NAO Data Science Internship – Technical Exercise- Simisola Odimayo

Introduction

In adherence to the pivotal 2018 Paris Agreement, the global community collectively pledged to cap the rise in global temperatures at 1.5 degrees Celsius. This commitment entails sustained global governmental efforts to mitigate greenhouse gas (GHG) emission. The United Kingdom has notably demonstrated dedication, with significant initiatives to address environmental impact, resulting in a substantial reduction in GHG emissions since 1990. This proactive stance aligns with the UK's commitment to combat climate change and contribute to the global goal of limiting temperature increases.

In response, I developed a machine learning model analysing decades of UK GHG emissions, examining historical trends, and forecasting future reductions. By harnessing data-driven insights, this study aims to offer valuable support for the UK's climate goals and contribute to the global initiative to mitigate the impact of climate change.

Materials and Methods

Dataset

The data was sourced from the Office for National Statistics, released on October 9, 2023. It encompasses information on carbon dioxide, methane, nitrous oxide, hydrofluorocarbons, perfluorocarbons, sulphur hexafluoride, nitrogen trifluoride, and total greenhouse gas emissions across various industries from 1990 to 2021. The data was pre-processed and analysed both quantitatively and qualitatively. Subsequently, predictive modelling, employing machine learning and long short-term memory (LSTM), was conducted to forecast greenhouse gas emissions for the next five years. This comprehensive approach aims to provide a nuanced understanding of greenhouse gas trends and patterns.

Descriptive Analysis

Data Pre-processing

The data underwent pre-processing and analysis using Python. Initial cleaning involved the removal of non-numeric elements like commas to facilitate processing with pandas and numpy libraries. The dataset was initially analysed in its raw form to assess statistical variations over years. Subsequently, it was transposed, with years as the independent variable on the x-axis and industries as the dependent variable on the y-axis. This transformation provides a structured representation for further exploration of the relationship between years and industries in the dataset. The results can be seen in **Figure 1** and **Appendix 1.**

Time Series Model

Data Selection and Preparation

From the dataset, the years and total GHG emissions were extracted. This would allow for a one-series analysis using the model. The dataset then underwent an 80:20 split for training and test data. The data was then normalised using MinMaxScaler to scale the data.

To facilitate input for the neural network model, a generator was designed, taking inputs from three years to predict results for subsequent years. The choice of three years yielded optimal results among various input variations. Additionally, the number of features remained at one, as the prediction solely relied on total GHG emissions without considering other industry variables.

Model Architecture

The model architecture comprised two layers: an LSTM layer with 100 filters and a dense layer with a single neuron, employing 'relu' as the activation function. The model was compiled using the Adam optimizer, assessing the mean squared error (MSE) as the loss function, and trained over 30 epochs with 80% of the data (from 1990 to 2015).

Subsequently, the trained model predicted values for the years 2016 to 2020 and was then utilized to forecast GHG emissions for the 2022 to 2026.

Results

Descriptive Analysis

Qualitative and quantitative analysis demonstrated a yearly decrease in GHG emissions. **Appendix 1** provides a comprehensive overview, indicating a decrease in mean values, interquartile ranges, as well as maximum and minimum figures, collectively portraying an overall reduction trend. Additionally, **Figure 1** visually represents the declining trend in GHG emissions, highlighting variations among industries, suggesting that some sectors are more successful in reducing their GHG emissions compared to others.

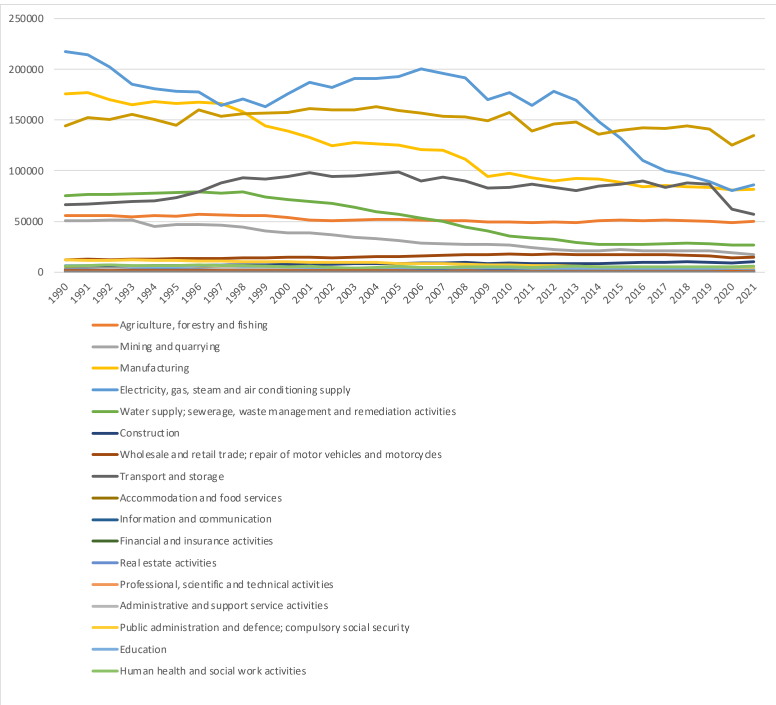

Figure 1 GHG emissions emitted from 1990 to 2021

Model's Predictions

The model was trained on 80% (1990 to 2015) of the dataset and the features and trends learnt by the model were used to predict GHG emissions for 2016 to 2021. **Table 1** illustrates the model's outcomes, revealing a close resemblance to the actual results.

Date	Actual	Predictions
2016	578877.4	591257.3
2017	563160.5	574898.1

2018	564144.5	566201.2
2019	550583.9	559627.7
2020	488596.4	554132.9
2021	502786.6	551010.3

Table 1 demonstrates the predictions achieved by the model vs. actual results of the dataset.

Evaluation of Model Performance on Dataset

The model was evaluated using the Mean Absolute Percentage Error (MAPE). The model performed well and had a low a MAPE of 4.87% indicating that the predictions have an error of approximately 4.87%.

Forecasting

Following a successful evaluation, the model was applied to forecast GHG emissions for 2022 to 2026. **Table 2** showcases the model's predicted values juxtaposed with the provisional values from the Office for National Statistics.

Date	Predicted	Provisional
2022	551010.4	512544.5
2023	548922.9	-
2024	547477.7	-
2025	546657.0	-
2026	546124.1	-

Table 2 illustrates the predictions achieved by the model vs. provisional values from the Office for National Statistics

Discussion

The descriptive analysis revealed a consistent decline in greenhouse gas (GHG) emissions across various industries, as evidenced by reductions in mean, minimum, interquartile ranges, and maximum figures, indicative of the success of climate change initiatives. However, a closer examination via graphs highlighted notable variations among industries, emphasizing a need for targeted efforts. While sectors like electricity, gas, steam, and air conditioning supply exhibited significant decreases, others such as transport and storage, and accommodation and services showed less progress, suggesting a potential focus for governmental interventions.

For forecasting, the LSTM model emerged as the preferred choice due to its proficiency in capturing complex time series patterns and handling sequential data. Despite its suitability for large datasets, a deliberate choice was made for a one-series analysis of total GHG emissions instead of a multivariate approach, considering the complexity and potential errors associated with the latter within a constrained timeframe.

The LSTM model demonstrated high performance, predicting values within a range of actual figures with minimal error. However, it is acknowledged that a multivariate analysis would provide more nuanced insights, especially considering the divergent trends observed across industries. Such an approach could inform more targeted governmental interventions and resource allocation for further improvements in specific sectors.

In summary, the study reveals an overall decline in greenhouse gas emissions, showcasing the success of climate initiatives. While the LSTM model proved effective for short-term predictions, a more detailed multivariate analysis could offer deeper insights for tailored interventions in specific industries, contributing to a more sustainable future.

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
6848	64202.1039	61769.9573	59492.8872	58790.2535	57879.7444	59600.8631	57683.6594	58049.7202	55588.152	56707.3736	58042.1482	56540.5547	57880.1172	58066.0845	57852.9601	57893.9171	56980.0121	55421.6621	50456.7318
76.3	75.1	76.8	78.2	81.9	83.4	81.6	84.3	81	80.5	97.1	89.3	72.6	60.1	64.6	72.9	65.2	54.5	50.1	37.9
2143	40597.9238	39686.5286	38851.8429	38329.9476	38137.2429	39390.181	38634.081	38853.3619	37385.6714	37543.1952	37910.9286	36755.4524	37072.7619	37047.1238	36787.5667	36040.5	35617.4381	34589.2762	31754.7619
638.6	214348.2	202484.4	185006.1	180725	178167	177971.4	166017.9	170829.8	163215	176015.8	186999.9	182150.3	191226.6	191181	192672	200513.5	196066.5	191635.4	169809.7
21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21
978.4	55808.5	55510.4	54725.4	55487.3	55391.7	56706.2	56132.4	55674.8	55680.5	53597.6	51258.6	50997.8	51437.4	52171.6	52218.2	51150.6	49938.7	44604.6	40319.4
151.5	6471.1	6828.4	6124.1	6391	6424	6677.6	6857.2	6284.4	5891.6	5218.7	5526.6	4549.3	4018.7	4616.6	5652.2	4734.6	4705.7	5114.6	5025.2
797.7	1885.4	1712.9	1793.5	1739.3	1737.8	1793.2	1629	1550.4	1483.6	1470.7	1546.1	1382.8	1427	1405	1417.1	1231.4	1167.7	1286.8	1107.5