National College of Ireland

Predicting CO2 Emission from Power Industry using Machine Learning

MSc Research Project
Master's in data Analytics

Pooja Singh
Student ID: X19234236

School of Computing
National College of Ireland

Supervisor: Mr. Aaloka Anant

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Pooja Singh
Student ID:	X19234236
Programme:	Master's in Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Mr. Aaloka Anant
Submission Due Date:	31/01/2022
Project Title:	Predicting CO2 Emission from Power Industry using Machine Learning
Word Count:	7496
Page Count:	4

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:	
Date:	31st January 2022

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):	

Predicting CO_2 Emission from Power Industry using Machine Learning

Pooja Singh
X19234236

Abstract:

Carbon dioxide is the primary contributor to global warming, which has had a devastating impact on both economic growth and human well-being. People and the government are working together to address climate change as a global issue so that our descendants don't have to bear the consequences. Even though Greenhouse Gas (GHG) emissions have decreased, the United States (U.S.) continues to be one of the top GHG emitters. Approximately 80% of Greenhouse Gas (GHG) emissions come from carbon dioxide (CO2). As a result, the decrease in CO2 emissions will contribute to reducing GHG emissions produced in the United States each year. from the US Energy Information Administration (EIA). The pre-processed data shows that the coal, natural gas, and total energy electric power sectors are the three highest emitters of CO2. Seasonal Autoregressive Integrated Moving Average (SARIMA), Long Short-Term Memory (LSTM), Prophet, and exponential smoothing models employ the emissions from these three sectors to anticipate CO2 emissions. It was shown that the Triple Exponential Smoothing model surpasses the all the other models with a MAE of 1.97 when it comes to projecting CO2 emissions from different industries, according to the data. Therefore, the data will be presented to structural engineers, builders, and the government to assist them design successful plans and policies for decarbonization.

Keywords: Seasonal Autoregressive Integrated Moving Average (SARIMA), Long Short-Term Memory (LSTM)

Area: Data Mining & Machine Learning

1 Introduction

CO2 Emissions have become a cultural consensus in the reduction of greenhouse gas emissions. Its worldwide environmental effect is matched by its economic impact. Global energy-related CO2 emissions climbed 1.7% in 2018 to 33.1 Gt CO2. Environ two-thirds of the rise in fossil-fuel emissions came from power plants. The power industry was chosen for three reasons. When it pertains to carbon and resources use, it's one of its most carbon- and resource-intensive industries. And while CO2 emissions in the US power industry have risen over time, little research has been done to anticipate them, presenting a fresh research opportunity. Finally, few studies have shown that large-scale industrial and urban development projects involve environmental costs that should be addressed. This study seeks to anticipate carbon emissions from the electricity industry and identify the most carbon heavy sectors in power industry. Since about the 1960s, energy efficiency and carbon emissions have been the subject of several studies. A wide range of energy models and technologies have been utilized to analyse the relevant emissions in depth.

1.1 Motivation:

In 2020, the US electric power industry emitted 1447 million metric tons of CO2, or roughly 32% of total US energy-related CO2 emissions of 4,575 million metric tons (MMmt). US oil and gas output has risen dramatically in the last decade. The US presently leads the globe in both oil and gas production. The world must abandon fossil fuels, at least those that do not collect and store carbon. We can combat climate change as the world's greatest oil and gas producer by incorporating energy firms, regulating oil and gas production to limit GHG emissions, and implementing policies that cut GHG emissions domestically and globally. The historic Paris Agreement from 2015, which has now been signed by every country in the world except the United States under Trump's administration, seeks to keep global warming to no more than 2 degrees Celsius over pre-industrial levels. As a result, the United States must keep a close eye on its carbon emissions.

1.2 Research Question:

"How effectively can machine learning algorithms be used to forecast carbon emissions within United States (US) from various sectors of the power industry?"

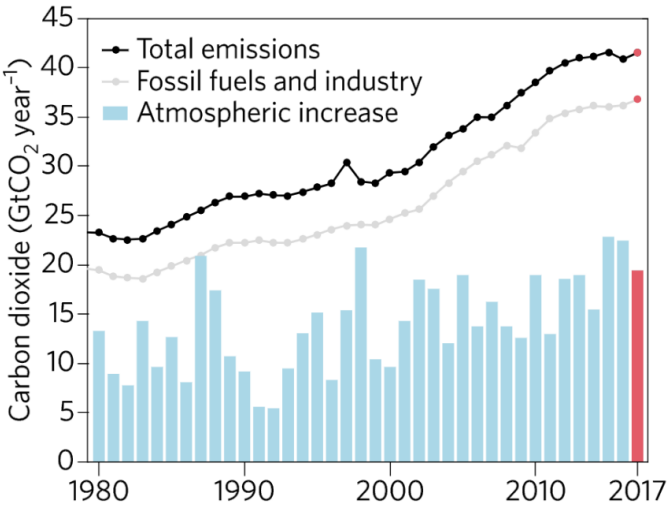

Figure 1. Carbon emission throughout world

The ever-increasing GHG emissions require immediate attention and action. This will assist government officials and city developers in making better decisions that will serve as future health insurance.

1.3 Research Objectives:

• To investigate the EIA dataset's significance and then perform pre-processing.

• Highlight the regions and industries with the highest CO2 emissions.

• Evaluate the performance of different machine learning models and choose the best one.

• To forecast potential CO2 emissions, use the best model available.

• Presenting the findings so that appropriate actions and policies may be implemented.

The work done on estimating carbon emissions in the power industry, as well as the machine learning approaches utilized, will be examined in Section 2. The many steps of the approach employed in this study are explained in Section 3. The design specification is discussed in Section 4. Section 5 delves into each stage of the process, from data collection through model development and assessment measures. It will go over both the design choice and the data processing procedures in detail. The implementation of the model is explained in section 6. The numerous assessment metrics and criteria that will be used to analyse the study findings and discuss overall conclusions will be discussed in sections 7 and 8. The conclusion is discussed in the Section 9.

2 Related Work

2.1 Introduction-

The development of systems for predicting Carbon emission has accelerated in recent times as a result of growing interest in estimating CO2 emissions. The power industry is a significant source of CO2 emission. Rapid industrialization may benefit a country's economy while also raising CO2 emissions. This research attempts to identify the world's most polluting nations and to accurately investigate and forecast carbon emissions from the most polluting sectors.
Previous work has been analysed to identify gaps and opportunities for improved performance and research. Carbon emission trends in various areas and approaches used to detect the trends are explored as part of this topical overview.

2.2 Application of Neural Networks(NN) to the energy consumption and carbon emissions from the power sector:

In the recent decade, ANNs have made their way into the power industry because of their greater practicality and dependability than classic regression methodologies. Many neural network topologies have been employed to estimate energy and emissions. With coefficients of variation between 2 and 40 percent in recurrent neural networks, auto-associative neural networks, general regression neural networks and back propagation, are the most effective ones. For example, the differences in prediction accuracy may be traced back to many different variables such as how frequently training data is collected and what kind of data-gathering tools were used to collect and analyse it.

Researcher (Jabreel et al.; 2020) utilized the ridge polynomial neural network with error feedback (RPNN-EF) to predict the next five values of carbon dioxide emissions in three countries. There were seven different forecasting systems compared to the RPNN-EF accuracy. The suggested technique, when compared to the other 7 machine learning forecasting techniques, delivers fair projections on the average. The results suggest that neural networks with error feedbacks may be used for recursive multi step forecasting. Experiments were conducted using a dataset representing the electric load of an educational facility.

Serbian CO2 emissions were calculated using artificial neural networks (ANNs) designed by (Radojevic et al.; 2013). Throughout their inquiry and assessment, they tried to discover if the technique could be used to forecast key parameters of sustainable development in Serbia. The research was done to address the lack of data and to predict alternative development scenarios and their environmental impacts on the environment. NeuroShell 2 was used to build and train the neural network. The years 1999 to 2007 were considered when making the calculations. On the basis of the data, it is reasonable to conclude that artificial neural networks (ANNs)

may be used to mimic greenhouse gas emissions, one of the environmental issues affecting sustainable development. Furthermore, ANN models may be a significant tool for modelling alternative development scenarios, the impact of government and industry policies and regulations, and, as a consequence, for aiding in decision-making on sustainable development at the national and international levels.

(Aggarwal, Kumar, and Sharma et al.; 2013) employed back propagation of ANN to estimate carbon emissions. Data for prediction, estimate, or assessment were calculated using statistical methods that utilised back propagation of ANNs to train data for prediction. To anticipate a large amount of data, an artificial neural network is the ideal way to use, based on discoveries from the least squares and back propagation algorithms. Using back propagation, for example, the average relative error percentage are always fewer than that are using the least square technique. This illustrates how the back propagation method outperforms the least squares method. To evaluate, estimate, and forecast a huge collection of data, back propagation, an artificial neural network (ANN), is superior to analytical techniques such as the least squares method.

2.3 Other Machine Learning techniques used for estimating carbon emissions:

(Kuruppuarachchi et al.; 2021) used machine learning to enhance the prediction of business carbon emissions for investors' risk assessments. When it came to the optimal emission prediction method, it used a Meta-Elastic Net learner to combine forecasts from several base learners. On average, it improved accuracy by up to 30 percent when compared with earlier models. Extra predictors (energy production pr consumption statistics) and additional corporate disclosures in some industries like power sector, according to the research, might improve prediction accuracy even further.

To estimate carbon emissions, ML regression models were employed by (Akbarzadeh et al.; 2020) conventional regression, deep learning and shallow learning, all of which were applied in the research. Increasing the depth of your understanding LSTM was shown to provide the best results. The LSTM model beat the other ML models with the highest R coefficient and the least root mean squared error (RMSE). (Juan Wang et al.; 2021) researched China's energy structure, which is heavily reliant on coal. Five variables were used to investigate the factors that impact power emissions of co2 (how much coal is used as a major energy source, how much CO2 is emitted per unit of energy used, and how many people live in cities). Pearson coefficient and correlation analysis are used to compare different carbon emission scenarios. Chinese energy-related CO2 emissions have a major impact on world emissions, although the country's per capita emissions are extremely low, and the intensity of its emissions is steadily decreasing.

China's Jiangsu province was forecasted using a novel technique for the GM(1,1) model by (Wang and Dang et al.; 2013). Comparing the new model to the original GM(1,1) model reveals superior prediction accuracy and less relative errors than the latter. A growth in carbon emissions of 53316.14 ten thousand tons is predicted for 2020 based on the provided scenario. PC-RELM (Principal Component-Regularized Extreme Learning Machine) was used by Sun & Sun (2017) to predict the amount of CO2 created as a result of energy consumption in China, using data from the China Statistical Yearbook. When measured by the metrics median absolute percentage error (MdAPE) and maximum absolute percentage error (MAPE), the model outperforms the competition when it comes to estimating emissions (MaxAPE). Nevertheless, statistical methods continue to be employed by academics for time series analysis, despite the arrival of machine learning models in forecasting (Hosseini et al. 2019, Akcan et al. 2018).

2.4 Statistical analytic approaches, Nonlinear intelligent models and Grey forecasting:

Forecasting CO2 emissions may be broken down into three main areas as shown in the image below. These really are: statistical analytic approaches; nonlinear intelligent models; and grey forecasting. In recent years, some combination models have been used for predicting carbon dioxide emission based on the 3 models. (Dai et al.; 2018) proposed a carbon dioxide emission prediction model employing GM(1,1) as well as the least squares support vector machine (LSSVM) which is improved using the modified shuffle frog jumping algorithm. The researcher (Li et al.; 2018) developed an improved extreme version of learning machine prediction models relying on the grey prediction theory with the help of support vector machine to estimate the carbon emissions linked to electricity usage in the area of Beijing-Tianjin-Hebei. Long-term memory networks, grey relational analysis, and principal component analysis were used to anticipate China's carbon emissions in (Huang et al.; 2019). To better anticipate renewable energy consumption over the short term, (Moonchai and Chutsagulprom.; 2020) used a modified multivariable grey prediction model along with the Kalman filtering. According to these findings, the combination of the models are more accurate in predicting carbon emissions than single models (Pino-Mejas et al.; 2017) , (Zhao et al.; 2018).

Many industries, including energy forecasting, have used FGMs because of their ability to anticipate time series with tiny sample sizes. (Wu et al.; 2018) employed the innovative FAGMO(1,1,k) to forecast China's use of nuclear energy. Greypower-based water usage prediction was implemented by Yuan et al. in 2019. China's Chongqing used new fractional time-delayed grey model for predicting the natural gas and coal usage. Fractional derivative accumulation GM(1,1)model FAGM(1,1,D) was developed by (Gao et al.; 2015) and was applied to China's carbon dioxide emissions. The present FGM model employs fractional order accumulation (FOA), which is proven to be effective in minimizing the mistakes of grey models. It's simple to apply statistical models, but you must collect a lot of historical data first (Rao et al.; 2020). However, the non-linear intelligent model have a poor resilience and require a large amount of information to develop a sample database, but they are very flexible and powerful when it comes to making predictions, estimations, and dealing with noisy data. An alternate fore-casting strategy for CO2 emissions might be the grey forecasting method, which uses fractional order accumulation to improve the alignment stability and high flexibility of grey models when input data is ambiguous or scarce. As a result, making use of these models to their full potential is quite difficult. In the meanwhile, the extension models are derived from the time series of carbon emission, without taking into account the processes of carbon modelling.

Summary of the literature on forecasting models.

Author	Model type	Model	Application
Meng and Niu (2011)	statistical analysis model	logistic equation	China's carbon emissions
Köne and Büke (2010)	statistical analysis model	trend analysis	CO2 emissions from fuel combustion
Aydin (2015)	statistical analysis model	multiple linear regression	CO2 Emissions in Turkey
Sutthichaimethee (2018)	statistical analysis model	ARIMAX model	Industrial CO2 emissions in Thailand
Sun and Liu (2016)	nonlinear intelligent model	support vector machine	Industries and residential CO2 emissions
Wei et al. (2018b)	nonlinear intelligent model	extreme learning machine	CO2 emissions in Hebei
Azadeh (2009)	nonlinear intelligent model	fuzzy regression algorithm	oil consumption estimation
Xu et al. (2019a)	nonlinear intelligent model	artificial neural network	China's CO2 emissions
Qiao et al. (2020	nonlinear intelligent model	lion swarm optimizer	carbon dioxide emissions
Wang and Li (2019)	grey forecasting method	non-equigap grey Verhulst	CO2 emissions and economic growth
Pao et al. (2012)	grey forecasting method	nonlinear grey Bernoulli	China's CO2 emissions
Wu et al. (2015)	grey forecasting method	multivariable grey model	CO2 emissions in BRICS
Ding et al. (2020)	grey forecasting method	discrete grey model	Chinese energy-related CO2 emissions
Xiang et al. (2020)	grey forecasting method	accumulated grey model	carbon dioxide emission
Xu et al. (2019b)	grey forecasting method	grey rolling model	Chinese greenhouse gas emissions

Figure 2. Summary on some of the forecasting models from literature

2.5 Summary of Literature Review:

In the field of energy forecasting, the focus has shifted from old statistical approaches to today's more advanced machine learning algorithms like neural networks, deep learning, and support vector machines, among others, as these prediction methods continue to advance in sophistication. To curbing carbon emissions, this research fills up some of the knowledge gaps left by previous studies. Despite the increased interest in ARIMA-based machine learning hybrid approaches, most empirical investigations used only annual or monthly data with no consideration of seasonality or time of year. When compared to other time series models, research has shown that the conventional ARIMA model is the best at capturing cycles (Reikard, 2019). The ARIMA model's expertise is capturing trends and seasonality in demand for air travel, and this study's monthly data did just that. In addition, the ARIMA model's accuracy declined with increasing forecasting steps, making it suited for short- to medium-term forecasts (Liu et al., 2014). As a result, the chronology of data and the scope of the inquiry warrant the use of the ARIMA model in this investigation. In spite of its simplicity, the ARIMA model was selected as the principal tool for predicting carbon emissions.

Moreover, all models were successful in forecasting, but exponential smoothing utilizing trends was selected since it had the fewest forecast errors. Also, there hasn't been any research done on American carbon emissions from electricity generation, which is the main source of GHG emissions in the United States. Final suggestions for energy saving and reduction in emissions from power generation will be made.

3 Methodology

In this research, we are attempting to highlight which countries are emitting the highest levels of carbon dioxide, as well as how to anticipate future emissions from those countries using historical data from the EIA's dataset (Energy Information Administration). Described in this part are the scientific method and the architectural design employed in this investigation. In order to answer the study objectives, we will employ Data Mining (DM) methods. Anomalies, hidden patterns, and correlations in huge data sets can be discovered by data mining. An industry standard approach for data mining known as CRISP-DM has been used in this study because it encourages best practices and makes it easy for projects to be repeated. A total of six phases are included in the CRISP-DM technique. Data Preparation, Modelling, Evaluation and Deployment are just a few of the many aspects of this. This technique generates a standardized framework for the management and planning of projects. The crisp method used in this study is shown in the below figure 3 given by (Cornellius Yudha Wijaya.; 2021).

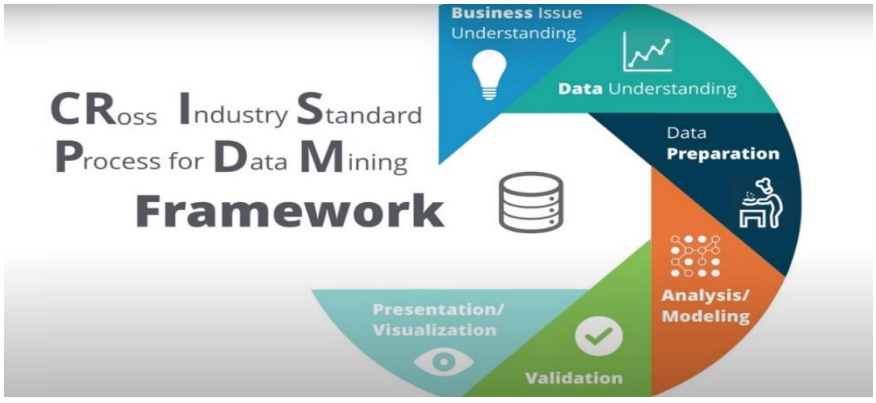

Figure 3. Proposed Methodology for carbon emission

3.1.1 Business Understanding

Research goals are defined in this phase and then a business strategy is developed to help attain them, which is the most critical part of the whole project cycle. The fight against climate change has already begun with individuals teaming forces to develop alternative solutions that might assist to improve the environmental condition. Adopting environmentally friendly lifestyles is a matter of urgency if we are to slow the rate of global warming. Due to its high GHG emissions, the United States must work with other high GHG emitting countries to avert the extinction of all living organisms as a result of rising global temperatures. For the research project's purpose, it is to anticipate US CO2 emissions from various sectors, in order to identify sectors with growing CO2 emission and use this knowledge to create policies to regulate emissions and fulfil the Paris Agreement's reduction target of 25% to 28% by 2025. Four machine learning models have been developed to anticipate CO2 emissions, and the best model will be selected for future projections since it has the highest accuracy.

3.1.2 Data Understanding

CRISP-second DM's phase sees the analysis of all of the acquired data for any patterns or trends that may have emerged. EDGAR[1] - Emissions Database for Global Atmospheric Research is used to identify the nation and industry with the greatest emissions. On the EIA[2] - Energy Information Administration website, the second dataset used in this research study is freely available for download and reuse and this dataset does not have any ethical implications. Data from January 1973 to August 2021 comprises monthly CO2 emissions from 9 sectors in million metric tons of CO2 unit (MtCO2). There are 6 columns and 5256 rows in the data, which includes NA values.

3.1.3 Data Preparation & Data Cleaning

The cleaning of the dataset is the third process, and Python was utilized to accomplish this task. Research projects can only go smoothly when defects like missing values and inconsistencies are taken care of before they can be used in the analysis of the real-world dataset. Data knowledge is essential before beginning any pre-processing of the data set (exploratory data analysis). The researcher can have a better understanding of the dataset by plotting a correlation graph between independent and dependent variables, and independent variables that have a strong correlation had to be deleted. To get the results you want from exploratory data analysis, you'll need to clean up your data first. Data cleaning comprises the elimination of missing values, irrelevant data, and non-applicable (NA) data. Index numbers for sectors were deleted from a fourth column in the dataset and the column names of three other columns was changed as a result. When the year and months were turned into a date column, the thirteenth month was also eliminated using dropna() method to remove all year-to-year emissions sums. Eventually, the cleaned dataset was saved as a CSV file and made ready for any further analysis. We will reshape the dataset from EDGAR into a vertical data frame since the present form of the data frame does not meet our objectives. This will find things simpler for us to analyse and visualize the data.

3.1.4 Modelling

To achieve the research objective, the use of a suitable data mining algorithm is important. Algorithm selection is a critical decision based on the type of data being analysed. Time series models were chosen since the data set was comprised of a variety of data points that were collected over a period of time. Pre-processed data is modified using several models, and the resulting data is then fed into different models for distinct time series patterns. They

[1] https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

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

are given training on 90% of the data in order to predict the remaining 10% of data, which is used for testing purposes.

3.1.5 Evaluation

Model performance and efficiency computation is an important part of this research since it helps assess the model's effectiveness in fulfilling the defined issues or goals. Knowledge gathered from the use of four time series forecasting algorithms is examined to get further insight into the intended outcome in this stage. As a result of this knowledge, a visual representation of the electrical demand information is shown to end stakeholders. There are four evaluation measures employed in this research: the Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Mean Absolute Error (MAE) and the Root Mean Square Error (RSME).

3.1.6 Deployment

The level of complexity and ease of use of the deployment depends on the requirements. An established research document is presented, as well as a setup manual and a solutions document. The entire implementation can be found in Github[3].

4 Design Specification

Figure 3 depicts the project's two-part design architecture. There is a Data Layer at the top and an Application Layer and a Business Logic Layer at the bottom. Using the Emissions Database for Global Atmospheric Research website, country-by-country emission data is retrieved from the US Energy Information Administration website. Python programming language is used to sanitize this data. EDA is performed on the cleaned data in order to gain a deeper understanding of the data's distinct characteristics. Python's Scikit learn library was used for data encoding after EDA. The implementation is graphically shown in figure 4 below.

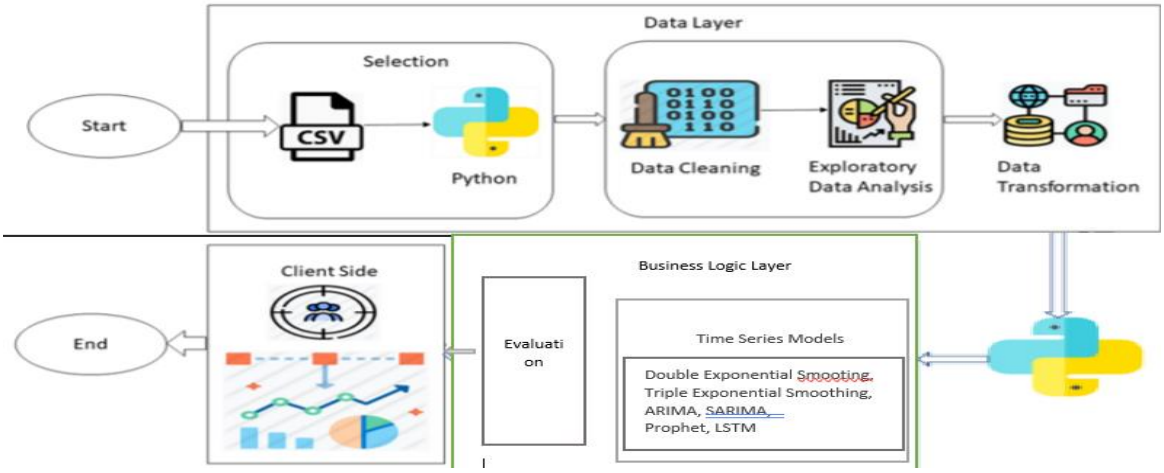

Figure 4. Implementation flow diagram

Mean Absolute Percentage Error, Mean Absolute Error, Mean Squared Error, and Root Mean Square Error were used to evaluate the effectiveness of four machine learning algorithms on encoded/transformed data at the Business logic layer. Visualizations of performance results were used at the application layer to make it apparent to end-users what they might expect.

[3] https://github.com/Isro4488/MSC-Data-Analytics---Thesis-Submission

5 Implementation

This phase includes a complete explanation of the various procedures involved in constructing ARIMA, Double Exponential Smoothing, Prophet, SARIMAX and LSTM algorithms to anticipate carbon emission from power sector. Python was used to program the entire project. Various statistical programs in Jupyter Notebook have been used to preprocess data and build ARIMAX, LSTM, Exponential Modelling, SARIMAX, and Prophet models. The LSTM model is also built in Jupyter notebook using the Tensorflow and the Keras Python tools. In addition to being easy to use, the Keras and Tensorflow libraries are ideal for neural network techniques.

5.1 Exploratory Data Analysis

5.1.1 Data Selection

The carbon emission dataset, accessible as a.csv file from the U.S. Energy Information Administration (EIA), was downloaded via that agency's website. Dataset about emission from all countries is extracted from Emissions Database for Global Atmospheric Research (EDGAR) and used to determine which countries and industries have the highest levels of emitted CO2. Monthly CO2 emissions from nine industries in million metric tons of carbon dioxide unit were retrieved from the EIA website between January 1973 and August 2021. (MtCO2). Within the dataset, there are six columns and 5689 rows, which includes the NULL values (which are not included).

Motivation for choosing the power sector and the Natural gas emission is highlighted below: Fossil fuels accounted for 93 percent of total U.S. energy production in 1966. Non-fossil fuel sources, such as renewables and nuclear power, have also seen an increase in output during the last few decades. Thus, fossil fuels have made up around 80 percent of US energy output in the previous decade, as a result. There has been a 15 quadrillion British thermal units (quads) rise in U.S. crude oil, and natural gas plant liquids (NGPL) and dry natural gas output growth since 2008. The figure 5 below shows the distribution of carbon emission from different sectors in United States, it shows the renewable sources of energy are also contributing hugely towards the emission.

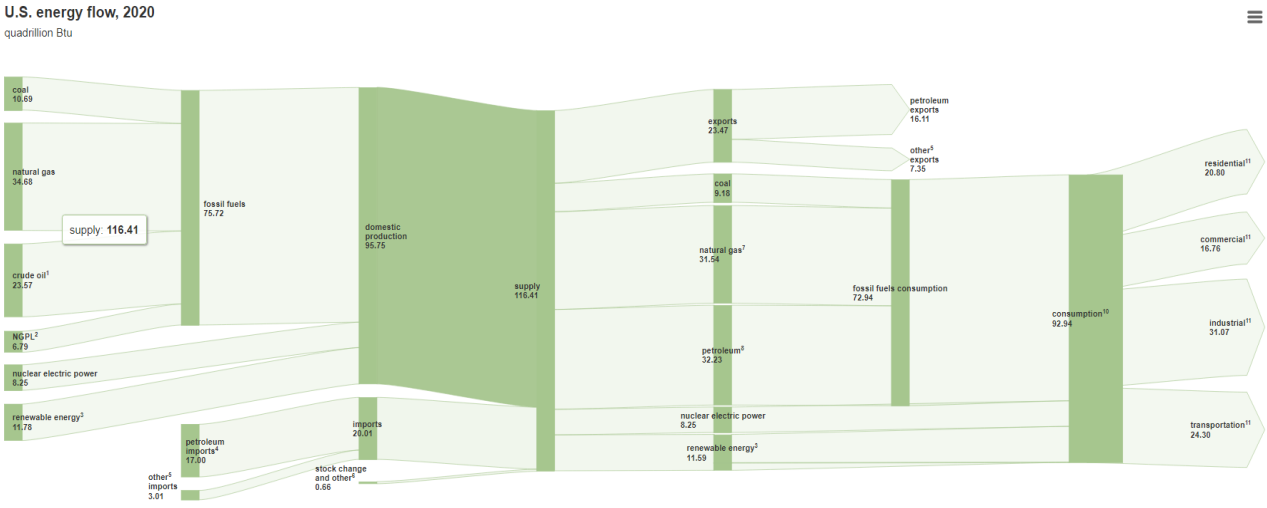

Figure 5. U.S Energy Information Administration, Carbon emission contribution

5.1.2 Data Preprocessing

Impurities like as special characters, missing values (NA), or incorrect values might make the real-world data unusable for training models; hence pre-processing is always required to

eliminate these impurities before training models. Open-source software Python was used to clean the dataset for research purposes. The following are the procedures that were performed in order to clean the dataset:

1. In the dataset, the column 'Column Order' (column number 4), which held index numbers for the sectors, was deleted.
2. As seen in Figure 8, the column names for YYYYMM, MSN, and Description were changed.
3. The month and year column was converted into a timeseries object using date_parser argument in read_csv funtion.
4. Removing the non-datetimeindex rows.
5. Convert the index to a datetime format, coerce errors, and filter out NaTs (not applicable).
6. Emission value is converted from object to numeric type.
7. CO2 emissions from the previous 12 months were added together to form a 13th month. The year was transformed to NA using the gsub function to remove it from the time series analysis because it was unnecessary.
8. The additional 418 NA values added to the dataset are then removed, and the cleaned data is saved as a CSV file for use by the models.
9. The data from EIA has been shifted to a vertical dataframe to meet our needs.
10. Check all the nations in the dataframe that are unique. Next, a few rows that are regions but not nations were omitted.

First, using data from the EDGAR database, the country with the maximum emissions is selected. Using the Python plot function, it is possible to acquire a better understanding of the EIA data and identify the nations with the greatest levels of emission.

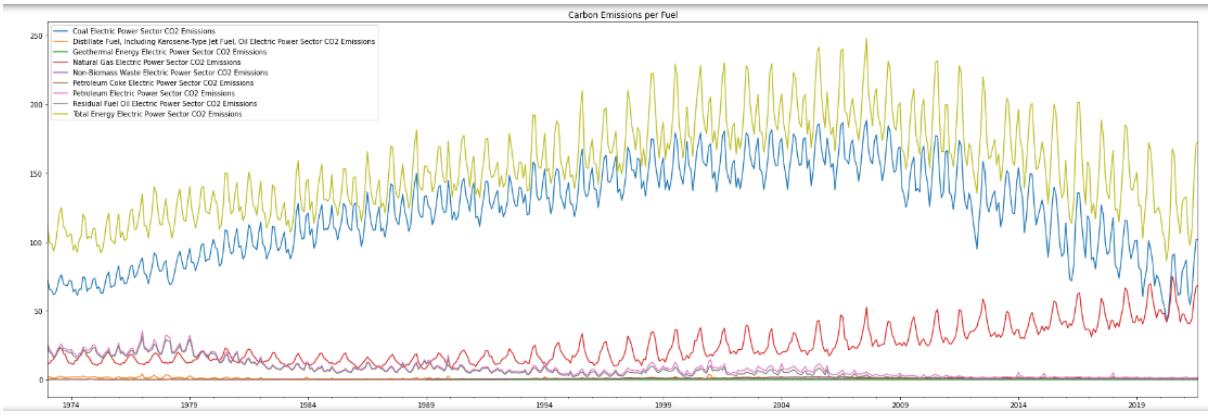

Figure 6. Carbon Emission per fuel from US

For the forecast of CO2 emissions, the data from the EIA is displayed, and three sectors with the largest emissions are chosen for analysis. As shown in Figure 6, the highest CO2 emitting sectors are coal, natural gas, and total energy. Despite the fact that emissions from the total energy electric power sector and coal electric power sector and have decreased over time, emissions from the natural gas electric power sector have increased.

5.2 Exploratory Data Analysis

After plotting the dataset taken from EDGAR, its clearly evident that United States has the highest level of CO2 emissions and choropleth library is used for visualizing it in figure 7 below.

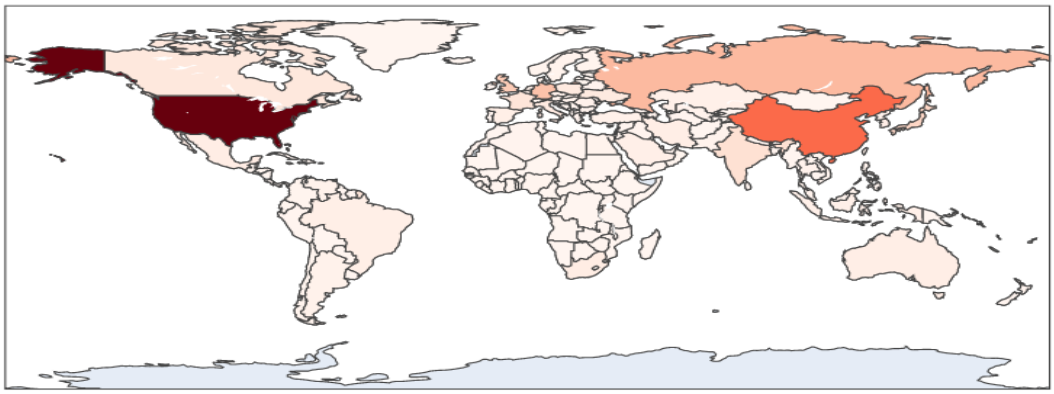

Figure 7. Carbon Emission Intensity in the world

All the greenhouses are seen and as seen from the graph below in figure 8, carbon dioxide contribution is much more than any other greenhouse gases.

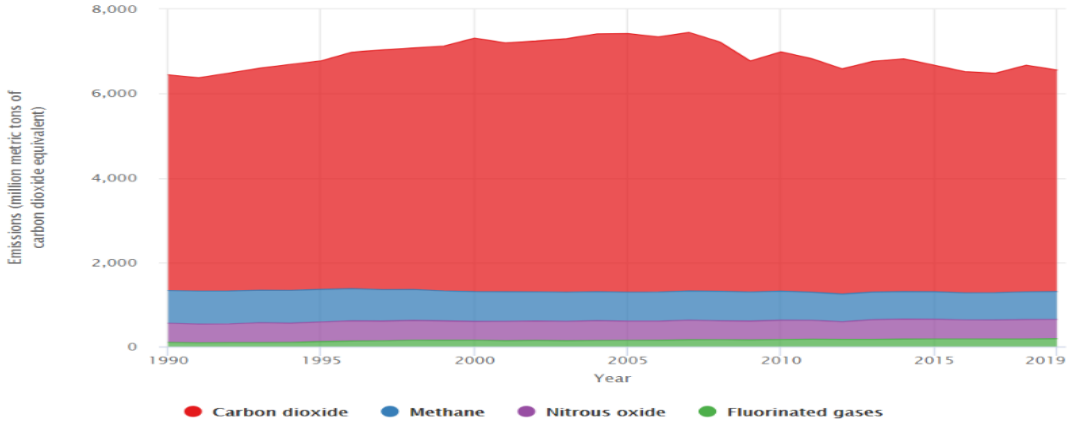

Figure 8. GHG Emissions from US

The usage of natural gas has been steadily growing in recent years. The usage of coal for electricity generation, on the other hand, has been falling. This is demonstrated by the plots of CO2 emissions from coal and natural gas, which indicate that, although the CO2 contribution from coal is decreasing, the CO2 contribution from natural gas is increasing. The below figure 9 shows how the natural gas usage in power industry is increasing rapidly.

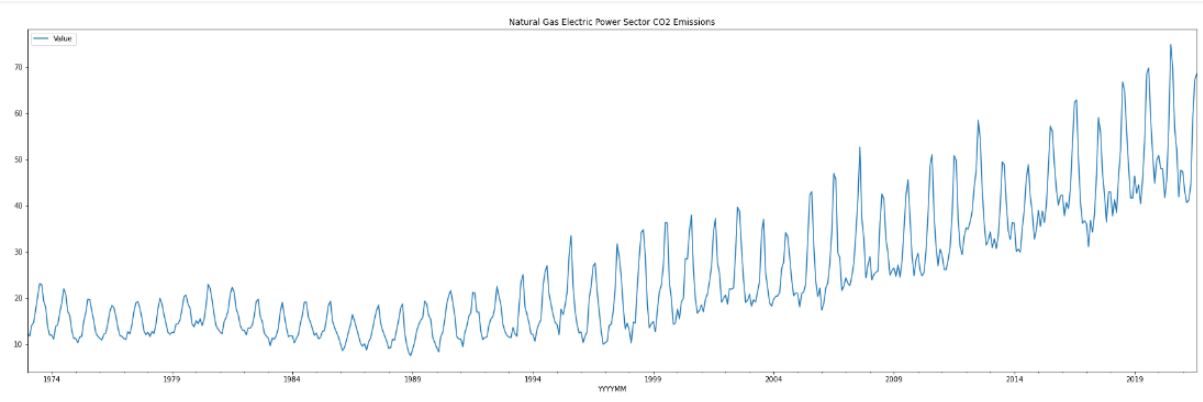

Figure 9. Natural gas emission from power sector in US

6. Evaluation

Anaconda's Jupyter Notebook software is used to write the Python code for model creation. Scikit-learn or sklearn, numpy, pandas, prophet, and matplotlib are the common libraries utilized by all of the models implemented in this research. Each model's training and testing data are separated.
A visualization of our time series data is the first step we need to take. The narrative will provide us a sense of the series' general pattern and seasonality. We will then apply a statistical approach to examine the dataset's trend and seasonality. The nonstationary dataset is transformed into a stationary dataset by eliminating the trend and seasonality from the dataset, and the residuals are then studied further.

6.1.1 Experiment 1 - ARIMA:

To convert the year column to a datetime format, the data is first loaded into Jupyter Notebook and processed using a date parse function. Time series here is tested using the Dickey-Fuller test, which involves rolling mean plot and then conducting its test. The null hypothesis can be ruled out if rolling mean value is straight and also the Dickey fuller's (DF) value < 0.05.

The TestStationaryAdfuller() function will take care of this test giving

```
p-value                         0.665268
```

According to the figure, there is inadequate evidence to reject the null hypothesis because the time series has a unit root. This demonstrates that the series follows a pattern. So, it's not a stationary, as previously said. In addition, the Test Statistic is larger than the critical values with confidence levels of 90, 95, and 99 percent. In other words, there is no evidence to support the null hypothesis. It is thus nonstationary. To make the time series stationarity we use moving average technique. Testing the stationarity again using Dickey Fuller Test gives results as:

```
p-value                         0.000015
```

It is necessary to draw the ACF and PACF (partial Autocorrelation Function and Partial Autocorrelation Function respectively) plots in order to obtain the values for p and q from the PACF plot and the ACF plot respectively. The plot is shown in figure 10 below.

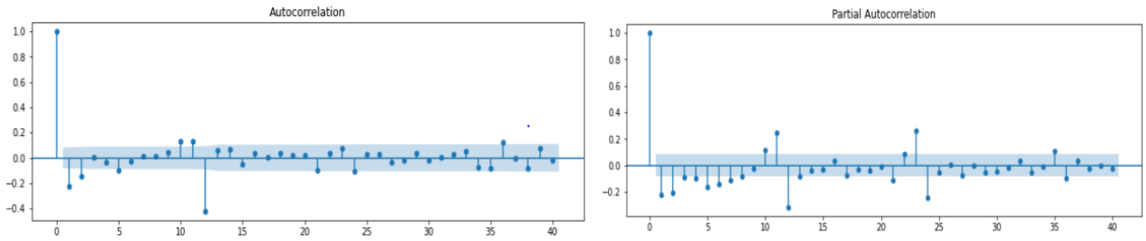

Figure 10. ACF and PACF plot

The.fitted values function in plots is used to verify the model's fit. Forecasting values is done using the.forecast() function which is then transformed to real numbers using the np.exp() function because these data are log-formatted. The RMSE, MAE, and MAPE values for the test and predicted values are shown against each other. With an MAE of 7.6, this model predicts emissions for natural gas emission, and the projections are a bit close to the actual test data.

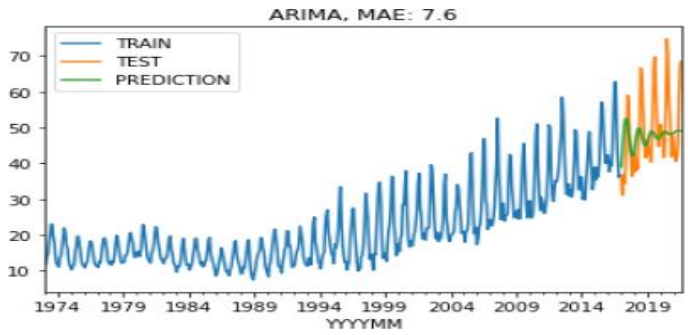

Figure 11. The actual and predicted values comparison from ARIMA

6.1.2. SARIMA:

After that, we'll add the integration order I. (d). The d parameter specifies the lot of variations necessary to stabilize the series. Lastly, we include seasonality S(P, D, Q, s), where s is just the duration of the season. The parameters P and Q, which are the same as p and q, except for the seasonal component, are required for this component. Seasonality may be eliminated from a series using D, which indicates the seasonal integration order. The SARIMA(p, d, q)(P, D, Q, s) model is the result of all of this. In order to model using SARIMA effectively, we must first eliminate seasonality and non-stationary behaviour from our time series using time series transformation. The below figure shows the plot of actual and predicted values and the MSE is as low as 2.95.

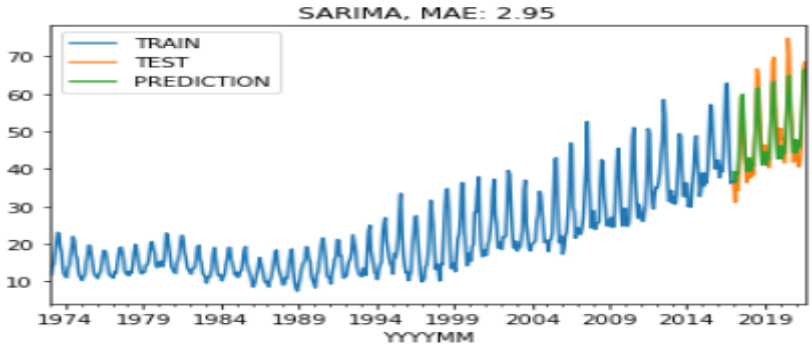

Figure 12. The actual and predicted values comparison from ARIMA

6.2 Experiment 2 - Exponential Smoothening:

Exponential smoothing is used with all of the variants, and hyperparameter tuning has been used to further increase the performance of the model even further. Weights of most recent to the old observations are assigned in an exponentially decaying order. Greater recent data is given more priority ("weight") since it is considered to be more relevant and so is given lower priority ("weight"). A simple exponential smoothing method is not used since the series contains trends. All the exponential model used here are implemented using “additive” trend.
Due to the fact that it indicates a pattern, the Double Exponential Smoothing approach is considered more credible for assessing this dataset. It's a more sophisticated approach and uses below equation:

$$y = \alpha x_t + (1 - \alpha)(y_{t-1} + b_{t-1})$$
$$b_t = \beta(y_t - y_{t-1}) + (1 - \beta)b_{t-1}$$

Below figure compares the actual and predicted trend.

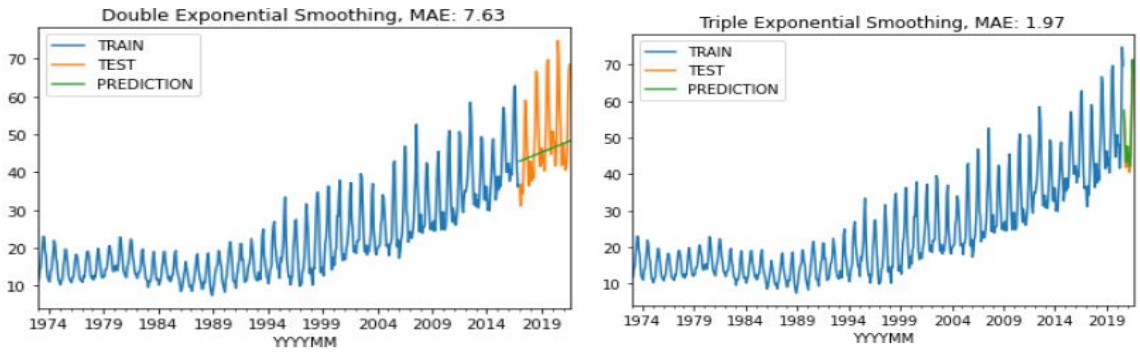

Figure 13. The actual and predicted values comparison from DES and TES

The results from double exponential model show that the predicted values very much align to the test data, that is this model accurately predicts the emissions, with MAE of 7.63. Finally, the triple Exponential smoothing is considered and implemented, this method predicts dynamically using level, trend and seasonality. And the results from triple exponential model indicate an improved performance with MAE of 1.97. The projections for both models are ompared in figure 13.

6.3 Experiment 3 - LSTM:

LSTM stands for long-term memory. It's an architecture or framework that enables recurrent neural networks to store more information. For recurrent neural networks, short term memory refers to the ability to re-use previously learned information. When it comes to the current task, the past knowledge is utilised. For the neural node, that means we don't have a record of all of the past data. In neural networks with recurrent connections, long-term memory is introduced through the LSTM approach. The vanishing gradient problem, when the neural network learning stops because the updates to the various weights in a particular neural network get less and smaller, is mitigated by this. A succession of "gates" is used to accomplish this. Layers of memory are used to store these data, as shown in the figure 14:

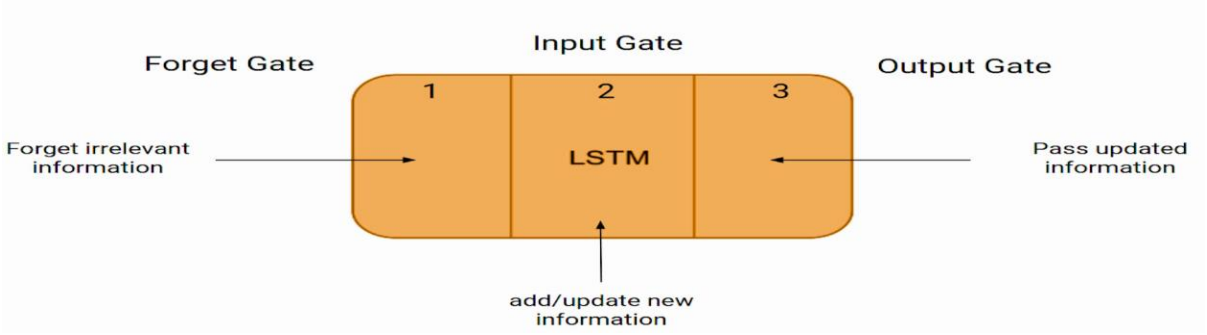

Figure 14. LSTM Architecture

Inside a unit, there are three kinds of gates: Cell Scaling Input Gate (write) Cell Scaling Output Gate (read) To forget: Scales old cell values to new ones (reset) The long-term memory function is incorporated into the model by using each gate as a switch that controls the read/write function. The below graph compares the actual emission value and projected value of emission throughout this 2021, and results MAE of 3.53.

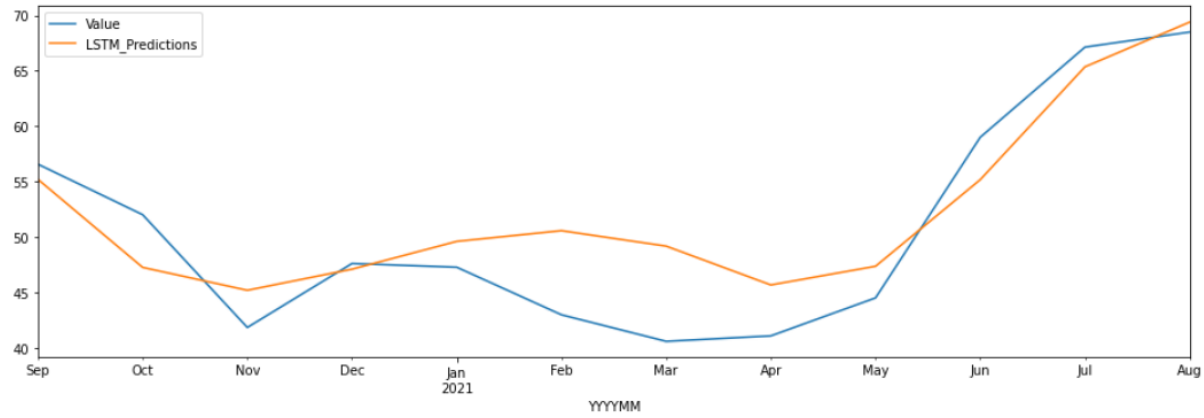

Figure 15. The actual and predicted values comparison from LSTM

6.4 Experiment 4 - Prophet:

Time series data may be automatically detected using prophet modelling, which is why it is so very often used. Python's fbprophet module is used to import the Prophet model. NA Data is removed from the dataset when the date parse() method converts the year column 'YYYYMM' to datetime. It is necessary to alter the column names "YYYYMM" for years and "Value" for values so that the Prophet model can use the data for forecast. The model's forecast frequency is set to 'M,' meaning monthly. The emission values for future dates can be derived from a dataframe created using the .make future dataframe() function and sent as an input to the.predict() method.

$$\hat{y} = \text{trend} * (1 + \underbrace{\text{multiplicative terms}}_{\text{Seasonal features}\quad\text{Exogenous regressors}}) + \underbrace{\text{additive terms}}_{\text{Seasonal features}\quad\text{Exogenous regressors}}$$

The graph below indicates the comparison of actual and forecasted values of carbon emission, and resulted in MAE value of 4.10.

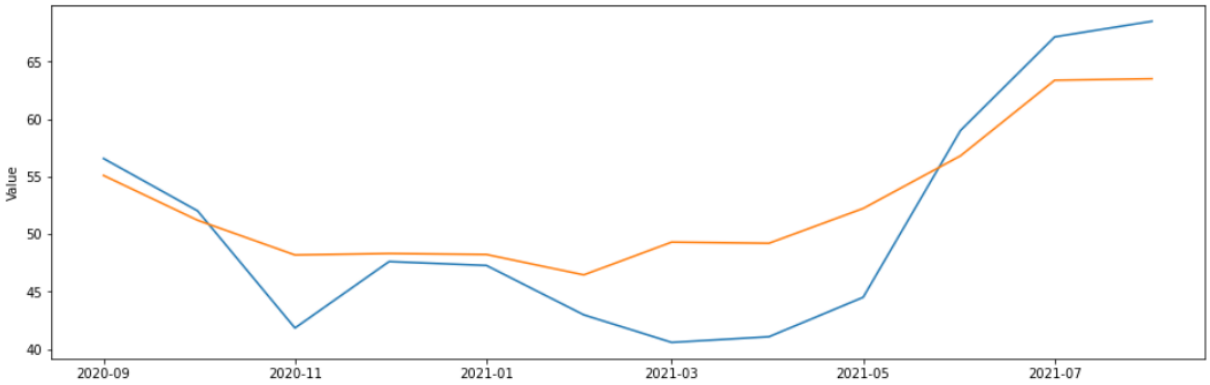

Figure 16. The actual and predicted values comparison from ARIMA

7 Results

Given that the study was conducted using four distinct models, it is required to analyse the models' outcomes using proper metrics in order to find the model with the greatest fit. The model's hyperparameters are produced from the data, which are then modified to have the best match. As an alternative to hit-trail searches, grid searches can save time for hyper parameter optimization. All of the data was split in 90:10 proportion for training the model and testing

it, and this process is followed throughout this research. Calculations are finally made by making a comparison of the actual value to what had been expected.

	Models	RMSE Errors	MSE Errors	MAPE	MAE
0	DES	9.547775	91.160000	14.610182	7.807289
1	TES	2.435051	5.929473	4.030617	1.974551
2	ARIMA	8.570032	73.445449	15.437757	7.164513
3	SARIMA	4.351342	18.934174	7.960879	3.727696
4	LSTM	8.181985	66.944879	7.777783	16.007639
5	Prophet	5.018344	25.183780	4.107524	8.846700

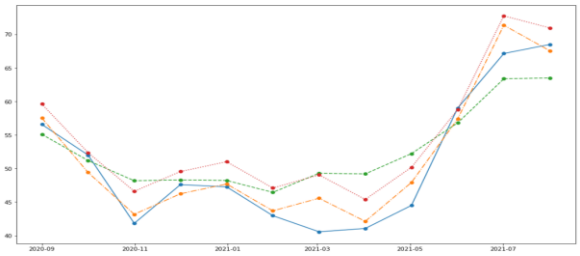

Figure 17. Comparative Analysis of the applied models

The figure 17 above shows the different evaluation criteria values for all the applied models and Triple Exponential Smoothening models provides the best results followed by SARIMA model. The figure on right hand side gives a quick glance in the performance of top 4 models. The green line shows actual value, and orange line shows the prediction made by Triple exponential model. Thus, we can see that the forecasts were correct, which helps us meet the purpose of this research and might assist governments discover an alternative for energy or devise methods to minimize it. About 5 percent higher CO2 emissions are projected to be emitted by the United States' energy sector in 2050 than they were in 2020, according to the EIA. In 2021, we expect an increase of around 7% in energy-related CO2 emissions from 2020 to 2021 as economic activity grows and energy demand rises after the lockdown due to COVID.

8 Discussion

As a result of this study, researchers were able to identify the world's top carbon emitters and the best model for estimating emissions from the world's most carbon emitting sector. Using evaluation metrics, the best model is identified. Our findings are easy to describe. Dueto a paucity of dynamic factors such as fuel prices and energy sector demand, they may be inconclusive in long term. With the incorporation of dynamic parameters such as fuel costs and energy sector power demand, the accuracy of prediction may be enhanced further. World CO2 emissions are expected to fall by roughly 25% by 2050 under a more realistic poverty and demographic growth scenario, which is consistent with current global trends in emissions. Because our estimates are cautious, we are unable to account for recent events in the world economy because we lack the necessary data. Furthermore, the suggested technique may be used to a wide range of real-world prediction challenges in renewable and non-renewable energy fields. According to our findings, reducing carbon emissions and bolstering global climate policy are both viable options for limiting global warming. Developing nations' economic costs and fairness under various burden-sharing arrangements might be examined in further research.

This study will raise awareness of the rising amounts of carbon emissions and assist architects, governments, and other stakeholders devise strategies and policies to minimize them. The results of this research may be used by architectural design firms to calculate their carbon emissions ahead of time, and they can then choose energy-efficient solutions that not only decrease emissions but also boost their productivity and profitability.

9 Conclusion and Future Work

The objective of this research is to determine the best fit model for projecting carbon dioxide emissions from various sectors so that the released data may be used to monitor the sectors and adopt new regulations for the future. The preprocessed data is obtained from the Energy Information Administration(EIA) website and EDGAR websites and entered into the Python environment. Electric power sectors such as the Natural Gas sector, Coal Electric sector, and total Energy Sector have the greatest Carbon emissions as compared to any other sector, according to Exploratory analysis. ARIMA, SARIMA, Exponential Smoothing, Prophet, and LSTM models were used to forecast CO2 emissions in these industries. Models are compared using RMSE, MAE, and MAPE, and the Triple exponential model has the greatest performance. It appears that the emission levels of Natural Gas Electric Power Sector will continue to rise, and this rise can be attributed to low cost of natural gas.

Carbon emissions from urban development projects may be predicted using this study, which is applicable to a wide variety of fields and can be reused over and over again. To achieve a low-carbon built environment, it is critical that all stakeholders work together, and this study does just that. Future studies can take into account other influences on carbon emissions, such as the price of fuel, the GDP, and the degree of trade openness. Developing nations' economic costs and fairness under various burden-sharing arrangements might be examined in further research. In future research, it is recommended to estimate CO2 emissions based on numerous variables such as renewable energy consumption, electricity consumption and other variables in many sectors and industries, including income, urbanization, and international commerce as well as financial development and trade openness. Other greenhouse gases like methane and nitrous oxide may be studied using similar models in the near future.

Ethical Considerations:

EDGAR and EIA websites are used to download this research datasets, and the data is made available to the general public on these websites. Public access to all of the material does not raise any ethical issues. The research will get written authorization for any restricted data retrieval. The supervisor and caretakers shall be informed in the event of a mishap during the implementation phase. Keeping data private while it's obtained from several sources is a major concern. Obtaining very sensitive information necessitates much greater care. Any potential threat is avoided at all costs.

Acknowledgement:

Without the consistent support of my supervisor, Professor Aaloka Anant, this research work would not be possible. His enthusiasm, knowledge, and rigorous attention to detail have helped pave the way.

References:

W. Waheeb, H. Shah, M. Jabreel and D. Puig. (2020), "Ridge Polynomial Neural Network with Error Feedback for Recursive Multi-step Forecast Strategy: A Case Study of Carbon Dioxide Emissions Forecasting," 2020 2nd International Conference on Computer and Information Sciences (ICCIS), pp. 1-6, doi: 10.1109/ICCIS49240.2020.9257685.

Radojević, D., V. Pocajt, I. Popović, A. Perić-Grujić, M. Ristić. (2013) , "Forecasting of Greenhouse Gas Emission in Serbia Using Artificial Neural Networks. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects"

Rajesh Kumar, R.K. Aggarwal, J.D. Sharma. (2013), 'Energy analysis of a building using artificial neural network: A review, Energy and Buildings', Volume 65, Pages 352-358

Quyen Nguyen, Ivan Diaz-Rainey, Duminda Kuruppuarachchi, (2020), 'Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach', Energy Economics, Volume 95, 105129, ISSN 0140-9883

Abderrachid Hamrani, Abdolhamid Akbarzadeh, Chandra A. Madramootoo. (2020), 'Machine learning for predicting greenhouse gas emissions from agricultural soils', Science of The Total Environment, Volume 741, 140338, ISSN 0048-9697

Juan Wang, Sulan Zhang, Qingjun Zhang, (2021) "The relationship of renewable energy consumption to financial development and economic growth in China, Renewable Energy", Volume 170, Pages 897-904, ISSN 0960-1481

Li M, Wang W, De G, Ji X, Tan Z, (2018) "Forecasting Carbon Emissions Related to Energy Consumption in Beijing-Tianjin-Hebei Region Based on Grey Prediction Theory and Extreme Learning Machine Optimized by Support Vector Machine Algorithm". Energies.

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

Dai, S., Niu, D., Han, Y., 2018. Forecasting of energy-related CO2emissions in China based on GM(1,1) and least squares support vector machine optimized by modified shuffled frog leaping algorithm for sustainability. Sustain 10 (4), 958.https://doi.org/10.3390/su10040958.

Li, M., Wang, W., De, G., Ji, X., Tan, Z., 2018. Forecasting carbon emissions related to energy consumption in Beijing-Tianjin-Hebei region based on grey prediction theory and extreme learning machine optimized by support vector machine algorithm. Energies 11 (9), 2475.https://doi.org/10.3390/en11092475.
Huang, Y., Shen, L., Liu, H., 2019. Grey relational analysis, principal component analysis and forecasting of carbon emissions based on long short-term memory in China. J. Clean. Prod. 209, 415e423.https://doi.org/10.1016/j.jclepro.2018.10.128.

Moonchai, S., Chutsagulprom, N., 2020. Short-term forecasting of renewable energy consumption: augmentation of a modified grey model with a Kalman filter. Appl. Soft Comput. 87, 105994.https://doi.org/10.1016/j.asoc.2019.105994.

Pino-Mejías, R., PDC3erez-Fargallo, A., Rubio-Bellido, C., et al., 2017. Comparison of linear regression and artificial neural networks models to predict heating and cooling energy demand, energy consumption and CO2emissions. Energy 118,24e36.https://doi.org/10.1016/j.energy.2016.12.022

Ma, X., Mei, X., Wu, W., Wu, X., Zeng, B., 2019. A novel fractional time delayed grey model with Grey Wolf Optimizer and its applications in forecasting the natural gas and coal consumption in Chongqing China. Energy 178, 487e507.https://doi.org/10.1016/j.

Gao, M., Mao, S., Yan, X., Wen, J., 2015. Estimation of Chinese CO2emission based on a discrete fractional accumulation grey model. J. Grey Syst.-UK. 27 (4), 114e130.

Wu, W., Ma, X., Zeng, B., Wang, Y., Cai, W., 2018. Application of the novel fractional grey model FAGMO(1,1,k) to predict China's nuclear energy consumption. Energy 165, 223e234.https://doi.org/10.1016/j.energy.2018.09.155

Rao, C., Lin, H., Liu, M., 2020. Design of comprehensive evaluation index system forP2P credit risk of "three rural" borrowers. Soft Computing 24 (15),11493e11509.https://doi.org/10.1007/s00500-019-04613-z.