Forecasting Carbon Dioxide Emissions in the United States using Machine Learning

Shreeya Namboori
x18128947

Abstract

Climate change has been realized as a major concern worldwide and people along with government are coming together to combat this worldwide issue to ensure that our future generation doesn't have to suffer. United States(U.S.) has been one of the top emitters of GHG (Green House Gas) emissions for a long time, although the emissions have been decreasing. Carbon dioxide(CO_2) is a major GHG gas that makes up 80% of the GHG emissions. Therefore the reduction of CO_2 emissions will help in reducing GHG emissions produced by the U.S. per year. The data for monthly C02 emission is taken from the U.S. Energy Information Administration (EIA). The analysis of the preprocessed data reveals three highly CO_2 emitting sectors that are Coal Electric Power Sector, Natural Gas Electric Power Sector, and Total Energy Electric Power Sector. The emissions from these three sectors are used by ARIMA, SVM, SVM-PSO, and Prophet models for forecasting CO_2 emissions. The performance of the models are compared with each other using RMSE, MAE and MAPE metric and the results show that the Prophet model outperforms all the other models in forecasting CO_2 emissions from different sectors. Therefore the Prophet model is used for forecasting future emissions for the next 36 months for all three sectors.

Keywords : SVM, SVM-PSO, ARIMA, PROPHET, GHG , Carbon dioxide

1 Introduction

1.1 Background

Global warming is a worldwide problem that needs to be addressed as it is affecting the environment around us in a negative way. The droughts, melting of glaciers and the rise in sea level are all the consequences of global warming that has threatened the survival of all the living beings. The reduction of Greenhouse gas emissions (GHG) has become the primary focus of many countries as they have realized that the adverse climatic conditions caused by global warming are the consequence of these emissions. Paris agreement signed under UNFCCC (United Nations Framework Convention on Climate Change) focuses on reducing GHG emissions and was signed in 2016 in Paris, France. This agreement aims to keep the increase in global average temperature below 2 degrees and to further aim at reducing it to 1.5 degrees so that the impacts encountered due to climate change can be reduced.

Carbon dioxide is one of the most important GHG gas and the main source for these emissions are human activities, vehicles, fossil fuel combustion like coal, oil and natural gas and different industrial sectors. The rate at which CO_2 is being produced by various human activities is far greater than the rate at which it is being absorbed, primarily due to the decreasing forest cover. The United States alone accounts for one-fourth of the total GHG emissions in the world. According to a report the total GHG emission in the United States in 2016 composed of 81% of carbon dioxide[1] emissions while in 2017[2] it was 82% of the total GHG emission. Carbon dioxide plays a major role in the increase of GHG emission therefore policies that can help in limiting these CO_2 emissions are required. An independent research firm the Rhodium Group tracks the CO_2 emissions and has reported an increase in the emissions by 3.4% in 2018[3] after years of decline which is a concerning issue. The forecasting of carbon dioxide emissions can help in identifying sectors that need monitoring or new policies for reducing GHG emissions.

1.2 Research Question

The research question for this research project is:
RQ: "How efficiently can the machine learning algorithms perform time series forecasting of carbon dioxide emissions from different sectors in the United States?"

1.3 Research Objectives

The U.S. is one of the top emitters of carbon dioxide as can be seen from Figure 1, the U.S. was the second-highest CO_2 emitting country in 2017 [4] fter China. Figure 2 shows the yearly GHG emission by the United States in the past years[5]. Although the emissions have decreased in the past there have been reports of increased GHG emission in 2018.

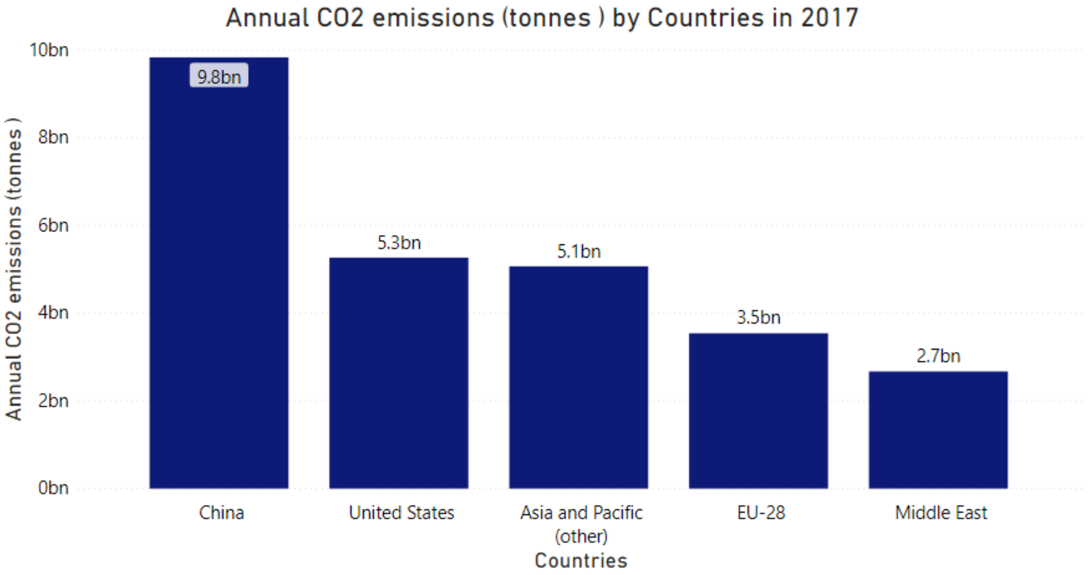


Figure 1: Highest CO_2 emitting countries

[1] https://www.sciencenewsforstudents.org/article/explainer-co2-and-other-greenhouse-gases

[2] https://www.c2es.org/content/u-s-emissions

[3] https://www.npr.org/2019/01/08/683258294/u-s-carbon-dioxide-emissions-are-once-again-on-the-rise?t=1575314288216

[4] https://ourworldindata.org/CO_2 -and-other-greenhouse-gas-emissions

[5] https://stats.oecd.org/Index.aspx?DataSetCode=AIR_GHG#

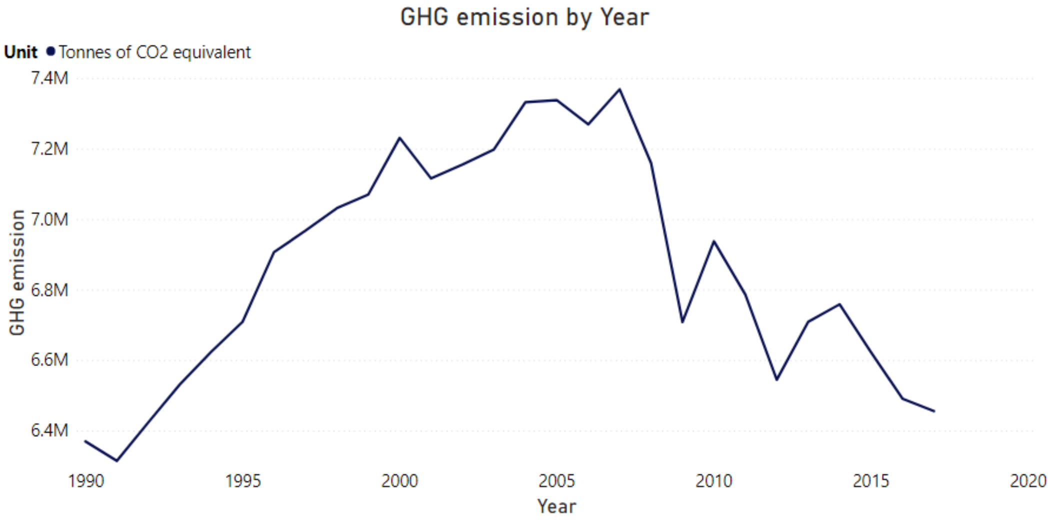


Figure 2: U.S. yearly GHG emission

The objectives this research aims to achieve are:

- Presenting the sectors with highest $CO2$ emission.
- Comparing the performance of all the models to find the best forecasting model.
- Forecasting the future CO_2 emissions using the best model.
- Publishing the information so that appropriate policies can be implemented.

2 Related Work

2.1 SVM

SVM has gained popularity in the classification and regression problems as it's based on the principal of structural risk minimization (SRM) which prevents overfitting of the data (Chuentawat & Kan-Ngan 2019). Saleh et al. (2016) used Support Vector Regression (SVR) algorithm of SVM model for predicting Carbon dioxide (CO_2) emissions from energy and coal consumption data. The C and epsilon values for SVM was chosen using the trial and error method and the values having the lowest RMSE was chosen for the prediction. For a C value of 0.1 and epsilon value of 0 the RMSE error for prediction was 0.004. Du et al. (2017) discusses the importance of data preprocessing and normalization to get accurate results and further states the advantages of PSO over GA as it is easier to implement, understand and need very few parameters for adjusting. They used SVM optimized by PSO-SVM for predicting precipitation. The metrological data for a year from Nanjing Station was collected and preprocessed and it's dimensionality was removed using PCA (Principal Component Analysis) before modeling. The performance of SVM-PSO was compared with SVM, SVM-GA(Genetic Algorithm) and Ant Colony SVM (ACSVM) using MSE, SVM-PSO demonstrated the best prediction as it had the lowest MSE value. Chuentawat & Kan-Ngan (2019) used SVM-GA for forecasting PM 2.5 emission using data from UCI Machine Learning Repository for Beijing, China to compare the performance of univariate and multivariate time series. Univariate time series demonstrated a better performance as it had lower MAPE and RMSE values compared to the multivariate time series for the same C and epsilon values. Tang & Zhou (2015) highlights the problem of overfitting a data when using ANN (Artificial Neural Network) therefore emphasizing the advantage of using SVM. These researchers used PSO-SVM for short-term inflation rate forecast by taking the CPI (Consumer Price Index) data from National Bureau of Statistics of China. The proposed model was compared with BPNN (Back-Propagation Neural Network), SVM and SVM-GA and the results showed that SVM-PSO had the best forecasting as it's MSE was lowest. Sudheer et al. (2014) discusses how important it is to find optimal parameters for SVM to get an efficient performance which led them to use SVM-PSO for forecasting streamflow of Swan River and St. Regis River. SVM-PSO model's performance was compared with ARMA (Auto Regressive Moving Average) and ANN using correlation coefficient and normalized mean squared error(NMSE). SVM-PSO exhibits a better prediction as it has the highest R and lowest NMSE value. Li et al. (2009) proposed a SVR model that predicts the Chinese real estate price using the quarterly data from 1998 to 2008. A total of five indicators were given as input to the SVR model and real estate price was used as the output. The performance of the proposed model was compared to BPNN using the metrics RMSE, MAE and MAPE and the results exhibited the superiority of the proposed SVR model for real estate price prediction. Kunda & Phiri (2017) forecasted CO_2 emissions in Zambia from 2017 to 2021 using SVM model from the dataset taken from World Bank from 1960 to 2016. An open source data mining tool WEKA (Waikato Environment for Knowledge Analysis) was used in this research and SMOreg (Sequential minimal optimization regression) algorithm was used for performing trend analysis on time series data. The future forecast result presented an increase in CO_2 emissions for the years 2017 to 2021. Zhao et al. (2010) predicted mercury emission from the combustion of coal using SVM from the dataset obtained from of a coal using power plant in U.S. . The proposed model was compared with BPNN, Multiple nonlinear regression (MNR) and GRNN (Generalized Regression Neural Network) using evaluation metric of MSE. SVM demonstrated a better performance than all the other models by having the lowest MSE. Wang et al. (2009) discusses how SVM overcomes the shortcomings of an ANN based forecasting model and the importance of PSO algorithm in selecting the appropriate parameters.Therefore they had proposed a wavelet transformed PSO-SVM (W-PSO-SVM) for forecasting gas concentrations to get more efficient time series forecasting and compared it's performance with other models like ANN, SVM and W-SVM . W-PSO-SVM exhibited higher accuracy in forecasting by having lower residual mean square, lower maximum error percentage, lower minimum error percentage and lower average error percentage. Gu et al. (2011) also realised the need of selecting appropriate parameters for SVM without consuming too much time therefore they proposed GA (Genetic Algorithm) for optimizing SVM since it consumes less time. SVM-GA was used for predicting house prices in China and the performance of this model was compared with GM model using absolute relative error (ABE) metric. The results presented lowest ABE value for GA-SVM than GM proving it's accuracy in predicting the house prices.

2.2 ARIMA

ARIMA is the most popular model used for time series forecasting used by the researchers and choosing an appropriate ARIMA model is essential to get an efficient time series forecast. Sen et al. (2016) had the objective of finding the best ARIMA model for forecasting the energy consumption and GHG emission from the data taken from an Indian pig iron manufacturing company. The different ARIMA models chosen are compared using MPE(mean percentage error), MAE, ME, MAPE, RMSE, AIC (Akaike Information criterion) and SBIC (Schwarz Bayesian Information Criterion) and the model having the lowest value for all the evaluation metrics is considered a better performer. Lotfalipour et al. (2013) used Grey Model (GM) and ARIMA model for forecasting $C0_2$ emissions in Iran using the data from British Petroleum. GM had a better forecasting performance as it had lower RMSE, MAE and MAPE compared to the ARIMA model. The future prediction done using GM showed a 66 % increase in emission for the year 2020. Palomares-Salas et al. (2009) compared the performance of ARIMA and BPNN for forecasting wind speed. The best ARIMA model was identified and the performance of ARIMA and BPNN was found to be almost the same but ARIMA was faster in forecasting. Amin & Hoque (2019) used ARIMA and SVM short term load forecasting to ensure that the operation of the power system remains stable and reliable. The performance of SVM was better than the ARIMA models as it had lower RMSE and MAPE values in forecasting short term load. In today's world employment is a major concern for all the people and realizing this issue Xiaoguo & Yuejing (2009) used ARIMA for forecasting employment by taking data from the China Human Resource Market for supply demand. After making the time series stationary and selecting an appropriate ARIMA model the quarterly forecast for the year 2008 is done. which shows that the highest recruitment would in in third quarter of the year. They further discuss that ARIMA is not suitable for long term forecasting as the errors will become larger.

2.3 Prophet

Borowik et al. (2019) used the Prophet model by FB for crime time series forecasting in Poland. The researchers discussed the challenges while dealing with the crime data which usually is either incomplete or is full of errors therefore they used ETL (Extract Transform Load) as a preprocessing step to properly manage, clean and store the data. This research forecasted the frequency for six different crimes, police interventions , hooliganism, road trafc offence, burglary and theft, detention and other criminal offenses. The metric MAPE was used for predicting the accuracy of the time series model and the results showed that there was approximately 10% forecasting errors for road traffic crimes and police interventions, around 40% forecasting errors for hooliganism and around 20% forecasting errors for the rest of the crimes. Yenidogan et al. (2018) used both ARIMA and Prophet to forecast Bitcoin to find the most accurately forecasting model. The dataset used from Kaggle is splitted in three-k folds for training, testing and validation. Feature selection was done using correlation matrix where a strong relation was found with GBP (British pound sterling), EUR (Euro), and JPY (Japanese Yen) currencies. The forecast results from both the models were compared and it was found that Prophet had an accuracy of about 94.5% while ARIMA had an accuracy of 68%. The Prophet model performed better than the ARIMA in time series forecasting of Bitcoin. Even though Prophet has a good performance in time series forecasting, there are not many research papers that have employed this model.

2.4 Neural Networks

Neural networks has been popularly used for forecasting by many researchers due to their flexibility as they make fewer assumptions about the data. Mason et al. (2018) used the RNN (Recurrent Neural Network) that was optimized using the CMA-ES (Covariance matrix adaptation evolutionary strategy) algorithm for forecasting energy demand, generation of wind energy and CO_2 emissions in Ireland. The comparison of the proposed model by researchers was done with Differential Evolution (DE), Random walk forecasting (RWF), Moving Average (MA), Linear Regression (LR), Particle swarm optimization(PSO), Back-Propagation (BP) and CMA-ES algorithms using evaluation metrics MAPE, MSE, MAE, and RMSE and it was observed that CMA-ES had a good performance in forecasting wind energy and energy demands but was not able to forecast CO_2 emissions efficiently as the testing data contained smaller values which was absent in the training data. Sheta et al. (2015) used data from natural gas, coal, global oil, and primary energy consumption to predict the CO_2 emission using Product Unit Neural Network (PUNN) and Neural Network AutoRegressive eXternal (NNARX) models. It was observed that PUNN model was able to predict more efficiently as it had lower RMSE, MAE and high VAF (variance accounted for) value. Kai et al. (2011) states that wavelet networks are quickly able to converge and with just just a few training iterations they are able to find the global optima,therefore they have proposed a wavelet network for predicting carbon flux. The performance of the wavelet network model was compared with BPNN (Backpropagation neural network) and SVM using normalized square root of mean square error (NSRMSE) and correlation coefficient and it was observed that the proposed model outperformed both BPNN and SVM in prediction of carbon flux by having low NSRMSE and high correlation coefficient. Lu et al. (2017) targeted the transportation sector and forecasted the carbon emissions produced specifically by this sector using a three layer perceptron neural network using dataset that contained information about taxi trajectories, Point of Interest data and road network data from Auto Navi MAP for Zuhai city of China. After pre-processing of data feature selection was done using correlation matrix. The model's performance was compared with Deep Belief networks, Linear Regression, Gaussian Naive Bayes, and Stacked Denoising Autoencoder based on the prediction accuracy. The proposed model had the highest prediction accuracy of 90.86%. Kingsley Appiah et al. (2018) used the ability of ANN to recognize patterns and predict the future by learning from the historic data to predict carbon emissions. The carbon emissions of India, Africa, China and Brazil were predicted using a feed forward ANN model that was optimized using LM (Levenberg-Marquardt) algorithm on the data taken from FAOSTAT (Food and Agriculture Organization Corporate Statistical Database) database. The model exhibited a good performance by having a lower MSE and higher accuracy.Li et al. (2010) performed time series of CO2 emissions in China using a RBF neural network from MATLAB due to its advantage of fast learning and good generalization performance. The model had high predictive accuracy in forecasting the emissions and low absolute error. Zhang et al. (2018) proposed a time series wavelet transform LSTM model for forecasting hourly vehicle emission of CO (Carbon Monoxide), HC (Hydrocarbons) and NO (Nitric oxide) using the metrological data from China Meteorological Administration from May to December 2017. The performance of this model was compared with ARIMA using RMSE and MAE performance metrics and from the results it was observed that wavelet transform LSTM had better forecasting performance.

2.5 Other forecasting models

GM models have many advantage like they require little data to do forecasting and have higher accuracy compared to other models which is why they have been used by many researchers for forecasting (Ayvaz et al. 2017)(Wang & Dang 2013). Ayvaz et al. (2017) forecated the CO_2 emissions that are produced due to energy consumption in Turkey, Europe and Eurasia using six types of Discrete Grey Models(DGM) namely DGM, non-homogeneous DGM , optimized DGM, non-homogeneous DGM with rolling mechanism , DGM with rolling mechanism and optimized DGM with rolling mechanism. The data for the research was taken from British Petroleum and the models were compared with each other using RMSE, MSE and MAPE values. It was found that DGM had better prediction for Eurasia and Europe while non homogeneous DGM had better prediction for Turkey. The future forecast done from the best forecasting models from the year 2015 to 2030 revealed that the CO_2 emissions decreased in both Europe and Eurasia while Turkey faced an increase in CO_2 emissions. Philibert et al. (2013) used Random Forest(RF) model for predicting the emission of nitrous oxide (N20) using dataset from the experiments of Stehfest and Bouwman (2006) research paper. The model was compared with Stehfest and Bouwman(2006) linear model and a non linear model that used exponential function using RMSE and the results confirmed that RF outperformed all the other models by having lower RMSE. Wang & Dang (2013) used a new prediction method for GM(1,1) model to forecast carbon emission in Jiangsu province of China. The new model is compared with the traditional GM(1,1) model and it is found that the new model has better prediction accuracy and small relative error compared to the traditional one. The future forecast is done from the proposed model which shows an increase in carbon emission that reaches 53316.14 ten thousand tons in 2020. Sun & Sun (2017) used PC-RELM (Principal Component- Regularized Extreme Learning Machine) to forecast CO_2 produced due to energy consumption using data from China Statistical Yearbook. The model presents the best performance in forecasting the emissions when compared with ELM, REML, GM(1,1), BPNN and logistic model using the metrics median absolute percentage error (MdAPE) and maximum absolute percentage error (MaxAPE).

Even with the introduction of machine learning models in forecasting statistical methods keeps on being used by the researchers for time series analysis and gives good results(Hosseini et al. 2019, Akcan et al. 2018). Akcan et al. (2018) forecasted GHG emission in Turkey using statistical methods such as moving average, exponential smoothening method and exponential smoothening with trend method using data from Turkish Statistical Institute. The evaluation metrics used in this research were RMSE, MAE and ME (mean error). Although all the models had a good performance in forecasting but EST method was preferred as it had lowest errors. Hosseini et al. (2019) used multiple linear regression (MLR) and multiple polynomial regression (MPR)for forecasting CO_2 emissions in Iran using data from the World Bank. The forecasting was done for two scenarios one where business is carried as usual (BAU) and other where implementation of SDP(Sixth Development Plan) is done successfully. The forecast presented by both the models for 2015 to 2030 showed that in the BAU scenario Iran would not be able to meet the CO_2 emission reduction while in the SDP implemented scenario the target CO_2 reduction would be achieved as promised in the Paris Agreement.

3 Methodology

3.1 Introduction

The process of extracting useful information from the huge amount of data to gain some insights is known as Data mining. KDD(Knowledge Discovery in Databases), CRISP-DM (Cross-industry standard process for data mining), and SEMMA (Sample, Explore, Modify, Model, and Assess) are the three most popular data mining methodologies. Each methodology differs from each other in terms of the number of steps they have. CRISP-DM methodology has been used for this research as it provides flexibility and structure to the project Wirth & Hipp (1995), is easier to comprehend and has well designed steps that covers all the aspects of the project.

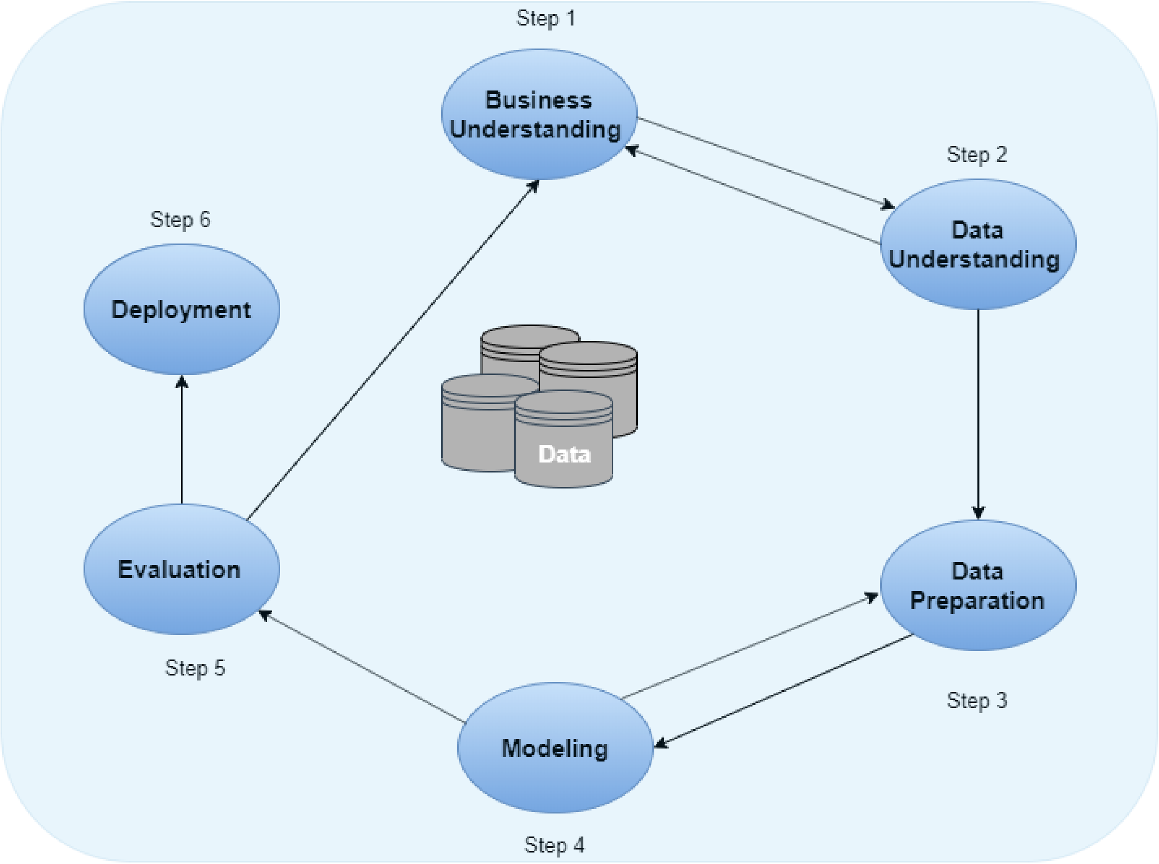


Figure 3: CRISP-DM Methodology

3.2 CRISP-DM Overview

CRISP-DM is one of the most popularly used data mining methodologies that was found by SPSS in 1996 (Bošnjak et al. 2009). CRISP-DM has six phases or steps (as shown in Figure 3[6]) that covers the entire life-cycle of the data mining project. The six phases of CRISP-DM are: Business understanding, data understanding, data preparation, modeling, evaluation and deployment.

[6] https://www.draw.io

3.2.1 Business Understanding

It is the initial and most crucial phase of the project where the objectives of the research are identified and highlighted and after that, a business plan is formulated to achieve those objectives. The combat against climate change has already started with people joining forces to find different solutions that can help in improving the environmental condition. There is an immediate need for addressing the issue of global warming by adopting sustainable living methods. United States is one of the highest GHG emission producing country and therefore it needs to reduce its GHG emission along with other high GHG emission producing countries to prevent the endangerment of all the living beings due to increasing global warming. The objective of the research project is to forecast the CO_2 emissions by the United States from different sectors so that the sectors with increasing CO_2 emission can be identified and this information when published can help in implementing policies to control the emissions and meet the target reduction of 25% to 28 % by 2025 as promised in the Paris Agreement. The business plan formulated is to use four machine learning models for forecasting the CO_2 emissions and choose the best model with the highest forecasting accuracy to do future forecasts.

3.2.2 Data Understanding

This is the second phase of the CRISP-DM where the data that has been collected is thoroughly analyzed to gain an understanding and insight about the data. The dataset for the research project is collected from U.S. Energy Information Administration (EIA) website that has made the dataset public for reuse. The dataset contains monthly CO_2 emissions from January 1973 till August 2019 from nine sectors in million metric tons of carbon dioxide unit. The dataset contains six columns and 5445 rows which even includes the NA values. After careful observation, it is found that emission values of the Non-Biomass Waste Electric Power Sector and the Geothermal Energy Electric Power Sector are not available from January 1973 till December 1988.

3.2.3 Data Preparation

The third phase of CRISP-DM consists of cleaning the dataset and for this purpose, the RStudio tool was used. The dataset had 416 NA values that were omitted as they were from the Non-Biomass Waste Electric Power Sector and the Geothermal Energy Electric Power Sector from January 1973 till December 1988. The fourth column in the dataset had index numbers for sectors therefore it was removed and the column names for three columns were modified. The year and months were separated using a separator by using substring function and the thirteenth month of the year that contained the sum of emissions for all the months in a year was also removed using the gsub function. The cleaned dataset was finally written as a CSV to be used for modeling.

3.2.4 Modeling

The fourth phase of CRISP-DM is modeling where the data that is thoroughly cleaned is used by the selected models to give the desired output. The models selected for the research are SVM(Support Vector Machine), SVM-PSO(Support Vector Machine with Particle Swarm Optimization), Prophet, ARIMA and LSTM (Long Short Term Memory). These machine learning models are trained on 90% of data to give a forecast on the test data which is 10% of the remaining data.

i **SVM**

SVM is a supervised machine learning algorithm that was found by Vladimir Vapnik and his teammates in 1992 and it has been popular in classification and regression problems because of its simplicity and the fact that it overcomes some of the shortcomings of ANN(Artificial Neural Network) like overfitting. This model is based on the principle of structural risk minimization making it have a good generalization ability (Tang & Zhou 2015). While SVM is used for classification problems, the SVR (Support Vector Regression) algorithm in SVM is used for regression problems.

ii **SVR-PSO**

When using SVM a great challenge faced by many people is the appropriate selection of the SVM parameters like C and epsilon. If the C parameter is large there is a high penalty resulting in overfitting and if the value is too small it results in underfitting. The epsilon value gives a margin of tolerance, if it's zero there is overfitting and if it's too large the results would not be satisfactory. To find the optimal parameters C and epsilon in SVM, the PSO algorithm is used. PSO algorithm was found by Eberhart and Kennedy in 1995 and it can deal with continuous optimization problems while supporting multi-point search (Du et al. 2017).

iii **Prophet**

Prophet is a relatively new forecasting model that was developed by Facebook's data science team and it is available as open-source software in both R and Python (Yenidogan et al. 2018). Unlike ARIMA the missing values in Prophet need not be interpolated as it can deal with the missing data while giving higher prediction accuracy. It is optimized for business forecasting scenarios that were encountered by Facebook having hourly, daily, weekly and monthly frequency. One of the advantages of Prophet is that it doesn't require prior knowledge therefore it is easier to use and can automatically find the seasonal trends residing in the data while giving efficient results.

iv **ARIMA**

ARIMA(Auto Regressive Integrated Moving Average) is one of the most popularly used time series forecasting models. It gives the future forecast based on the past values of the time series. An ARIMA model requires three input parameters known as the order of the AR term (p), the order of the MA term (q) and the number of differencing (d). The prerequisite for using the ARIMA model is that the time series should be stationary (Sen et al. 2016).

3.2.5 Evaluation

In the fifth phase of CRISP-DM, the performance of all the models used in the research is evaluated. The performance metrics used for this research are RMSE (Root Mean square Error), MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) based on which a model's performance is compared with the other models to identify the best one. The lower value of RMSE, MAPE, and MAE is desirable as it means that the model is more accurate in forecasting.

3.2.6 Deployment

According to Wirth & Hipp (1995), the simplicity and complexity of the deployment depend on the requirement. In this research, this phase consists of the presentation and submission of a well-established research document, a configuration manual, and ICT solutions document.

3.3 Conclusion

The rationale for choosing CRISP-DM as the research project methodology is justified and discussed briefly along with the six phases present in the CRISP-DM in correspondence to the research.

4 Implementation

4.1 Introduction

This section briefly discusses the implementation flow and the model implementation done to achieve the objectives of the research project. Figure 4 shows the implementation flow followed in the project to achieve the future forecasting results.

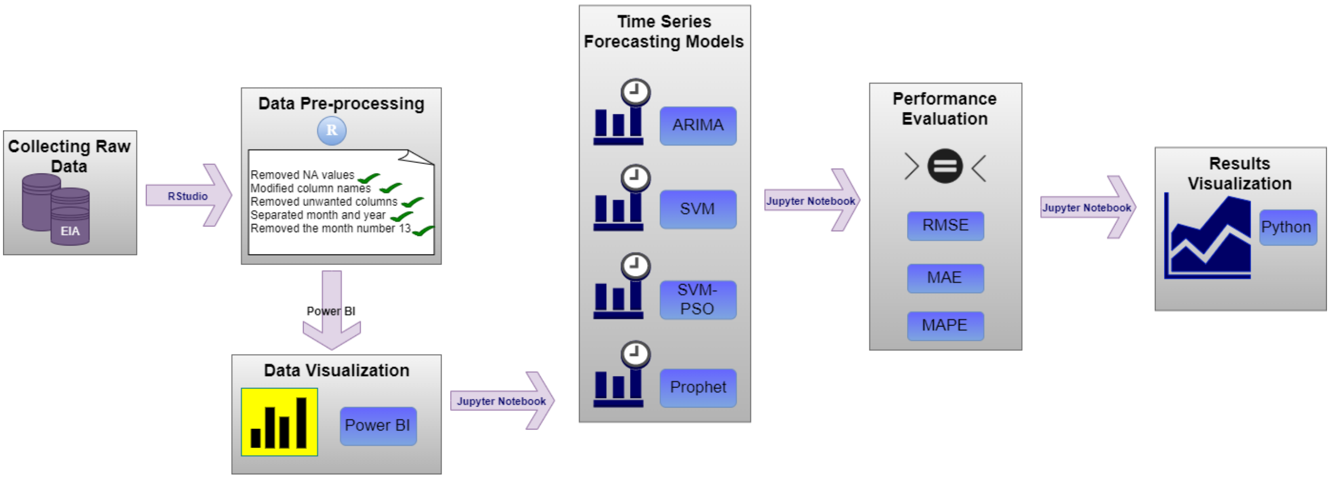
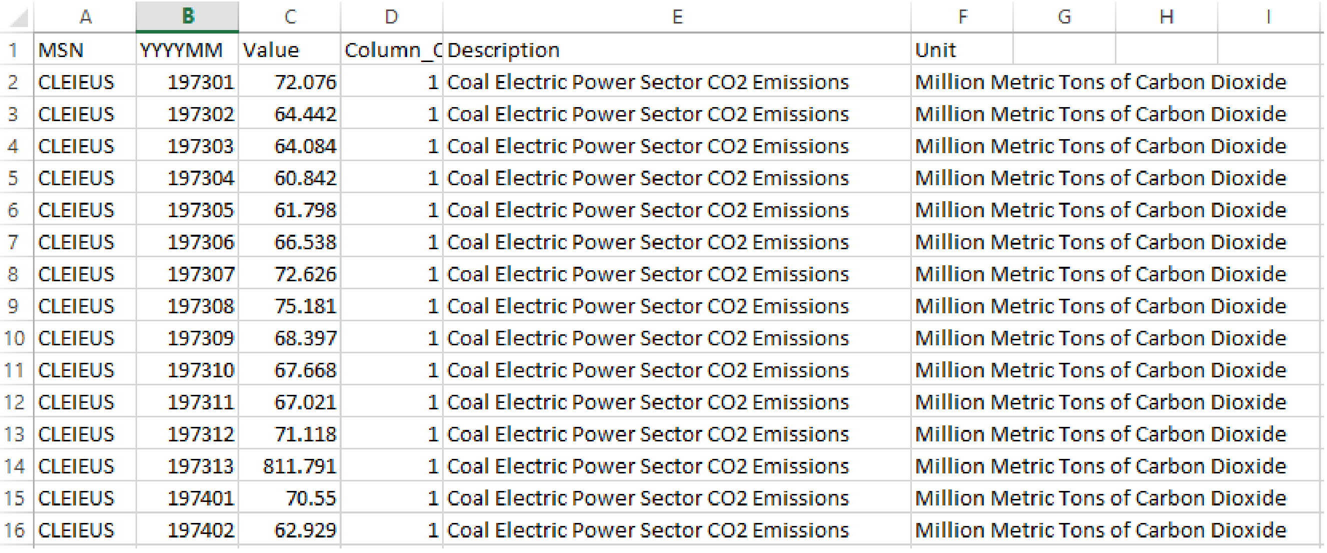


Figure 4: Implementation flow diagram

4.2 Acquiring Data

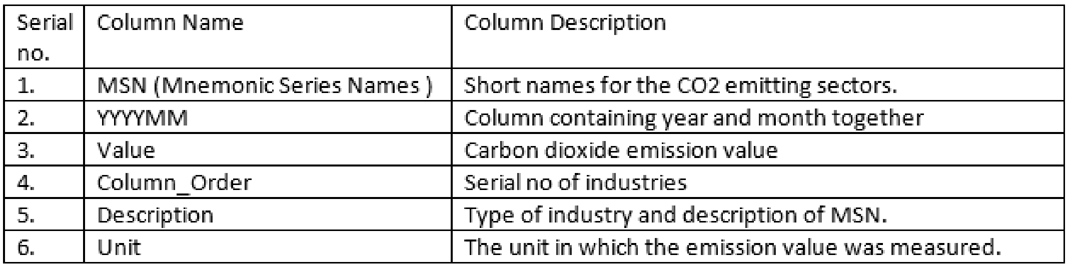
The dataset is downloaded from the U.S. Energy Information Administration (EIA)[7] that contains the monthly CO_2 emissions from January 1973 to July 2019. This raw dataset contains 5445 rows which include the emissions from nine sectors and it has six columns. The brief description of all the columns is given in the table present in Figure 6 while all the sectors present in the dataset are mentioned in the table present in Figure 7.

[7] https://www.eia.gov/totalenergy/data/browser/?tbl=T11.06#/?f=M



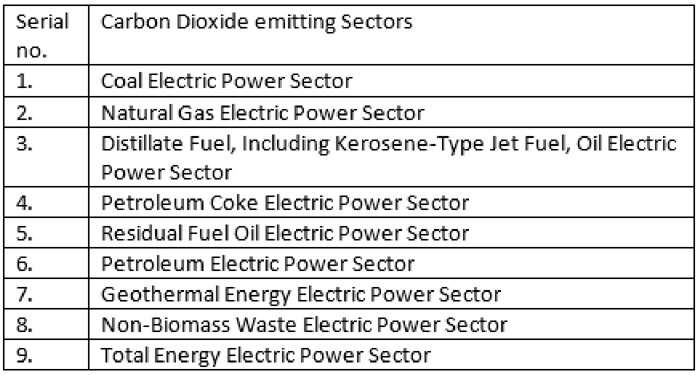
	A	B	C	D	E	F	G	H	I
1	MSN	YYYYMM	Value	Column_C	Description	Unit			
2	CLEIEUS	197301	72.076	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
3	CLEIEUS	197302	64.442	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
4	CLEIEUS	197303	64.084	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
5	CLEIEUS	197304	60.842	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
6	CLEIEUS	197305	61.798	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
7	CLEIEUS	197306	66.538	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
8	CLEIEUS	197307	72.626	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
9	CLEIEUS	197308	75.181	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
10	CLEIEUS	197309	68.397	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
11	CLEIEUS	197310	67.668	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
12	CLEIEUS	197311	67.021	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
13	CLEIEUS	197312	71.118	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
14	CLEIEUS	197313	811.791	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
15	CLEIEUS	197401	70.55	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			
16	CLEIEUS	197402	62.929	1	Coal Electric Power Sector CO2 Emissions	Million Metric Tons of Carbon Dioxide			

Figure 5: Raw Data



Serial no.	Column Name	Column Description
1.	MSN (Mnemonic Series Names)	Short names for the CO2 emitting sectors.
2.	YYYYMM	Column containing year and month together
3.	Value	Carbon dioxide emission value
4.	Column_Order	Serial no of industries
5.	Description	Type of industry and description of MSN.
6.	Unit	The unit in which the emission value was measured.

Figure 6: Column Description



Serial no.	Carbon Dioxide emitting Sectors
1.	Coal Electric Power Sector
2.	Natural Gas Electric Power Sector
3.	Distillate Fuel, Including Kerosene-Type Jet Fuel, Oil Electric Power Sector
4.	Petroleum Coke Electric Power Sector
5.	Residual Fuel Oil Electric Power Sector
6.	Petroleum Electric Power Sector
7.	Geothermal Energy Electric Power Sector
8.	Non-Biomass Waste Electric Power Sector
9.	Total Energy Electric Power Sector

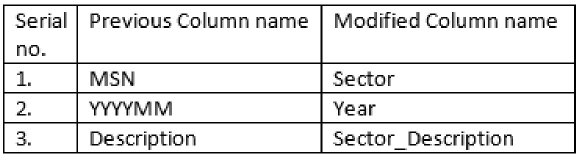
Figure 7: CO_2 emitting Sectors

4.3 Data Pre-processing

The raw data cannot be directly used for training the models as it could contain impurities like missing values (NA), special characters or wrong values, therefore, the acquired data is always pre-processed to remove the data impurities. RStudio which is an open source software, was used for cleaning the dataset for the research. The steps followed for cleaning the dataset are defined below:

i The column ’Column_Order’ which is column no. 4 in the dataset was removed as it contained index number for the sectors.

ii The column names YYYYMM, MSN and Description were modified as shown in Figure 8.



Serial no.	Previous Column name	Modified Column name
1.	MSN	Sector
2.	YYYYMM	Year
3.	Description	Sector_Description

Figure 8: Modified column names

iii There were 416 NA values in the dataset which were omitted since the CO_2 emission data for the Geothermal Energy Electric Power Sector and the Non-Biomass Waste Electric Power Sector are not present from January 1973 to December 1988.

iv The year and month were separated from each other by slash ’/’ using the substring function.

v Each year had a 13th month which is the sum of all the CO_2 emissions present in 12 months of a year. Since it was not required in time series analysis, it was removed during the cleaning by using gsub function where the year was converted to NA.

vi The new 382 NA values introduced in the dataset are then omitted from the data and the cleaned data is written down as CSV to be used by the models.

All the modifications have been performed on the duplicate data leaving the original dataset intact. To gain a better understanding of the data it was plotted in the Power BI tool and three sectors were selected for forecasting CO_2 emissions.The sectors chosen for forecasting are the Coal Electric Power Sector, the Natural Gas Electric Power Sector, and the Total Energy Electric Power Sector since they are the highest CO_2 emitting sectors as seen from the graph in Figure 9. Although the emissions from Coal Electric Power Sector and Total Energy Electric Power Sector have been decreasing, the emissions from Natural Gas Electric Power Sector has been rising over the years. The first objective of the research to present the the highest CO_2 emitting sectors is achieved here.

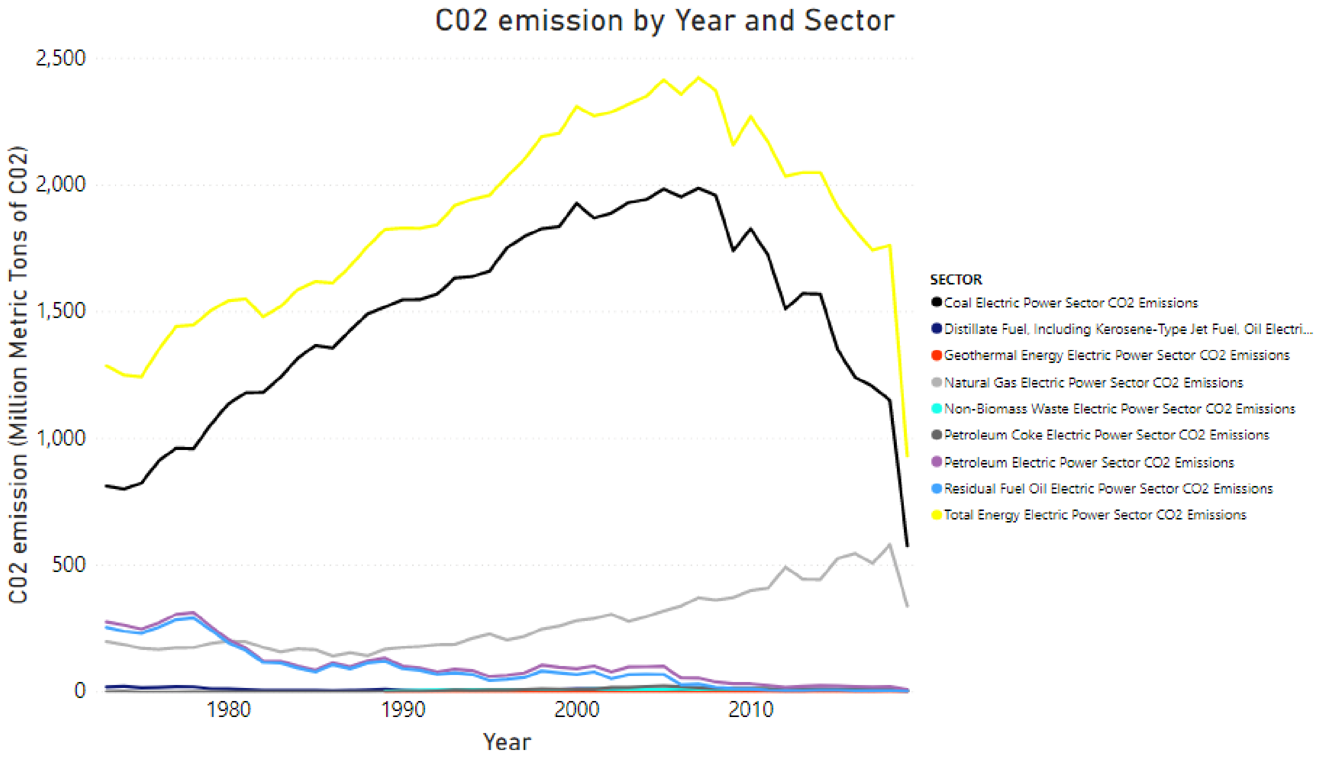


Figure 9: CO_2 Emission from all Sectors in U.S.

4.4 Model Construction

The code for model construction is written in Python programming language using Jupyter Notebook by Anaconda. The common libraries used by all the models are scikit-learn or sklearn, numpy, pandas, and matplotlib library. The data for each of the models are divided into the ratio of 90:10 for training and testing.

4.4.1 ARIMA

The data is first loaded in Jupyter Notebook and the year column is parsed using a parser function into the datetime format. The stationarity of time series is checked using test_stationarity function that checks if a time series is stationary by plotting the rolling mean and performing the Dickey-Fuller test on the time series. If the rolling mean is straight and Dickey fuller's value is less than 0.05 then the null hypothesis can be rejected and the time series is assumed to be stationary. The time series is made stationary by converting the actual value to log and subtracting the obtained log time series with its own log shifted version using the .shift() function. The number of times subtraction is done to make the time series stationary becomes the d (number of differencing). Once it is confirmed that the time series is stationary using the test_stationarity function, the ACF(partial Autocorrelation Function) and PACF(Partial Autocorrelation Function) plots are plotted to get the value for p from PACF plot and q from ACF plot. The ARIMA model of order p,d,q (ARIMA(p,d,q)) is then trained and fitted using .fit() function. The model fit is checked by using the .fittedvalues function in plots. The .forecast() function is used for forecasting the values till the given timesteps and these values along with test values are converted to actual numbers using the np.exp() function as they are in the log format. Test and prediction values are plotted against each other and the RMSE, MAE, and MAPE values are calculated for them.

4.4.2 SVM

SVR model and MinMaxScaler for feature scaling are imported from the sklearn library. A dataframe is created that contains index numbers along with the emission value. The index and the emission values go to two separate lists creating two sets for training and two sets for testing (one containing the index number and the other containing the emission value). Data normalization is done using MinMaxScaler and then it is transformed using .fit_transform method. After trying out many kernels like linear, sigmoid and rbf, the rbf kernel was chosen as the best kernel for forecasting. The model with the rbf kernel is then trained and fitted using .fit() function. The .predict() function is used to get the SVR prediction for the given test index. The predicted and test values are converted to their original format using the function .inverse_transform and these values are then plotted against each other and used for getting RMSE, MAE and MAPE metric values.

4.4.3 SVM-PSO

The particle swarm optimization algorithm is imported from the pyswarm library and KFold is imported from the sklearn library. A KFold of 10 splits is defined for creating 10 folds from the data where 9 folds would be used for training and each fold would be used once as a validation using the .split() function and this data is then normalized using MinMaxScalar and transformed using .fit_transform method. The model is then trained and fitted using .fit() function. A function svrPso is created for finding the optimized C and epsilon parameters for the SVR model with rbf kernel and lower bound and upper bound values for the C and epsilon are initialized in the main function from where the PSO algorithm is called. A calsMAPE function is created that gets called for each epsilon and C pair values returning MAPE values. Once the optimization parameters are found their value is passed to another function that uses SVR with rbf kernel with the optimal C and epsilon values to forecast the emissions.

4.4.4 Prophet

The Prophet model is imported from fbprophet in Python. The year column is parsed to datetime using a parser function and the unnecessary columns are removed from the data. The column name year is changed to 'ds' and value is changed to 'y' so that the data can be used by the Prophet model for forecasting. The frequency of forecast by the model is set to M which means monthly forecast. A dataframe containing future dates is made using .make_future_dataframe() function and the output of this function is given as input to the .predict() function to get the emission values of future dates.

4.5 Conclusion

The Implementation section gives a brief overview of the implementation flow, the data cleaning process, the steps followed for constructing the models, the libraries used, and the functions used for getting the forecast from the models. The references for model code was taken from the works of Brownlee (2017), Vincent (2017), Dabakoglu (2019), Etienne (2019), Singh (2016) and Nguyen (2019).

5 Evaluation and Results

5.1 Introduction

In this section, the performance of a model in forecasting the CO_2 emission values efficiently is compared with the other models based on the evaluation metrics RMSE, MAE and MAPE (Lotfalipour et al. 2013).

RMSE: It is the standard deviation of the errors that occur during CO_2 emissions forecasting. This metric has more sensitivity to the outliers present in the data.

MAE: It is the summed difference between the actual and the forecasted value which is divided by the total number of observations. This metric gives the absolute average difference between the two values.

MAPE: It is a metric that gives the error in forecasting as percentage and is a statistical measure to see how accurate the forecast is by the model. A smaller value of MAPE is desirable as it shows that the model has higher predicting power.

The equations for all the evaluation metrics are given below (Lotfalipour et al. 2013) where x_i is the observed value, y_i is the predicted value and n is the no. of observations.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(y_i - x_i\Big)^2} \tag{5.1.1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{5.1.2}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{x_i - y_i}{x_i}\right|\% \tag{5.1.3}$$

5.2 Experiment 1

This experiment uses the CO_2 emission data from Coal Electric Power Sector for training (90%) and testing(10%). The table below shows the evaluation metrics for all the models when forecasting Coal Electric Power Sector CO_2 emissions.

Metrics	ARIMA(12,1,3)	SVM	SVM-PSO	Prophet
1 RMSE	38.93	19.24	19.364	4.11
2 MAE	35.55	16.34	16.29	14.29
3 MAPE	38.68	17.205	16.96	16.13

Table 1: Evaluation metric table for Coal Electric Power Sector

It could be observed from the table that ARIMA has the highest value while Prophet has the lowest value for all three evaluation metrics. The lower value is preferred for all the metrics as they calculate the errors in forecasting therefore model having the lowest RMSE, MAE and MAPE value is considered to have more accuracy in prediction. From

Table 1 it can be inferred that Prophet is the best model for forecasting while ARIMA is the worst model for forecasting this time series.

5.3 Experiment 2

This experiment uses the the CO_2 emission data from Natural Gas Electric Power Sector for training (90%) and testing (10%). The evaluation metric values for all the models is given in the table below.

Metrics	ARIMA(8,1,12)	SVM	SVM-PSO	Prophet
1 RMSE	7.04	9.45	9.61	2.36
2 MAE	6.13	6.93	6.93	4.24
3 MAPE	13.07	14.25	14.06	8.62

Table 2: Evaluation metric table for Natural Gas Electric Power Sector

Table 2 shows that the Prophet has the lowest value for all the metrics while SVM and SVM-PSO having nearly the same values are outperformed by the ARIMA model.

5.4 Experiment 3

This experiment uses the emission values from Total Energy Electric Power Sector. The evaluation metrics for all models are given in the table below.

Metrics	ARIMA(11,1,9)	SVM	SVM-PSO	Prophet
1 RMSE	26.58	24.88	25.44	3.82
2 MAE	23.36	20.65	20.90	11.94
3 MAPE	17.04	13.84	13.81	8.92

Table 3: Evaluation metric table for Total Energy Electric Power Sector

Table 3 shows that the Prophet has the lowest value for all the metrics while SVM and SVM-PSO have nearly the same values and have better performance than the ARIMA model. ARIMA is the worst forecasting model for this time series while Prophet is the best.

5.5 Discussion

The prophet model had the best accuracy in forecasting the time series for all the three experiments outperforming the ARIMA, SVM and SVM-PSO model. Yenidogan et al. (2018) research also had similar results where the performance of ARIMA and Prophet were compared for Bitcoin forecasting and the results showed that the Prophet model outperformed ARIMA by a huge margin by having an accuracy of 94.5%. SVM-PSO model had outperformed ARMA when forecasting monthly streamflow of two rivers in the research done by Sudheer et al. (2014). In another research SVM had outperformed ARIMA for short term load forecasting (Amin & Hoque 2019). Therefore SVM models were assumed to be better than ARIMA which was only true for Experiment 1 and

Experiment 3. ARIMA showed the worst performance for Experiment 1 and Experiment 3 where the time series data had a somewhat parabolic trajectory (Figure 9). ARIMA was not able to accurately forecast such data. While in the case of Experiment 2 the data had a linearly increasing trend which is where ARIMA performed better than both the SVM and SVM-PSO model.

5.6 Results

The results of experiment 1,2 and 3 indicate that the performance of Prophet model in forecasting is the best compared to all the other models as it has lower RMSE, MAE and MAPE values, therefore, it is used for forecasting CO_2 emissions from the three sectors for 36 months into the future.

The results show that the CO_2 emissions from Coal electric Power Sector (Figure 10) and Total Energy Electric Power Sector (figure 12) are going to gradually decrease over the years compared to the 2018 and 2019 emissions while the CO_2 emissions from Natural Gas Electric Power Sector (Figure 11) are steadily going to increase with each passing year.

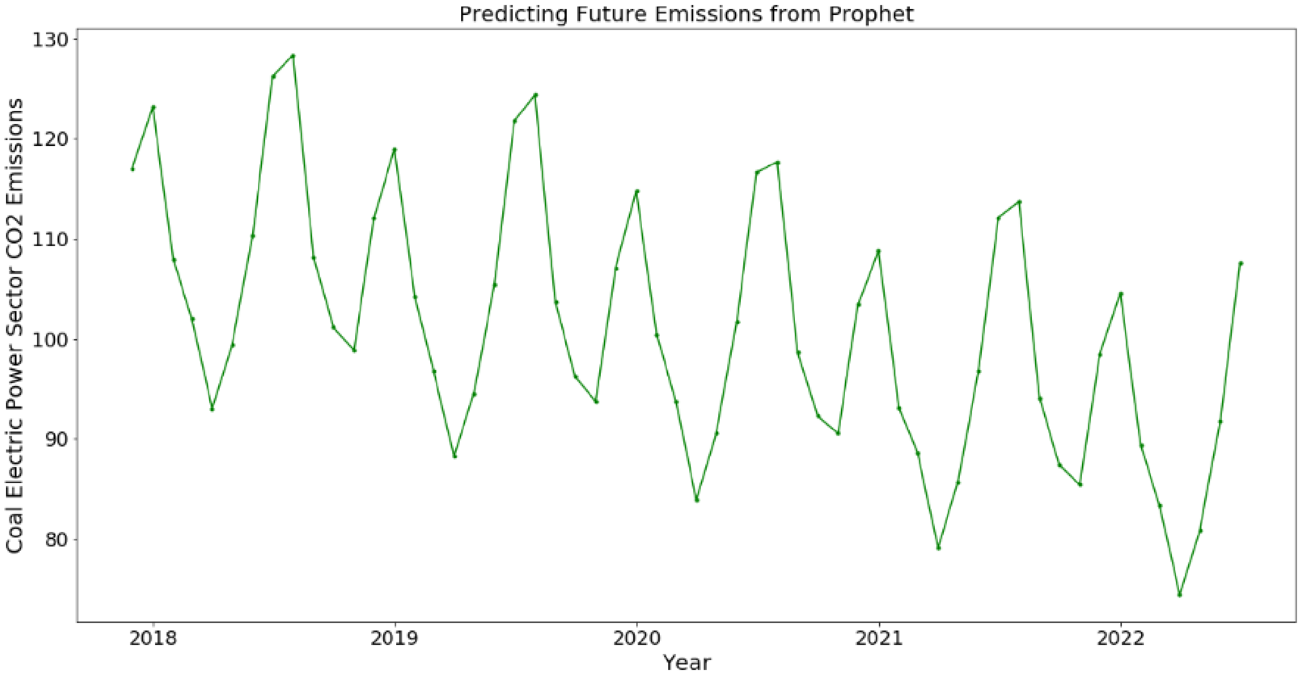


Figure 10: Future emission forecasting from Coal Electric Power Sector

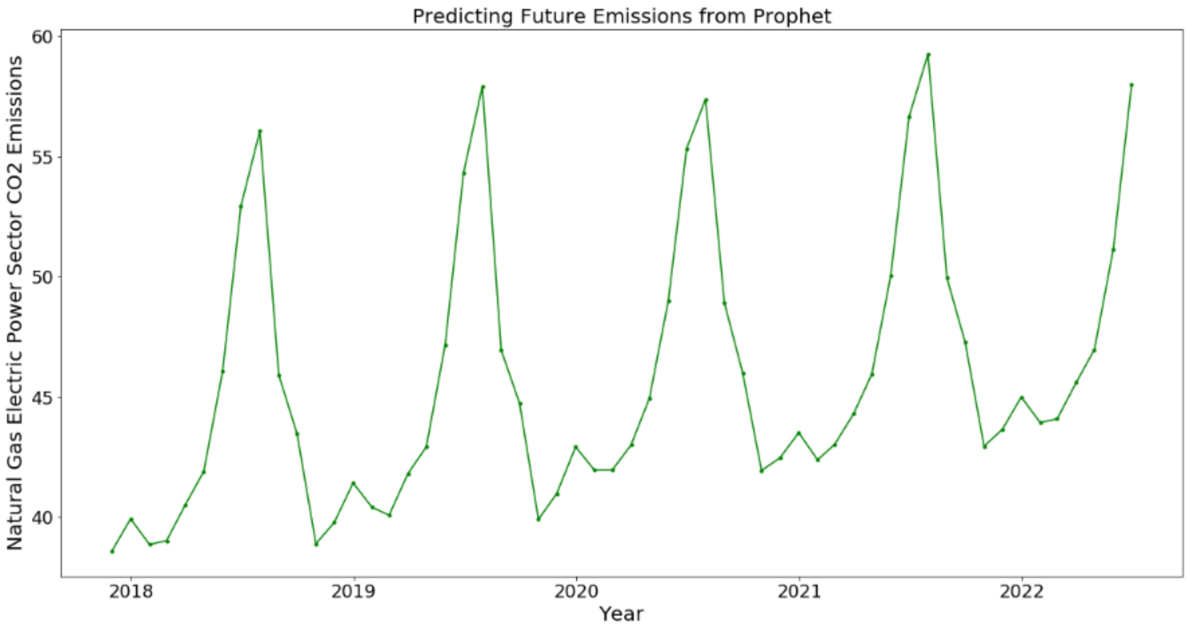


Figure 11: Future emission forecasting from Natural Gas Electric Power Sector

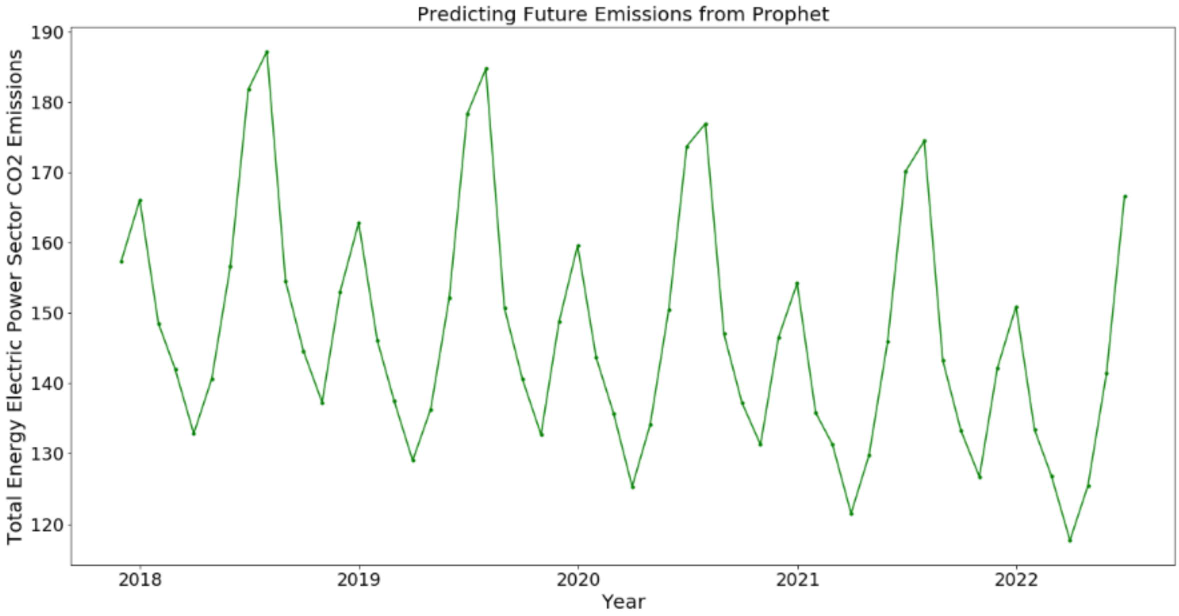


Figure 12: Future emission forecasting from Total Energy Electric Power Sector

5.7 Conclusion

This section compared the forecasting ability of four models ARIMA, SVM, SVM-PSO and Prophet using the evaluation metrics RMSE, MAE and MAPE and it can be observed that Prophet is the best model for forecasting emissions for all the three sectors. Therefore Prophet was used for forecasting future CO_2 emissions for 36 months. The research objective of finding the best model and forecatsing future CO_2 emissions using it is achieved during evaluation.

6 Conclusion and Future work

This research focuses on finding the best model for forecasting CO_2 emission from different sectors so that the published information could be used for monitoring the sectors and implementing new policies in the future. The data taken from the EIA website is preprocessed using RStudio. The exploratory analysis of the data using the Power BI tool reveals that the Coal Electric Power Sector, the Natural Gas Electric Power Sector, and the Total Energy Electric Power Sector has the highest $C0_2$ emissions compared to the other sectors. Therefore CO_2 emission forecasting for these sectors was done using ARIMA, SVM, SVM-PSO, and Prophet models. The performance of all the models are compared with each other using RMSE, MAE, and MAPE and the performance of Prophet in prediction is observed to be the best. Therefore Prophet is used for forecasting future emissions for 36 months. The forecasting shows that the emissions from Natural Gas Electric Power Sector are going to be on the rise while the emissions from Coal Electric Power Sector and Total Energy Electric Power Sector are going to decrease over those 36 months.
For future work, other models like LSTM (Long Short Term Memory), CNN-LSTM (Convolution Neural Network Long Short Term Memory) and Grey Model (GM) can be used for the time series forecasting. Methane is another most important GHG gas after CO_2 that could be forecasted using the same models in the future.

Acknowledgement

I would like to sincerely thank my research supervisor, Dr. Muhammad Iqbal for his invaluable input and feedback during the development of my research project. I would also like to express my gratitude to my family for constantly encouraging and supporting me.

References

Akcan, S., Kuvvetli, Y. & Kocyigit, H. (2018), 'Time series analysis models for estimation of greenhouse gas emitted by different sectors in Turkey', *Human and Ecological Risk Assessment* **24**(2), 522–533.
URL: *https://doi.org/10.1080/10807039.2017.1392233*

Amin, M. A. A. & Hoque, M. A. (2019), 'Comparison of ARIMA and SVM for short-term load forecasting', *IEMECON 2019 - 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference* pp. 205–210.

Ayvaz, B., Kusakci, A. O. & Temur, G. (2017), 'Energy-related CO 2 emission forecast for Turkey and Europe and Eurasia: A discrete grey model approach', *Grey Systems: Theory and Application* **7**(3), 436–452.

Borowik, G., Wawrzyniak, Z. M. & Cichosz, P. (2019), 'Time series analysis for crime forecasting', *26th International Conference on Systems Engineering, ICSEng 2018 - Proceedings* .

Bošnjak, Z., Grljević, O. & Bošnjak, S. (2009), 'CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data', *Proceedings - 2009*

5th International Symposium on Applied Computational Intelligence and Informatics, SACI 2009 (114), 509–514.

Brownlee, J. (2017), 'How to Create an ARIMA Model for Time Series Forecasting in Python'.
URL: *https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/*

Chuentawat, R. & Kan-Ngan, Y. (2019), 'The comparison of PM2.5 forecasting methods in the form of multivariate and univariate time series based on support vector machine and genetic algorithm', *ECTI-CON 2018 - 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* pp. 572–575.

Dabakoglu, C. (2019), 'Time Series Forecasting ARIMA, LSTM, Prophet with Python'.
URL: *https://medium.com/@cdabakoglu/time-series-forecasting-arima-lstm-prophet-with-python-e73a750a9887*

Du, J., Liu, Y., Yu, Y. & Yan, W. (2017), 'A prediction of precipitation data based on Support Vector Machine and Particle Swarm Optimization (PSO-SVM) algorithms', *Algorithms* **10**(2).

Etienne, B. (2019), 'Time Series in Python Exponential Smoothing and ARIMA processes'.
URL: *https://towardsdatascience.com/time-series-in-python-exponential-smoothing-and-arima-processes-2c67f2a52788*

Gu, J., Zhu, M. & Jiang, L. (2011), 'Housing price forecasting based on genetic algorithm and support vector machine', *Expert Systems with Applications* **38**(4), 3383–3386.
URL: *http://dx.doi.org/10.1016/j.eswa.2010.08.123*

Hosseini, S. M., Saifoddin, A., Shirmohammadi, R. & Aslani, A. (2019), 'Forecasting of $CO2$ emissions in Iran based on time series and regression analysis ', *Energy Reports* **5**, 619–631.
URL: *https://doi.org/10.1016/j.egyr.2019.05.004*

Kai, W., Xue, Y. J., Ji, K., Chen, H. M. & Chen, Q. (2011), 'Prediction of carbon flux based on wavelet networks', *2011 International Conference on Electric Information and Control Engineering, ICEICE 2011 - Proceedings* pp. 1553–1556.

Kingsley Appiah, Jianguo Du, Rhoda Appah & Daniel Quacoe (2018), 'Prediction of Potential Carbon Dioxide Emissions of Selected Emerging Economies Using Artificial Neural Network', *Journal of Environmental Science and Engineering A* **7**(8).

Kunda, D. & Phiri, H. (2017), 'An Approach for Predicting CO2 Emissions using Data Mining Techniques', *International Journal of Computer Applications* **172**(3), 7–10.

Li, D. Y., Xu, W., Zhao, H. & Chen, R. Q. (2009), A SVR based forecasting approach for real estate price prediction, *in* 'Proceedings of the 2009 International Conference on Machine Learning and Cybernetics', Vol. 2, IEEE, pp. 970–974.

Li, S., Zhou, R. & Ma, X. (2010), 'The forecast of CO2 emissions in China based on RBF neural networks', *2010 2nd International Conference on Industrial and Information Systems, IIS 2010* **1**, 319–322.

Lotfalipour, M. R., Falahi, M. A. & Bastam, M. (2013), 'Prediction of CO2 emissions in Iran using grey and ARIMA models', *International Journal of Energy Economics and Policy* **3**(3), 229–237.

Lu, X., Ota, K., Dong, M., Yu, C. & Jin, H. (2017), 'Predicting Transportation Carbon Emission with Urban Big Data', *IEEE Transactions on Sustainable Computing* **2**(4), 333–344.

Mason, K., Duggan, J. & Howley, E. (2018), 'Forecasting energy demand, wind generation and carbon dioxide emissions in Ireland using evolutionary neural networks', *Energy* **155**, 705–720.
URL: *https://doi.org/10.1016/j.energy.2018.04.192*

Nguyen, D. (2019), 'Learning Data Science Predict Stock Price with Support Vector Regression (SVR)'.
URL: *https://itnext.io/learning-data-science-predict-stock-price-with-support-vector-regression-svr-2c4fdc36662*

Palomares-Salas, J. C., De La Rosa, J. J., Ramiro, J. G., Melgar, J., Agüera, A. & Moreno, A. (2009), 'ARIMA vs. neural networks for wind speed forecasting', *2009 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, CIMSA 2009* pp. 129–133.

Philibert, A., Loyce, C. & Makowski, D. (2013), 'Prediction of N2O emission from local information with Random Forest', *Environmental Pollution* **177**, 156–163.
URL: *http://dx.doi.org/10.1016/j.envpol.2013.02.019*

Saleh, C., Dzakiyullah, N. R. & Nugroho, J. B. (2016), 'Carbon dioxide emission prediction using support vector machine', *IOP Conference Series: Materials Science and Engineering* **114**(1), 012148.

Sen, P., Roy, M. & Pal, P. (2016), 'Application of ARIMA for forecasting energy consumption and GHG emission: A case study of an Indian pig iron manufacturing organization', *Energy* **116**, 1031–1038.
URL: *http://dx.doi.org/10.1016/j.energy.2016.10.068*

Sheta, A. F., Ghatasheh, N. & Faris, H. (2015), 'Forecasting global carbon dioxide emission using auto-regressive with eXogenous input and evolutionary product unit neural network models', *2015 6th International Conference on Information and Communication Systems, ICICS 2015* pp. 182–187.

Singh, R. (2016), 'PSO-Based-SVR to forecast potential delay time of bus arrival. Applied on City of Edmonton real data.'.
URL: *https://github.com/RamanSinghca/PSO-Based-SVR*

Sudheer, C., Maheswaran, R., Panigrahi, B. K. & Mathur, S. (2014), 'A hybrid SVM-PSO model for forecasting monthly streamflow', *Neural Computing and Applications* **24**(6), 1381–1389.

Sun, W. & Sun, J. (2017), ‘Prediction of carbon dioxide emissions based on principal component analysis with regularized extreme learning machine: The case of China’, *Environmental Engineering Research* **22**(3), 302–311.

Tang, Y. & Zhou, J. (2015), ‘The performance of PSO-SVM in inflation forecasting’, *2015 12th International Conference on Service Systems and Service Management, ICSSSM 2015* (2011), 1–4.

Vincent, T. (2017), ‘A Guide to Time Series Forecasting with Prophet in Python 3’.
URL: *https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-prophet-in-python-3*

Wang, X. L., Liu, J. & Lu, J. J. (2009), ‘Wavelet transform and PSO support vector machine based approach for time series forecasting’, *2009 International Conference on Artificial Intelligence and Computational Intelligence, AICI 2009* **1**, 46–50.

Wang, Z. & Dang, Y. (2013), ‘Research on carbon emission predictiona in Jiangsu Province based on an improved GM (1,1) model’, *Proceedings of IEEE International Conference on Grey Systems and Intelligent Services, GSIS* pp. 93–97.

Wirth, R. & Hipp, J. (1995), ‘CRISP-DM : Towards a Standard Process Model for Data Mining’, *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (24959), 29–39.

Xiaoguo, W. & Yuejing, L. (2009), ‘ARIMA time series application to employment forecasting’, *Proceedings of 2009 4th International Conference on Computer Science and Education, ICCSE 2009* pp. 1124–1127.

Yenidogan, I., Cayir, A., Kozan, O., Dag, T. & Arslan, C. (2018), ‘Bitcoin Forecasting Using ARIMA and PROPHET’, *UBMK 2018 - 3rd International Conference on Computer Science and Engineering* (September 2018), 621–624.

Zhang, Q., Li, F., Long, F. & Ling, Q. (2018), ‘Vehicle Emission Forecasting Based on Wavelet Transform and Long Short-Term Memory Network’, *IEEE Access* **6**, 56984–56994.

Zhao, B., Zhang, Z., Jin, J. & Pan, W.-P. (2010), ‘Modeling mercury speciation in combustion flue gases using support vector machine: Prediction and evaluation’, *Journal of Hazardous Materials* **174**(1-3), 244–250.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0304389409015039*