The Prediction of the Air Quality based on the Prophet Algorithm

Fangjie Cai *

Department of Environmental Science and Engineering, Tongji University, Shanghai, China

* Corresponding author email: 2254366@tongji.edu.cn

Abstract. Due to the fact that the environmental problems are increasingly focused and among them, air quality is closely related to human's health. It is necessary to have a better understanding on the surrounding air quality. In this paper, a dataset collected from the internet is used to exemplify the usage of the Prophet algorithm on the prediction of air quality. Compared with the conventional monitor and analytical methods, it can not only record the up-to-date air quality indexes, but it can also do some predictions on the air quality based on the dataset and plot the trend. Since Prophet algorithm can take seasonalities and holiday effects into considerations, it is suitable to employ it to predict the changeable air quality. The prediction results were output by the codes in the form of line charts. According to the experimental results in this study, the accuracy of the model is acceptable and reliable.

Keywords: Prophet Algorithm; Data Analysis; Air Quality Prediction.

1. Introduction

Air is the mixture of gases comprising the Earth's atmosphere and it is also a crucial natural resource for human's survivals. Air pollution has become a huge challenge and its main problem is poor air quality. The Air Quality Index (AQI) describes air quality based on the concentration of pollutants in the air in an area. There are six categories that make up the AQI. A varying level of health concern relates to each category. Five significant air pollutants that are subject to Clean Air Act regulation each have an AQI set by the Environmental Protection Agency (EPA). Ground-level ozone, particle pollution (also known as particulate matter, including PM2.5 and PM10), carbon monoxide, sulfur dioxide, and nitrogen dioxide all have national air quality standards established by the EPA to safeguard the public's health. Due to the strong connection between air quality and human health, it is crucial to create an accurate and trustworthy model for forecasting air quality. This approach can assist the government learn about regional environmental challenges and, to some extent, establish applicable legislation, as well as suggest and caution residents about their outdoor activities.

In the early field of monitoring the air quality, researchers need to do the sampling works on the spots, which is lack of timeliness. Latest method to monitor the air quality includes the gas sensors and electronic noses. Although the process of collecting the relevant data of the air quality has been simplified by the technology, processing the data becomes the new problem. Two common methods are traditional forecasting methods and machine learning algorithms. Traditional forecasting methods mainly include numerical model prediction methods and statistical methods. Based on atmospheric dynamics and chemical processes, the numerical model method uses the system of equations to construct a mathematical model to simulate horizontal or vertical pollution data. Statistical methods focus less on the physicochemical properties of air pollutants, but directly combine meteorological data and air quality data from the previous step, and predict air quality by using statistical-based models. These methods can effectively combine multi-domain knowledge, but they all have certain limitations, such as the lack of flexible multi-scale framework when processing large amounts of data, and a large number of operations may lead to computational errors.

To avoid such limitations, this research employs machine learning algorithms for predicting air quality. There are some existing researches that use machine learning algorithms to build the model of air quality prediction. For example, Wang et al. use ARIMA to predict the AQI in Suzhou [1];

Yang et al. construct the air quality prediction model based on the classification and regression functions of random forest algorithm and cross-validation method [2]; Chang et al. do the scale prediction of AQI based on Prophet-random Forest optimization model [3].

This research uses Prophet to do the time series analysis on air quality prediction. This dataset, which include answers from a gas multi-sensor system placed in the field in an Italian city, comes from an essay [4]. Averages of the hourly replies are noted together with references to gas concentrations from a licensed analyzer. Prophet is a technique for predicting time series data that uses an add-on model to explain nonlinear seasonal trends that occur annually, monthly, daily, weekend, and holiday periods. It is most suitable for highly seasonal time series and historical data spanning several seasons. Prophet usually does a good job of handling outliers and resists lost data and changes in trends. Therefore, it is suitable to use Prophet to predict the air quality because the air quality index changes from time to time and can be influenced by some seasonal factors or certain festivals. In the procedure of data cleaning in this research, dropna () is used to abandon some meaningless data. Data in the dataset are transformed into float for further process.

2. Method

2.1 Dataset

2.1.1 Dataset Description and Preprocessing

The dataset came from a single investigation that Vito et al. did. 9,357 samples of hourly averaged answers from a group of five metal oxide chemical sensors that are built into an air quality chemical multi-sensor device are included in the collection. In a very polluted part of an Italian city, the device was positioned on a field at street level. From March 2004 to February 2005, or nearly a year, responses from on-field deployed chemical sensor devices were recorded for the longest period of time that is publicly available. Ground Truth received hourly averaged data for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx), and Nitrogen Dioxide from a co-located reference certified analyzer (NO2). According to the essay, there are indications of cross-sensitivities as well as concept and sensor drifts, which can impair a sensor's ability to estimate concentration. Additionally, missing values are marked with a value of -200.

In the data preprocessing, the data cleaning was carried out to make sure that the data are meaningful in the practical problems. First, this study uses data. dropna() to abandon some missing value. Secondly, the data column was parsed and transformed the measurements of the dataset to floats.

2.1.2 Prophet Algorithm

Prophet is a forecasting method that was implemented in R and Python [5-7]. It runs quick and produces fully automated forecasts that may be manually adjusted by analysts and data scientists. Prophet is a technique for predicting time series data that uses an add-on model to explain nonlinear seasonal trends that occur annually, monthly, daily, weekend, and holiday periods. It is most suitable for highly seasonal time series and historical data spanning several seasons. Prophet usually does a good job of handling outliers and resists lost data and changes in trends. It was made available as open-source software by Facebook's primary Data Science team.

Prophet is a piecewise linear or logistic growth curve-tipped additive regression model. It has a weekly seasonal component that is represented using dummy variables and a yearly seasonal component that is modeled using Fourier series. Therefore, Prophet is particularly helpful for datasets that contain a long time period of detailed historical observations (month or years) at hourly, daily, or weekly intervals, those that have multiple strong seasonalities, include known significant but irregular events, have missing data points or large outliers, or have non-linear growth trends that are approaching a limit. Prophet is also swift and accurate. It permits parameter adjustments and the creation of unique seasonality components, which could enhance forecasts. Additionally, it can deal with outliers and other data problems on its own. When a holiday or significant event may affect the

prediction, Prophet can adapt forecasting using the holiday feature. Automatic change points detection is accessible.

2.2 Metrics

In this research, after data cleaning, the Prophet algorithm was imported. Next, this study defines a train set and make the model be fit to the dataset. Then this study can output the predictions with lower limits and upper limits, which are also known as 'yhat_lower' and 'yhat_upper". Finally, the linear regression line chart with the predicted outcomes can be plotted. When carrying out the prediction, two typical metrics called MAPE and MAE are employed.

Mean Absolute percentage error (MAPE) measures the accuracy of a company's forecasting process. It shows the average accuracy between the expected amount and the actual amount by the absolute percentage error for each entry in the average data set:

$$\text{MAPE}=\frac{100\%}{\text{n}}\sum_{i=1}^{n}\left|\frac{\hat{y}_i}{y_i}-1\right| \quad (1)$$

Mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as the sum of absolute errors divided by the sample size.

$$\text{MAE}=\frac{100\%}{\text{n}}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (2)$$

3. Result and Discussion

Table 1. The concentration of NOx

Numble	ds	trend	yhat_lower	yhat_upper	trend_lower	trend_upper
1	2004-01-04	946.009417	790.067680	1104.635574	946.009417	946.009417
2	2004-01-11	943.188093	787.485341	1088.642583	943.188093	943.188093
3	2004-01-18	940.366770	776.806100	1094.341314	940.366770	940.366770
4	2004-02-08	931.902799	770.489706	1085.760666	931.902799	931.902799
5	2004-02-15	929.081475	777.884113	1100.201295	929.081475	929.081475

There are several types of air quality indexes in the dataset, including CO, C6H6, CO2, NOx. In this research, NOx was used as an example to plot its trends. Table 1 above shows the trend measurements of the concentration of NOx and the upper predictions as well as the lower predictions. Both two-line charts below are plotted in Figure 1. and Figure 2. according to the trending measurements, daily and weekly. They are both nearly smooth lines after the linear regression. It can be noticed clearly from the scatter plot that the figures fluctuate in a rather irregular way. Therefore, some errors cannot be avoided, which influences the accuracy of the result. It is acceptable because the environment surrounded the people is changeable and not strictly regular.

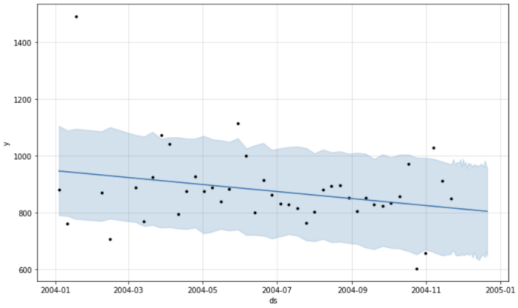

Fig 1. The daily prediction results.

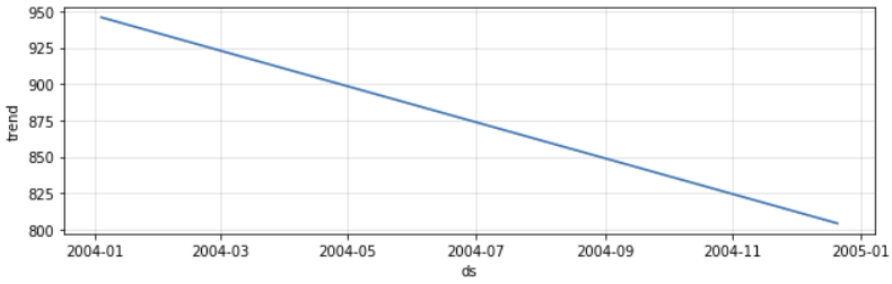

Fig 2. The weekly prediction results

Table 2. Comparison of the predictions and the realistic data

Numble	ds	yhat	yhat_lower	yhat_upper	y
1	2004-01-04	946.009417	790.067680	1104.635574	880.666667
2	2004-01-11	943.188093	787.485341	1088.642583	760.484990
3	2004-01-18	940.366770	776.806100	1094.341314	1490.333333
4	2004-02-08	931.902799	770.489706	1085.760666	869.108333
5	2004-02-15	929.081475	777.884113	1100.201295	706.395833

In the table, “yhat”, “yhat_lower”, “yhat_upper” refer to the predictions. And “y” refers to the real values. An auxiliary function was defined to make the comparisons, which helps making the evaluation on the model.

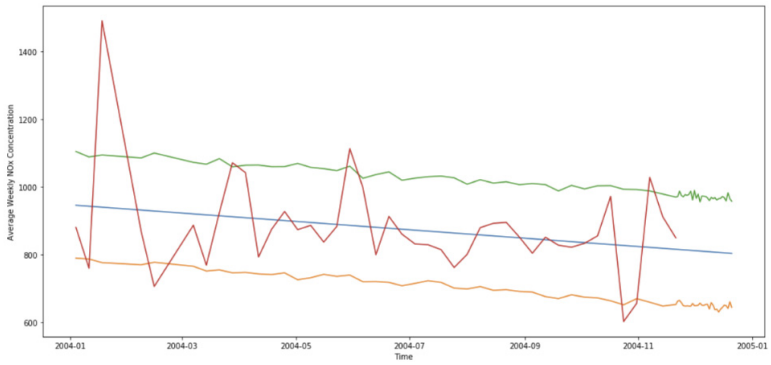

Fig 3. The prediction of the average weekly concentration of NOx

From this line chart shown in Figure 3, the upper and lower prediction made by the prophet algorithm can be observed clearly, which are almost linear. While the real values show in a polygonal chain with great uncertainty, the prediction cannot fit the actual result closely. The prediction plot can fit the trend of the original dataset, but cannot fit the turning points and drastic fluctuations closely, which causes the loss of accuracy.

Therefore, some improvements should be made. For example, on one hand, the prophet algorithm can be used to detect some potential changepoints and carry out some modification based on it. Also, the changepoint prior scale can be modified.

On the other hand, considering that the environment is changeable. Some great events may change the air quality index to a great extent, such as lightening the fireworks in the Spring Festival, which makes the concentration of CO2 and other gases climb greatly. The prophet algorithm can be used advanced to add some holiday effect, seasonal effect as well as considering some big events. All these modifications can improve the accuracy of this model and make the prediction more fit to the reality. In addition, some advanced machine learning algorithms can also be considered in this case for improving the performance in the future due to their satisfactory performance in many tasks [8-10].

4. Conclusion

In this research, a thorough and reliable model for predicting air quality is proposed to be realized. which can also assist the government in learning about regional environmental problems and, to some extent, in the development of useful policies by advising or cautioning individuals about their outdoor activities. Prophet algorithm and python are the main tools in this research. A dataset was employed in this research to train the model. According to the dataset, some prediction results about certain weekly air quality index were successfully made by the model. And the prediction results were output by the codes in the form of line charts. The accuracy of the model is acceptable and reliable. In the future, more specific issues can be discussed. For example, some drastic changes related to specific festivals and seasons should be considered to avoid the limitations posed by the dataset.

References

[1] WANG J S, WANG Y, ZHAO M X, et al. Application of ARIMA model in the prediction of air quality index in Suzhou. Journal of Public Health and Preventive Medicine, 2019, 30(2):18-20.

[2] Yang, S., and L. Zhao. Application of Random Forest Algorithm in Urban Air Quality Forecast. Stat. Decis 20, 2017, 83-86.

[3] Chang Tianjun, et al. Prediction of air quality index size based on Prophet-Stochastic Forest Optimization model, Environmental Pollution and Prevention, 41.07, 2019.

[4] S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical, Volume 129, Issue 2, 22 February 2008, Pages 750-757, ISSN 0925-4005.

[5] Gong, Feixiang, et al. Trend analysis of building power consumption based on prophet algorithm. 2020 Asia Energy and Electrical Engineering Symposium (AEEES). IEEE, 2020.

[6] Zunic, Emir, et al. Application of facebook's prophet algorithm for successful sales forecasting based on real-world data. arXiv preprint arXiv:2005.07575, 2020.

[7] Argue, C. J., et al. Robust Secretary and Prophet Algorithms for Packing Integer Programs. Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). Society for Industrial and Applied Mathematics, 2022.

[8] Yu, Q et al. Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV. China Communications 17.3, 2020, 46-57.

[9] Monil, Patel, et al. Customer Segmentation Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology (IJRASET) 8.6, 2020, 2104-2108.

[10] Kourou, Konstantina, et al. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal 13, 2015, 8-17.