National Audit Office

NAO Data Science Internship

Technical Exercise

Report by ASHISH KAKRAN

MSc (Data Science)

Teesside University

9 December 2022

The report can be accessed online here:

nao-internship-challenge

Summary

Introduction

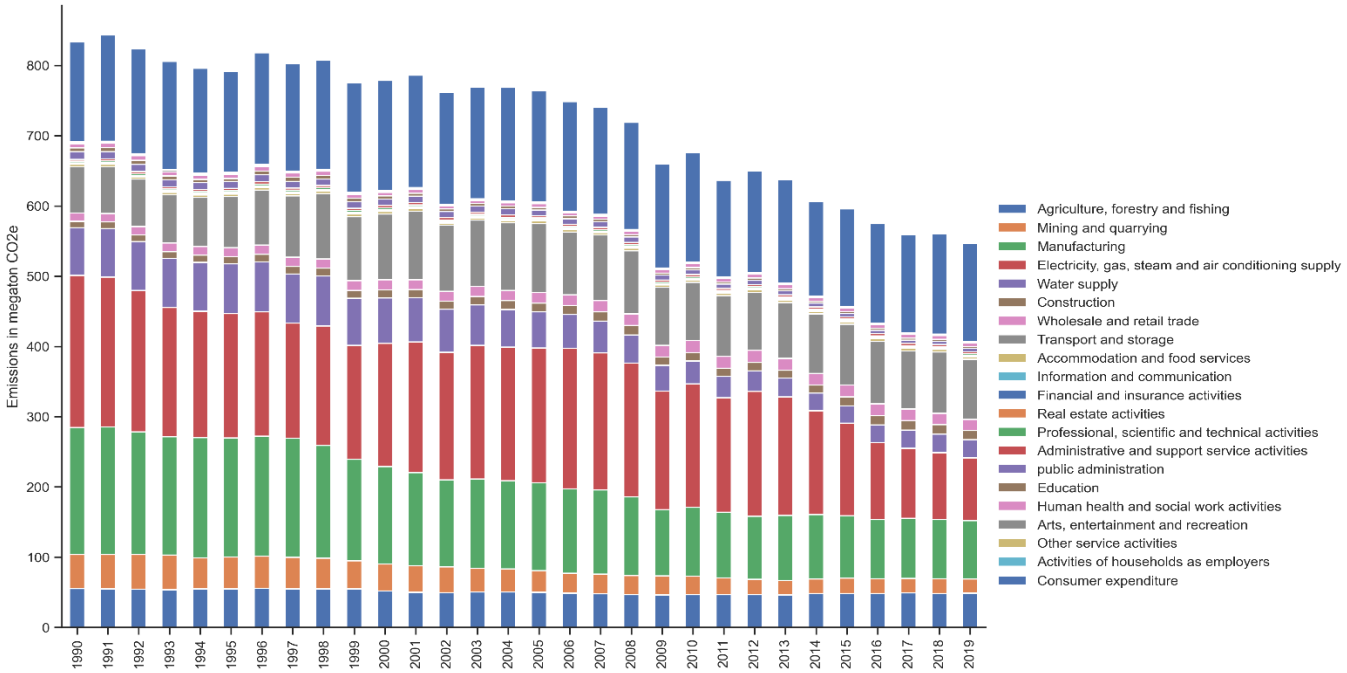
Over the last two decades, office of national statistics has collected atmospheric emissions (Clarke and Louise, 2022) of greenhouse gases (GHGs) across United Kingdom on residential basis. This data was analysed to forecast GHG emissions over the next 5 years.

The data was separated into two groups namely residential group and SIC group, latter being a more detailed version of the former. However, for modelling purposes only residential data was used since SIC group is subset of residential data and thus doing so avoids the need of training sophisticated nonlinear models when linear models can work reasonably accurate. The data is clean, and no anomalous values were found.

During the exploratory analysis, it was found that annual GHG emissions over the last two decades across all the sectors have been decreasing and thus UK seems well on track to meet the target of net zero emissions.

Figure 1

Annual GHG emissions across UK on residential basis (1990-2019)



The detailed residential data was summarized into a series of annual emissions for each year ignoring what sector these emissions came from.

Analysis

Data given here is time-series data (peixeiro, 2019) where annual emissions of GHG is given for each year from 1990 to 2020 inclusive. The GHG emissions value of year 2020 was not used in modelling since emissions were low due to covid-19 and using that value would have resulted in biased model. However, year itself cannot be used as feature to predict GHG emissions since these two are not correlated. That is, given a year, we cannot reasonably predict what the emission during that year will look like knowing only the year itself.

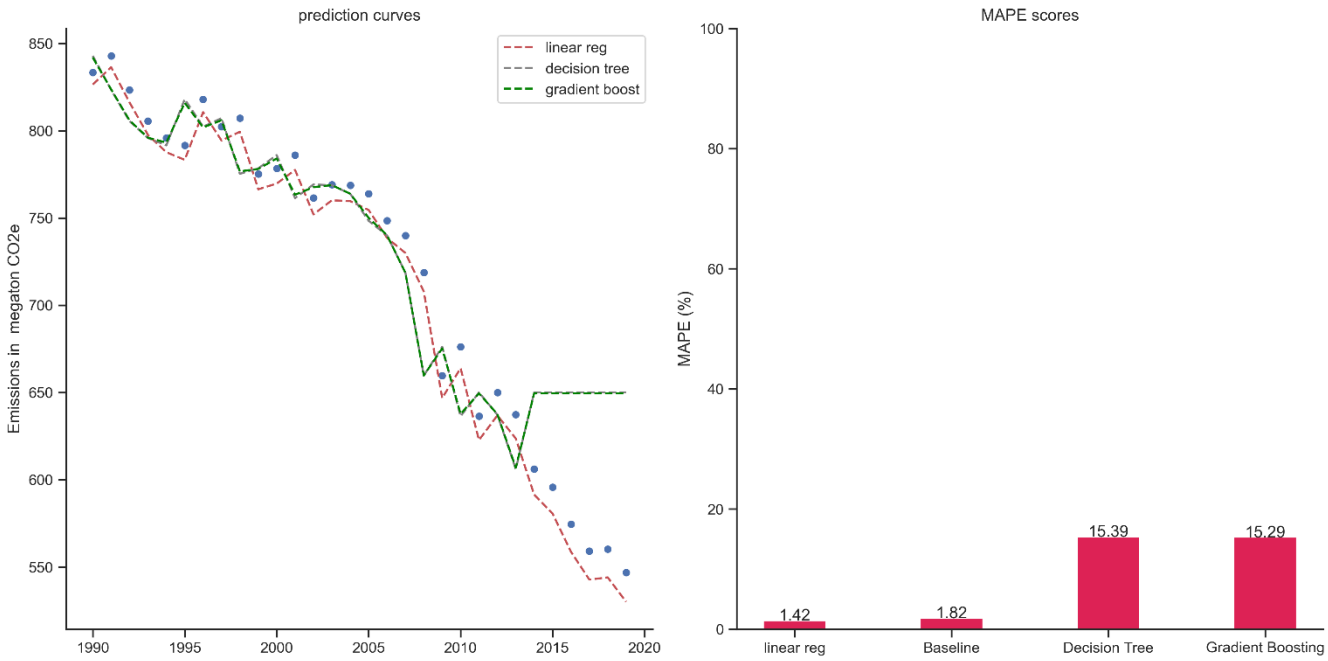
Emissions of a given year can however be used to predict emissions in next year as emissions do not change rapidly, unless something special happened during that year. For example, a rare case of covid-19 caused emissions to decrease which is why value for 2020 was neglected.

Due to lack of availability of enough data, only value of one year was used to predict value of next year. Alternatively, emissions values of series of previous years could have been used to predict emissions in next year. This gives a basis for simple baseline model where emissions value of current year can be used as it is for the next year. This baseline model is remarkably accurate and has mean absolute percentage error of 1.82.

Three models including linear regression, Decision tree regressor and gradient boost regressor were trained and compared. The mean absolute percentage error (MAPE) was used to compare these models and only linear regression was able to beat the baseline model. The R2 score of linear regression was highest at a somewhat reasonable value of 0.62. Thus, only linear regression was able to capture 62% of the variance in the data.

Figure 2

Comparison of trained models

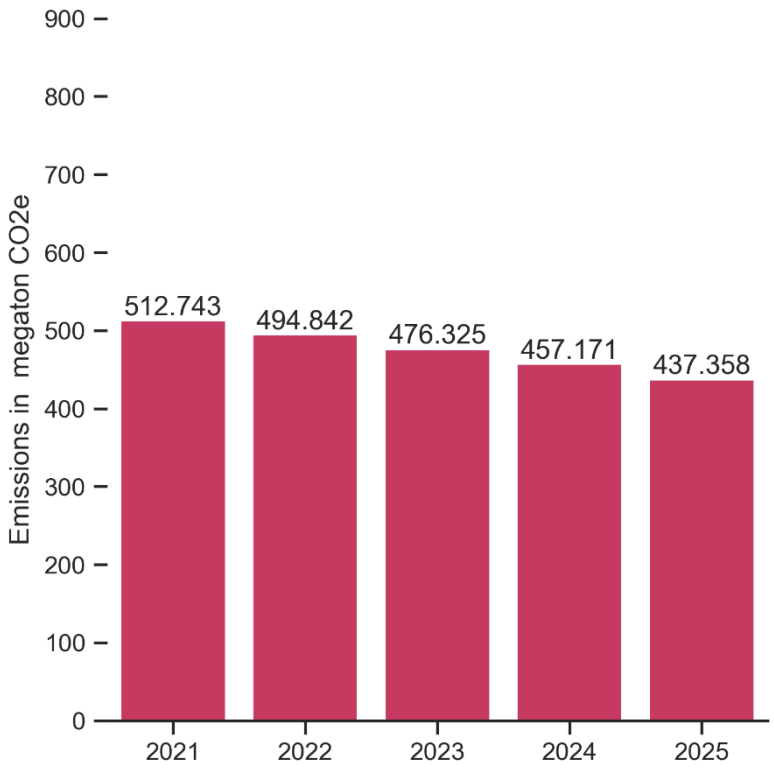


Forecasting

Linear regression model was then used as hypothesis to forecast the GHG emissions over the next 5 years. The GHG emissions are consistently decreasing by at least 10 megaton of carbon equivalents each year.

Figure 3

GHG emissions forecast (2021 – 2025)



Future Scope

More data is required to use powerful nonlinear models for better prediction as the current nonlinear models are simply overfitting the data. Instead of observing emissions once a year, these can be read each month to generate more data.

Alternatively, the current model trained here can be evaluated on validation test and be regularized accordingly along with hyper tuning to better predict the emissions. In addition, confidence interval can be made about confident the models are when it comes to forecasting emissions values for next five years.

References

Clarke & Louise (2022) *Atmospheric emissions: greenhouse gases by industry and gas.* Available at: https://www.ons.gov.uk/economy/environmentalaccounts/datasets/ukenvironmentalaccountsatmosphericemissionsgreenhousegasemissionsbyeconomicsectorandgasunitedkingdom .

Peixeiro, m. (2019) *Complete guide to time series analysis and forecasting.* Available at: https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775 .