Greenhouse Gas Emission Prediction

Katherine Williams 17kew2@queensu.ca
Molly Shillabeer 17mes10@queesu.ca

Table of Contents

Abstract 1

Introduction 1

Problem Statement 3

Proposed Method 3

Experimental Results 4

Summary or Conclusions 5

References 5

Abstract

Climate change has rapidly become an important topic as it causes destruction worldwide; using a model that can predict the future greenhouse gas (GHG) emission rates with other global development indicators can help find patterns within the GHG emission rates. An LSTM model was created to predict the GHG emission rates due to its ability to handle the time dimension of the data. The model predicted around 40% of the future GHG emissions correctly. Although the score is not optimal, with a larger indicator dataset and more recent GHG emission measurements, the model prediction rate should increase.

Introduction

Climate change has over time become one of the most important issues in politics and could harm people worldwide. The world has already seen some of the resulting effects, with natural disasters becoming more common, dangerous, and irregular. Recently, a snowstorm in Texas has demonstrated how unprepared many countries are to handle unpredictable and extreme weather. As a result, most countries have introduced emissions targets to try and reduce greenhouse gas emissions to avoid the worst of the consequences. Although Canada has had legislation regarding greenhouse gas emissions for more than 20 years, emissions continue to rise (figure 1).

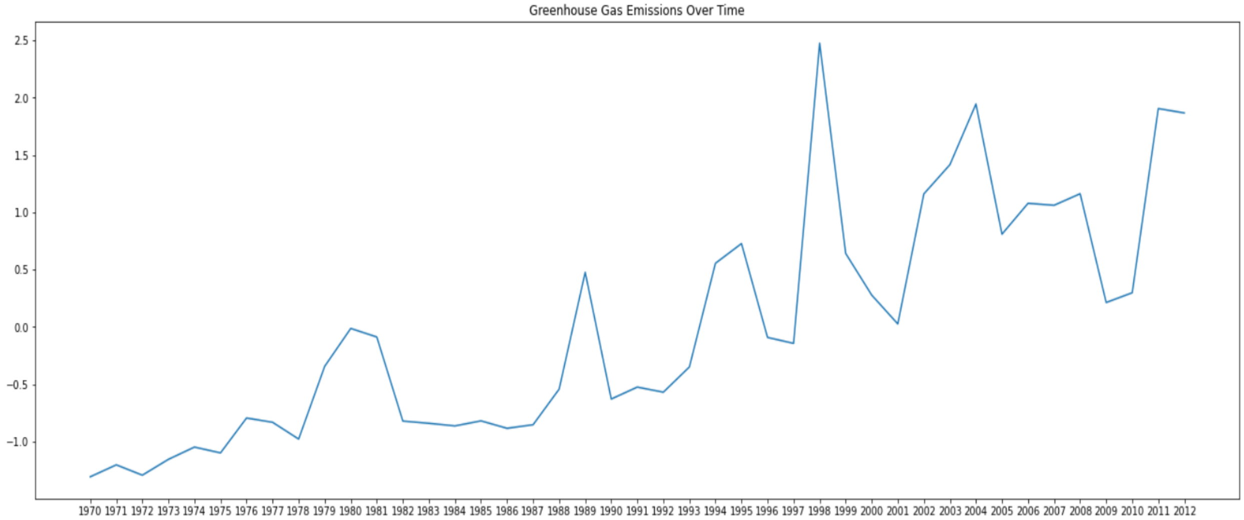

Figure 1: Graph showing the amount of greenhouse gas emissions for each year from 1970 to 2012

Canada's current target is to reduce GHG emissions 30% by 2030, compared to 2005 levels. Accurately measuring greenhouse gas emissions can be difficult, but there are also many other factors that influence emissions and are easier to measure. For example, a larger population will always create more greenhouse gases than a smaller population if other factors are equal. Many global development indicators can show similar patterns to emissions, or impact them directly. We chose to focus on indicators related to population, economy, electricity, and healthcare. Having a model that uses these indicators will allow us to predict what GHG levels should be without extra measures being taken. Knowing this information means that the actual GHG levels can be compared to predicted levels to see how effective the Canadian government has been at meeting their targets. With this model, we can also show how indicators are closely related to GHG emissions, and these indicators may show opportunities for further reducing emissions.

There are a number of difficulties in creating a model to achieve these goals - the most important being the lack of data and that each indicator can be measured in many different ways. Greenhouse gas emissions have only been tracked by the World Bank since 1970, meaning that there are under 50 samples of data that can be used for this model. This is a very small dataset, which limits the ability of the model to identify patterns. Another issue is that greenhouse gas emissions can be measured in many different ways, using different gases or including different activities. The measure used in this report includes forest fires, peat fires, and human sources of emissions, but excludes agricultural waste burning and savanna burning[1]. These issues also apply to the global development indicators, for example, there are dozens of ways to measure a country's economic output. This means that to use the model in the future, users will need to ensure that all indicators are measured in the same way the model was originally built for.

With this model, Canada will not only be able to see if their efforts to reach their goal are effective but they can also see how other aspects of their country could be an indicator of the GHG emissions rates. The indicators can be used as a warning to countries that their GHG emission rates may soon spike. We hope that this model can be used as a quick check to assess how high GHG emission rates are and to be used as a preventative measure to lower GHG emissions. This model can also be used to identify patterns that may cause the GHG emissions to spike in the first place, and identifying them can help decrease the overall GHG emission rates.

Problem Statement

GHG emission rates have been rising and causing destructive natural disasters all over the world. As much as each country has tried to monitor and keep track of the emission rates it is not enough to solve the global crisis. If there was a new factor measured within a country that could be an indicator for predicting the rates of GHG emission, it would give an immediate inside look that would allow for action to be taken faster than before.

It is clear that measuring GHG emission rates is not an easy task, there is not a large database of information that can make it difficult to predict future emission rates. By creating this model that can look at other global development indicators related to population, economy, electricity, and healthcare we can use indicators for GHG emissions that are easier to measure and have more data available. This will allow for a more accurate prediction of GHG emission rates and for patterns of the emission rate to emerge, which will help create impactful change.

Proposed Method

Global development indicators from The World Bank Open Data were used for this model. 118 features were selected from common development indicators that Canada tracks yearly. The feature being predicted is total greenhouse gas emissions in kilotons of CO_2 equivalent. The data used was for the year 1970 to 2012.

The first step in creating our model was preprocessing the data above. The features all had different units of measure, so the data was standardized to ensure that features with larger numbers would not be over-represented in the model. Missing values were filled with the median value of their feature. Initially, each feature was plotted to show the change from 1970 - 2012, and features that stayed the same through these 42 years were removed from the training set as it may have skewed the model to be less sensitive to the change of other features. Then the data was split into training and test sets.

Since this was a regression problem, a number of models were considered, including XGBoosting, LSTM neural network, and Facebook Prophet. XGBoosting was not chosen because it did not account for the time within the data. Facebook Prophet was an option but it required the datasets to be in a unique format that would not have been possible with the number of features that were in the training data. Facebook Prophet is also intended for data that shows seasonality, which does not apply to the data for this problem, which is reported on a yearly basis. LSTM was selected as the best model to use because it could account for the time dimension of the data.

The LSTM model was created with two LSTM layers; however, there were initial issues with the sizing of layers which could be solved with a layer to extend the dimensions and setting the GPU to allow growth. The LSTM was then trained with the metric mean squared error (MSE) because it shows the difference between the training and the testing data, so the lower the MSE the better the training data was able to predict the values of the testing data. The LSTM model was tuned to have a lower MSE loss rate to optimally fit the model to the data, without overfitting.

The R-squared value was used to show how well the model was able to predict the y_test data. This was used because it is able to assess how well a model fits in regression since it measures the percent of dependent variable variation within the model[2]. This indicates the lower the R-squared value the more error there is in the prediction. The optimal value that would show our model is accurate would be 1 and anything less than that shows worse performance. If overfitting were to occur then the model would not accurately measure any new data.

Experimental Results

The model resulted in good performance on test data. R-squared was the statistical measure used to evaluate the model. This measure represents the goodness of fit between the predicted values and actual test values, on a scale of 0 to 1. The final R-squared score achieved was about 0.47. This means that about 47% of the test data fit well to the actual data. Figure 2 shows that the model can identify general patterns in the data and fit test data reasonably.

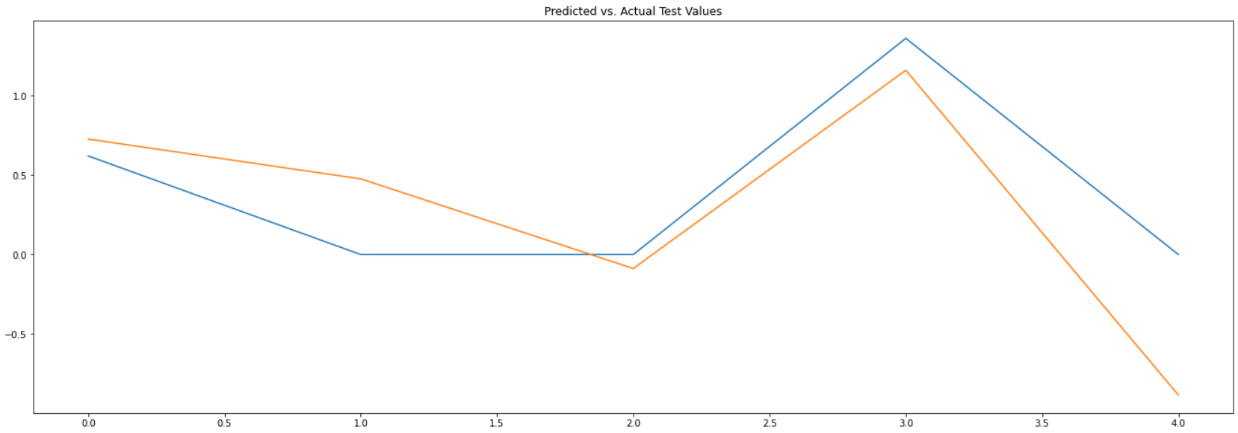

Figure 2: Diagram showing the predicted test values (blue) compared to the actual test values (orange).

Summary or Conclusions

After completing the regression model for predicting GHG emission from global development indicators, it is clear that the emission rates can be predicted using other global indicators. This shows that by monitoring indicators within the population, economy, electricity, and healthcare Canada will be able to roughly predict the future rates of GHG emissions. This may prove to be more useful than using previously recorded values of the GHG emissions since they only began recording the values in 1970. With such a small dataset, it may not be the more accurate method.

By using indicators from global development our model was able to accurately predict about 40% of the future GHG emissions. This prediction score is not ideal because we did not have a large enough database of global development indicators and needed more yearly measurements for each indicator. There are also limited measurements of GHG emissions ranging from only 2012 - 1970, if we were able to have access to the measurements taken in 2020 - 2013 then it may have increased the prediction rate. Although the prediction score is low, it shows that the idea of predicting future GHG emission rates based on other global development indicators may be possible with a larger dataset of indicators and with more measurements of GHG emissions.

References

[1] Total greenhouse gas emissions (kt of CO2 equivalent). (2021, March 19). Retrieved from https://datacatalog.worldbank.org/total-greenhouse-gas-emissions-kt-co2-equivalent-0

[2] Frost, J. (2021, April 13). How To Interpret R-squared in Regression Analysis. Retrieved from https://statisticsbyjim.com/regression/interpret-r-squared-regression/