the samples taken for prediction and the samples taken by observation. Root mean square error is sometimes called as root mean square deviation, the deviation means residuals when the results are taken from the data sample and when the sample is gone out of the records it is determine as errors. The accuracy in RMSE also serves by the comparing the prediction errors for the models used from the dataset being performed. RMSE can't be negative, as it is an error which can either be very high or very low i.e. 0. For the model to be perfect the RMSD is expected to be very less.

Equation for RMSE is:

$$RMSE = \sqrt{\overline{(f-o)^2}}$$

f = forecasts result,
o = observed result.

MAE: The mean absolute error is the gap between predicted error and the truth errors. Its formulae is:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{x}|$$

Where:
N = no. of errors,
Σ = summation for adding
|xi – x| = the absolute errors.

MSE: MSE is the mean squared error which simply justify the closeness of the line with the given points. The justification is been found by calculating the distance from points till the regression line and thereafter squaring it. The process of squaring the number is been done because of the reason than some numbers are in negative value which will be automatically comes in positive value.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Where:
N = no. of errors,
Σ = summation for adding
|yi−¯y|= the absolute squared errors.

As discussed in the previous section, the implementation is done in python. First all the packages are imported in python by using import 'packagename' as path. After that the project directory is been set-up, where all the changes in the file will be reflected ''C:/yash/ Airpollution". The data been imported and then the directory is been set

4.3 Implementation, Evaluations and Results of ARIMA Model

Arima which stands for 'Auto Regressive Integrated Moving Average' is a class that performs on a time series by considering its own past values. Past value defines the forecasted errors and the lags in its own value. ARIMA can therefore forecast future values by the equations came from the errors. The model's equation is represented as:

ARIMA(p,d,q)

Where,
P is the order of AR term
Q is the order of MA term
D is the order of differencing requred to make time series stationary

Implementation:

ARIMA model has been implemented by using 'ARIMA' function from statsmodel libary by the function auto.arima(). The values of (p,d,q)has been taken as (1,1,1). The ARIMA model has been fitted by the function model.fit(). After performing the implementation the summary of the result was been defined by results.summary() function. The mean absolute error and the mean squared error was needed which came by the function mae_arima and mse_arima. The values of mse_arima and mae-arima came out to be 0.73 and 0.706. These results are presented in table 3.

Table 3 Evaluation of ARIMA

Forecast	MAE	RMSE	MSE
1 year	0.70	0.42	0.73

MAPE was on the higher side because a lot of negative values were present in ARIMA predictions.

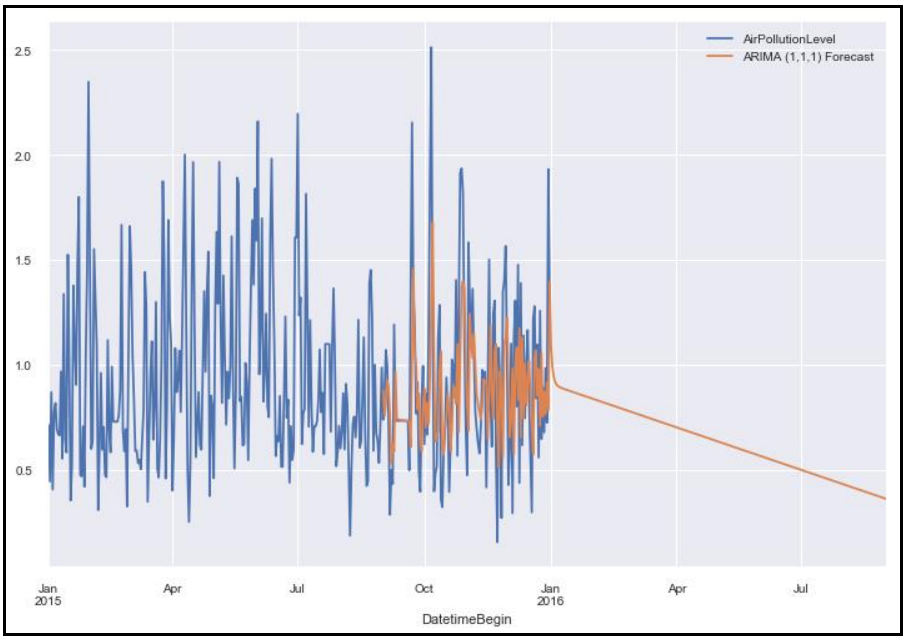

Figure 3: ARIMA PREDICTION

From Figure 3 it can be anticipated that ARIMA is not capable of handling data with rise and falls as well as trends and seasonalty, below per performace were received by applying ARIMA. SARIMA gave low errors than SARIMA as it has the feature of handling seasonalty in data. It shows that general time series models such as ARIMA is not suitable for pollution data sets with high variance.

4.4 Implementation, Evaluation and Results of SARIMA Model

SARIMA which stands for 'Seasonal Auto Regressive Integrated Moving Average'. As the name suggests it is formed by including seasonal terms in ARIMA. The SARIMA model's equation was represented by:

SARIMA (p,d,q)

Where,
P is the seasonal autoregressive order
D is the seasonal difference order
Q is the Seasonal moving average order
M is the number of time steps for a single seasonal period

Implementation:

SARIMA model has been implemented by using 'SARIMA' function from statsmodel libary by the function auto.sarima(). The values of (p,d,q)has been taken as (1,1,1). The SARIMA model has been fitted by the function model.fit(). After performing the implementation the summary of the result was been defined by results.summary() function. The mean absolute error and the mean squared error was needed which came by the function mae_sarima and mse_sarima. The values of mse_sarima and mae-sarima came out to be 0.177 and 0.34.

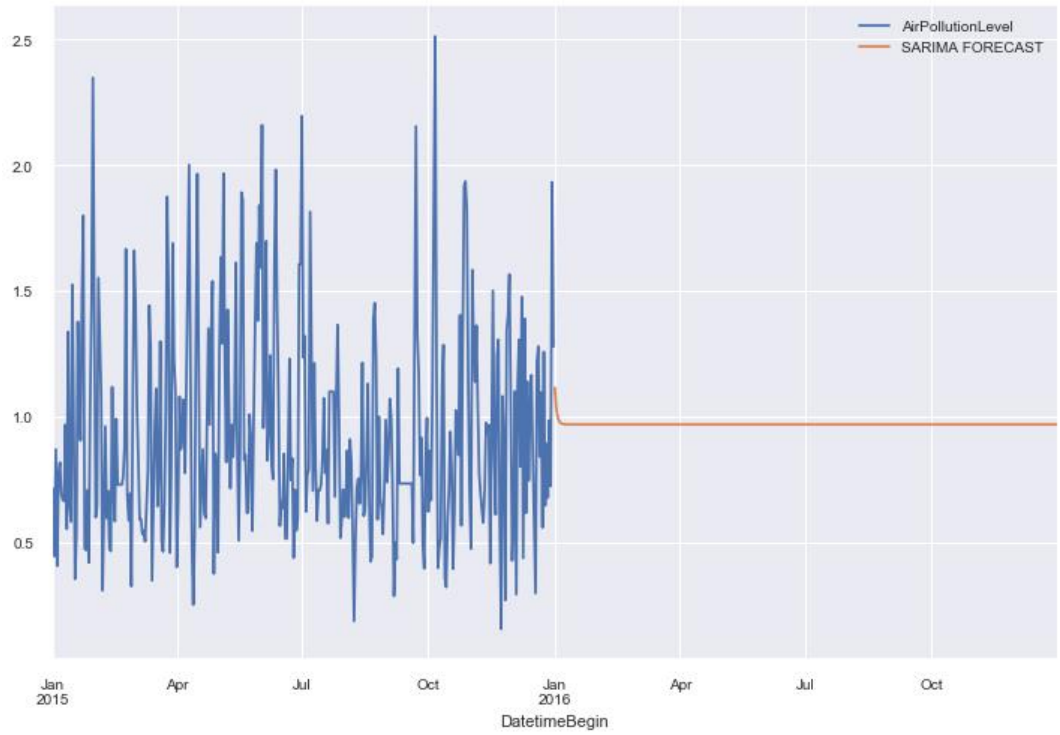

Figure 4 SARIMA Mean Forcasting

According to fig 4, it can be seen that SARIMA gave below average performance while predicting the pollution, after a rise in the values, a mean line was plotted by SARIMA. The reason can be that the data has frequent rise and falls and SARIMA wasn't able to model those points.

Table 4 Evaluation of SARIMA

Forecast	MAE	RMSE	MAPE	MSE
1 year	0.344	0.421	33.94%	0.17

By table 4, the results of Error are quite low which describe that the model is performing well. But from the predicted graph, the prediction went very variant. SARIMA still have very low error which can say that the model is fit.

4.5 Implementation Evaluations and Results of LSTM model

LSTM stands for long short-term memory networks. The problem of long -term dependency of remembering the data information is been neglected in LSTM. Feature scaling is been used in this model. As there is only one column its better to scale the numbers. LSTM is so called as it has the capacity to predict the future values by remembering the past values.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c)$$
$$h_t = o_t \circ \sigma_h(c_t)$$

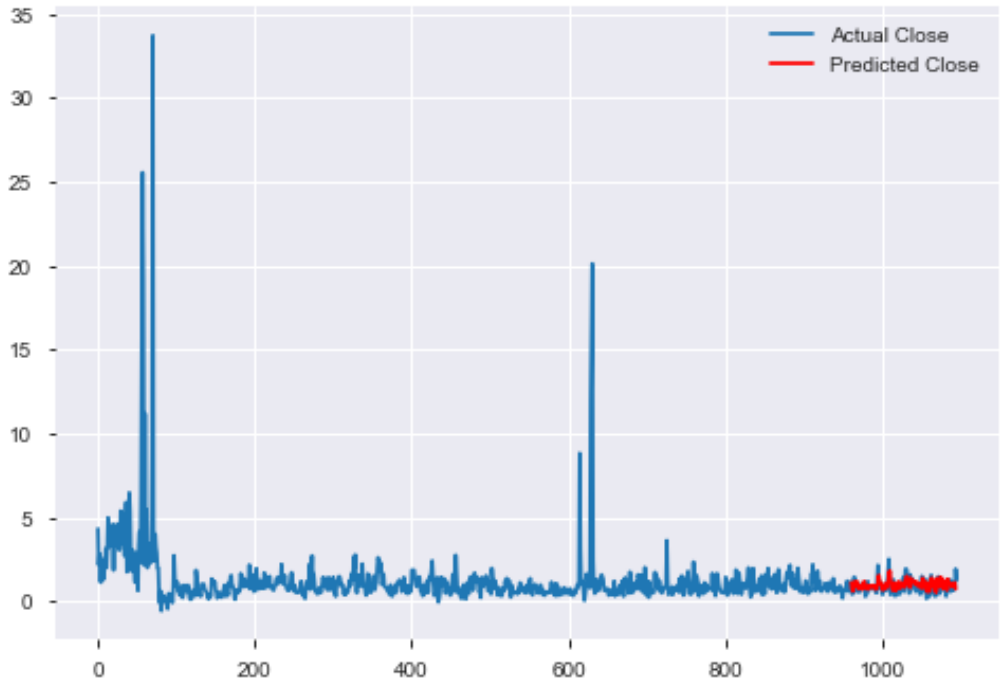

Figure 5 LSTM

As it can be seen from the above figure 5, the predicted values are overlapping the real values. LSTM due to its nature of remembering past values to predict the future value was able to provide low forecasting errors in comparison to other apllied models such as ARIMA and SARIMA.LSTM was applied using 'LSTM function'.'sequential was used for the input layer where as the 'dense function; was used as the output layer. Multiple training units where tested. However the best accuracy was obtained by 30 units. Min max scalar was used for feature scaling as LSTM was not compatable with non scale data. Inverse transform function was used to transform the scaled values back to their original format. The evaluation table has been represented below.

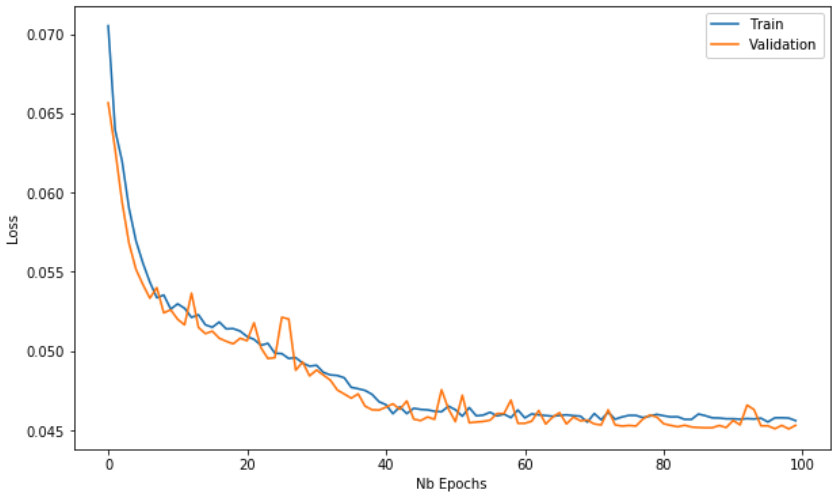

Figure 6: Simple LSTM – Train & Validation Loss

From the figure 6, it was evaluated that Loss function value has dropped across the training iterations which means that the model can predict general results. Hence, we can conclude that model is not subjected to overfitting and will perform significantly good on an unseen data.

Table 5 Evaluation of LSTM model

Forecast	MAE	RMSE	MAPE	MSE
1 year	0.347	0.43	36.25%	0.18

As seen in Table 5, The percentage error which is the MAPE came out to be 36.25%. The LSTM model is been fitted properly. It has predicted the result more efficiently than other models.

4.6 Implementation, Evaluation and Results of Stacked LSTM model

As discussed in the above section, LSTM has single hidden layers. Though the results were satisfactory but for better results stacked LSTM is used. As stacked LSTM uses multiple hidden layers which makes it easier for the memory to keep the previous input for the longer time and high inputs can be kept. It consists of the three layers input, output and the hidden layer in between both. The data will flow from one to another path. The shape of these hidden layers are directly proportional to the recurrent neural network. The loops start by passing the data say ‘k’ and the hidden state. The stacked LSTM model will then simply give back the output ‘k’ with the updated hidden state. The loop will go on until left with the last data say ‘k+100’, So, after sending ‘k+99’ the loop will stop and will produce an output to the FF layer with a predicted value say ‘P’. This way the for loops work.

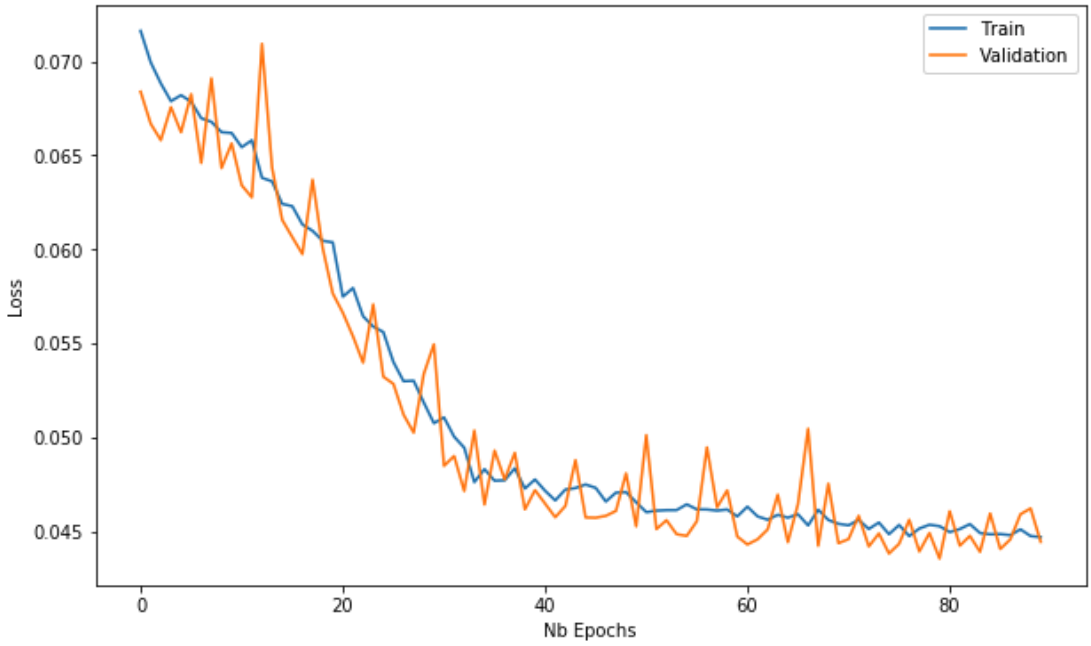

Figure 7 Stacked LSTM train & validation loss

In this figure 7 there are fluctuation in the validation loss function in the epoch axis. Since the value has dropped from the initial to the final epoch, we can assume that the model has a good prediction capacity. Hence, we expect that LSTM will perform slightly better than stacked LSTM as the nature of curve is different.

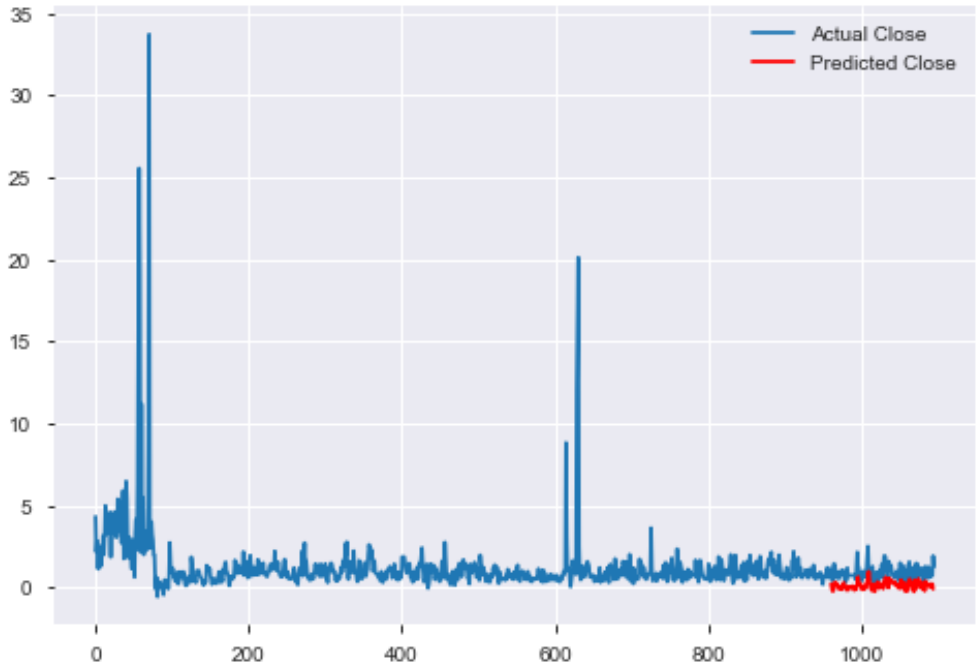

Figure 8 STACKED LSTM

The above figure 8 good performance in terms of forecasting air pollution as the predicted and the true values lie nearby. LSTM outperformed ARIMA, SARIMA in terms of evaluation

measures MAE, MSE, MAPE, RMSE. Therefore deep learning methods such as LSTM can be used for accurately forecasting air pollution data.

4.6 Implementation, Evaluations and Results of TBATS model

After implementing ARIMA and SARIMA, TBATS model was used using TBATS function which was imported from stats model library. In TBATS model, T stands for trigonometric regressors to model, B stands for Box-Cox transformations, A stands for ARMA errors, T stands for trend and S stands for the seasonality. It basically mixes the exponential smoothing with a Box-Cox transformation and whole this process is been automated. TBATS is considered to be different from other regression models by changing the pattern over given time period. As the pattern is high so the model tends to run slowly which can be considered as a negative side of the model. Since TBATS can handle multiple seasonality and trends it gave out better forecast than ARIMA and SARIMA as well. The final comparison of TBATS with Stacked LSTM has been explained in the results section.

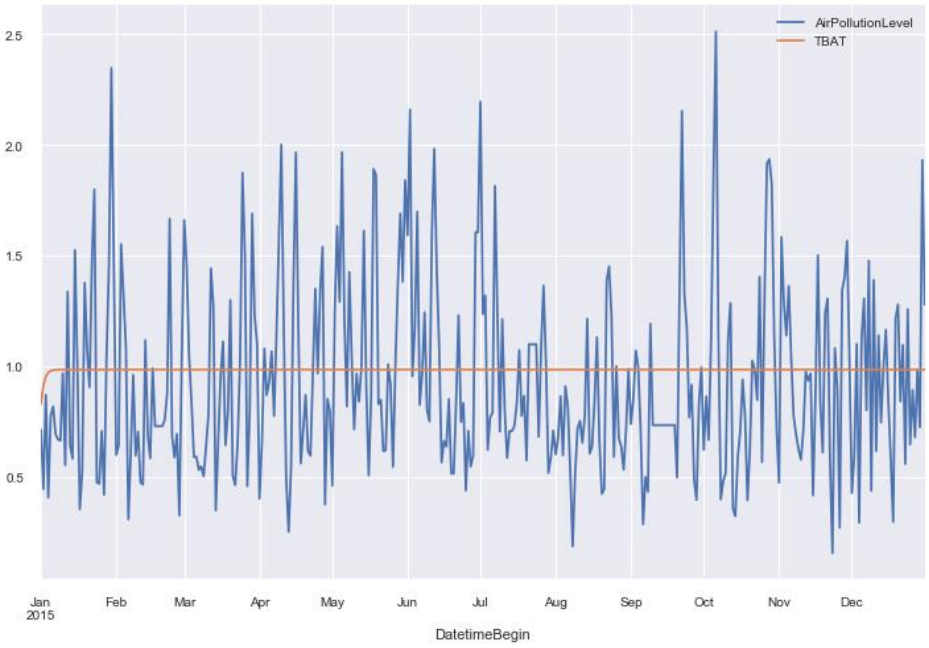

Figure 9 TBAT

According to figure 9, it can be seen that TBATS gave below average performance while predicting the pollution, after a rise in the values, a mean line was plotted by TBATS. The reason can be that the data has frequent rise and falls and TBATS wasn't able to model those points.

Table 6 Evaluation of TBATS

Models	Forecast	MAE	RMSE	MSE
TBATS	1 year	0.344	0.421	0.17

According to Table 6, the MSE came out to be the lowest which tells that the model has fit very well while comparing it with other models. As expected the RMSE and MAE also came out to be very low.

4.7 Discussion and Comparison of Developed Models

From the results shown in table 7 it is clear that the developed model was better than the other models. MAE, RMSE and MSE came out to be the highest for ARIMA. TBATS and SARIMA were the best performing model in terms of all evaluation metrics along with satisfactory values for LSTM. Since, only one input and output layer were used for LSTM, the performance could be enhanced by applying stacked layers on top of LSTM. However, the same is proposed for future. Therefore, it can be concluded that the developed TBATS model can accurately forecast the pollution with low RMSE, MSE and MAE values. Researcher (Chen et al., 2019) has tried to compare the learning models for NO2 for the same dataset as used in this project. From the result these models applied in this project had also performed well by taking different models.

Table 7 Comparison Table

Models	Forecast	MAE	RMSE	MSE
ARIMA	1 year	0.70	0.425	0.73
LSTM	1 year	0.347	0.43	0.18
SARIMA	1 year	0.30	0.39	0.15
TBATS	1 year	0.344	0.421	0.17

5 Conclusion and Future Work

Upon careful evaluation of the applied models, it can be seen that SARIMA and TBAT came out to be the best performing models in terms of MSE, MAE and RMSE. The reason behind this could be the capability of both the models with handling multiple seasonalties. The performance of LSTM was satisfactory in comparison to TBATS. Unsurprisingly, Arima was the worst performer among all.

Future Work: The best performing models could be evaluated and tested on different datasets, In order to check their performance on new data set. Additionally, different parameters could be try on SARIMA and TBATS to further improve their performance. Moreover, additional layers could be added to LSTM in order to increase their performance in forecasting pollution. So, stacked LSTM is proposed for future work.

Acknowledgement: I would like to thank Dr. Catherine Mulwa for providing me constant support to carry out my thesis. Her direction and constant motivation help me to comprehend the research objective in an efficient way. I would like to appreciate EEA for making the dataset publicly available which help me to undergo this thesis.

References

Air Pollution Monitoring and Prediction System. (2019). *International Journal of Recent Technology and Engineering*, 8(2S3), pp.648-651.

Arhami, M., Kamali, N. and Rajabi, M. (2013). Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. *Environmental Science and Pollution Research*, 20(7), pp.4777-4789.

BBC News. (2019). *UK's most polluted towns and cities revealed*. [online] Available at: https://www.bbc.com/news/health-43964341 [Accessed 4 Dec. 2019].

Beckerman B, Jerrett M, Brook JR, Verma DK, Arain MA, et al. (2008) *Correlation of nitrogen dioxide with other traffic pollutants near a major expressway.* Atmos Environ 42: 275–290.

Brauer M, Hoek G, Van Vliet P, Meliefste K, Fischer PH, et al. (2002) *Air pollution from traffic and the development of respiratory infections and asthmatic and allergic symptoms in children.* Am J Respir Crit Care Med 166: 1092–1098.

Brauer M, Lencar C, Tamburic L, Koehoorn M, Demers P, et al. (2008) *A cohort study of traffic-related air pollution impacts on birth outcomes*. Environ Health Persp 116: 680–686.

Chan, C. and Yao, X. (2008). Air pollution in mega cities in China. *Atmospheric Environment*, 42(1), pp.1-42.

Chaudhary, V., Deshbhratar, A., Kumar, V. and Paul, D., 2018. Time Series Based LSTM Model to Predict Air Pollutant's Concentration for Prominent Cities in India.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U., Katsouyanni, K., Janssen, N., Martin, R., Samoli, E., Schwartz, P., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Vermeulen, R., Brunekreef, B. and Hoek, G. (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environment International*, 130, p.104934.

Chiusolo, M., Cadum, E., Stafoggia, M., Galassi, C., Berti, G., Faustini, A., Bisanti, L., Vigotti, M., Dessì, M., Cernigliaro, A., Mallone, S., Pacelli, B., Minerba, S., Simonato, L. and Forastiere, F. (2011). Short-Term Effects of Nitrogen Dioxide on Mortality and Susceptibility Factors in 10 Italian Cities: The EpiAir Study. *Environmental Health Perspectives*, 119(9), pp.1233-1238.

Delavar, M., Gholami, A., Shiran, G., Rashidi, Y., Nakhaeizadeh, G., Fedra, K. and Hatefi Afshar, S. (2019). A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran. *ISPRS International Journal of Geo-Information*, 8(2), p.99.

Elangasinghe, M., Singhal, N., Dirks, K., Salmond, J. and Samarasinghe, S. (2014). Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering. *Atmospheric Environment*, 94, pp.106-116.

En.wikipedia.org. (2019). *Root-mean-square deviation*. [online]

https://en.wikipedia.org/wiki/Root-mean-square_deviation [Accessed 4 Dec. 2019].

Filluel L, Rondeau V, Vandentorren S, Le Moual N, Cantagrel A, et al. (2005) *Twenty five year mortality and air pollution: results from the French PAARC survey.* Occup Environ Med 62: 453–460.

Forecasting air quality time series using deep learning. *Journal of the Air & Waste Management Association*, 68(8), pp.866-886.

Freeman, B., McBean, E., Gharabaghi, B. and Thé, J. (2017). Evaluation of air quality zone classification methods based on ambient air concentration exposure. *Journal of the Air & Waste Management Association*, 67(5), pp.550-564.

Freeman, B., Taylor, G., Gharabaghi, B. and Thé, J. (2018).

GOV.UK. (2019). *Public Health England publishes air pollution evidence review*. [online] Available at: https://www.gov.uk/government/news/public-health-england-publishes-air-pollution-evidence-review [Accessed 4 Dec. 2019].

Grineski S, Bolin B, Boone C (2007) Criteria air pollution and marginalized populations: environmental inequity in metropolitan Phoenix, Arizona. Soc Sci Quart 88: 535–554.

Leontief, W. (1979). Population Growth and Economic Development: Illustrative Projections. *Population and Development Review*, 5(1), p.1.

Li, X., Peng, L., Hu, Y., Shao, J. and Chi, T. (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, 23(22), pp.22408-22417.

Lin, K., Pai, P. and Yang, S. (2011). Forecasting concentrations of air pollutants by logarithm support vector regression with immune algorithms. *Applied Mathematics and Computation*, 217(12), pp.5318-5327.

Lloyd, C. and Atkinson, P. (2004). Increased accuracy of geostatistical prediction of nitrogen dioxide in the United Kingdom with secondary data. *International Journal of Applied Earth Observation and Geoinformation*, 5(4), pp.293-305.

Maheswaran, R., Haining, R., Brindley, P., Law, J., Pearson, T., Fryers, P., Wise, S. and Campbell, M. (2005). Outdoor Air Pollution and Stroke in Sheffield, United Kingdom. *Stroke*, 36(2), pp.239-243.

Mahmood, I. (2015). Environmentâ€ ™s legal protection from air pollution in Iraq and the United Kingdom. *Epidemiology: Open Access*, 05(03).

Mayer, H. (1999). Air pollution in cities. *Atmospheric Environment*, 33(24-25), pp.4029-4037.

Neal, L., Agnew, P., Moseley, S., Ordóñez, C., Savage, N. and Tilbee, M. (2014). Application of a statistical post-processing technique to a gridded, operational, air quality forecast. *Atmospheric Environment*, 98, pp.385-393.

Osowski, S. and Garanty, K. (2007). Forecasting of the daily meteorological pollution using wavelets and support vector machine. *Engineering Applications of Artificial Intelligence*, 20(6), pp.745-755.

Prybutok, V., Yi, J. and Mitchell, D. (2000). Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. *European Journal of Operational Research*, 122(1), pp.31-40.

Qin, Z., Cen, C. and Guo, X. (2019). Prediction of Air Quality Based on KNN-LSTM. *Journal of Physics: Conference Series*, 1237, p.042030.

Russo, A., Raischel, F. and Lind, P. (2013). Air quality prediction using optimal neural networks with stochastic variables. *Atmospheric Environment*, 79, pp.822-830.
Russo, A., Raischel, F. and Lind, P. (2013). Air quality prediction using optimal neural networks with stochastic variables. *Atmospheric Environment*, 79, pp.822-830.
Singh, K., Gupta, S. and Rai, P. (2013). Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment*, 80, pp.426-437.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., de' Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., Viegi, G., Kloog, I. and Schwartz, J. (2019). Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Environment International*, 124, pp.170-179.

Wang, S. and Hao, J. (2012). Air quality management in China: Issues, challenges, and options. *Journal of Environmental Sciences*, 24(1), pp.2-13.

Who.int. (2019). *Air pollution*. [online] Available at: https://www.who.int/health-topics/air-pollution#tab=tab_1 [Accessed 12 Dec. 2019].

Who.int. (2019). *How air pollution is destroying our health*. [online] Available at: https://www.who.int/airpollution/news-and-events/how-air-pollution-is-destroying-our-health [Accessed 4 Dec. 2019].

Yanosky JD, Schwartz J, Suh HH (2008*) Associations between measures of socioeconomic position and chronic nitrogen dioxide exposure in Worcester*, Massachusetts. J Toxicol Env Heal A 71: 1593–1602.

Zhang, H., Chen, G., Hu, J., Chen, S., Wiedinmyer, C., Kleeman, M. and Ying, Q. (2014). Evaluation of a seven-year air quality simulation using the Weather Research and Forecasting (WRF)/Community Multiscale Air Quality (CMAQ) models in the eastern United States. *Science of The Total Environment*, 473-474, pp.275-285.

Zhao, X., Zhang, R., Wu, J.L. and Chang, P.C., 2018. A deep recurrent neural network for air quality classification. *J. Inf. Hiding Multimed. Signal Process*, *9*, pp.346-354.

Zhu, D., Cai, C., Yang, T. and Zhou, X. (2018). A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. *Big Data and Cognitive Computing*, 2(1), p.5.