Determining the Relationship of Region, Income, and Carbon Emission Rates to Average Country Temperature

Deja Dunlap, Nicole Tian, Fiza Shakeel

Problem statement:

Climate change is a global problem whose effects grow more noticeable with every year. One of the most pressing concerning trends scientists have started to notice is that the countries that are contributing the most to it, through excessive CO2 emissions, are being less affected than countries that are not. To examine this more closely, we would like to understand the relationship between the global surface temperature from 1961 to 2021, country income groups, and country carbon emission rates in predicting average country temperatures in 2022.

Data

We used two datasets. For country average temperature we are using the Global Surface Temperature Change from 1961-2022 dataset from Kaggle. For tracking country global carbon emission we are using the CO2 emission per capita dataset from the Oxford Martin School. From this, we will merge the two datasets and reconstruct the data so that it contains the following variables of interest:

- ISO3 Codes: Unique Identifier
- Country Name: Unique Identifier
- Area and Density: Continuous
- Sub-Region: Categorical
- Income Group: Categorical
- Annual Surface Temperature for each Year from 1961-2022 (61 separate columns): Continuous
- Country Carbon Emission Per Capita: Continuous
- High Emission Contributor: Categorical

This new dataset is useful, because it is filled with complete values and contains the information we want to be able to predict correlation between socioeconomic variables, like country income group and carbon emission per capita, and the country's global warming effects.

Methodology

This project used a Linear Regression model to understand whether rising global surface temperatures are disproportionately affecting certain countries. We decided to use a linear regression model because we wanted our variable of interest to be 2022 country temperature - a continuous value. Given this our hypothesis was:

$$h(x) = w^T x \; such \; that \; w \in R^{N \; X \; D}, \; x \in R^{N \; X \; D}$$

Where x represents the feature vector (the data) and w the weights. And our objective function was to

Minimize the distance between y and $h(x) = w^T x$

We can measure this distance using the following objective loss function, using the following loss equation

$$l = (1/N) \sum_{n}^{N} (h(x^n) - y(x^n))^2 = 1/N \, ||h(x) - y||_2^2$$

With some derivation becomes

$$1/N \, ||Xw - y||_2^2$$

Thus, we can find our best hypothesis by finding the one that minimizes loss, getting the following objective function

$$argmin \, l(w) = 1/N \, ||Xw - y||_2^2$$

To use the best possible hypothesis for testing, we implemented k-fold cross-validation testing, allowing for less bias within the model and avoiding overfitting to the test data.

Implementation

Given inputs of an average temperature from 1961 to 2021, country income level, and carbon emission levels, we predict the average temperature per country in 2022. We created three different datasets to be used on our Linear Regression to examine the effect of each of these variables:

1. The first was solely trained on the average country temperatures from 1961 to 2021 and country region, as these are the two variables most obviously tied to 2022 average temperature.
2. The second was trained with the country average temperatures, country region, and country income level.
3. The third was trained with the country average temperatures, country region, country income level, as well as country carbon emission level per capita.

Because the dataset was so small (only ~140 samples), we decided to use k-fold testing to increase model accuracy. We tested the models out on ks of size 2 to 10 to see which iteration gave us the best r-squared value and MSE score. We also ran the datasets through the scikit learn Linear Regression model as a sanity check to ensure that we were implementing everything properly. Then we calculatd mean squared error and r-squared of both the scikit and our model to determine which model most accurately predicts the 2022 country average temperature levels. Through this, we were able to determine if country income level and carbon emission levels are playing a role in determining the effects of climate change by looking at whether including those additional variables increased model accuracy.

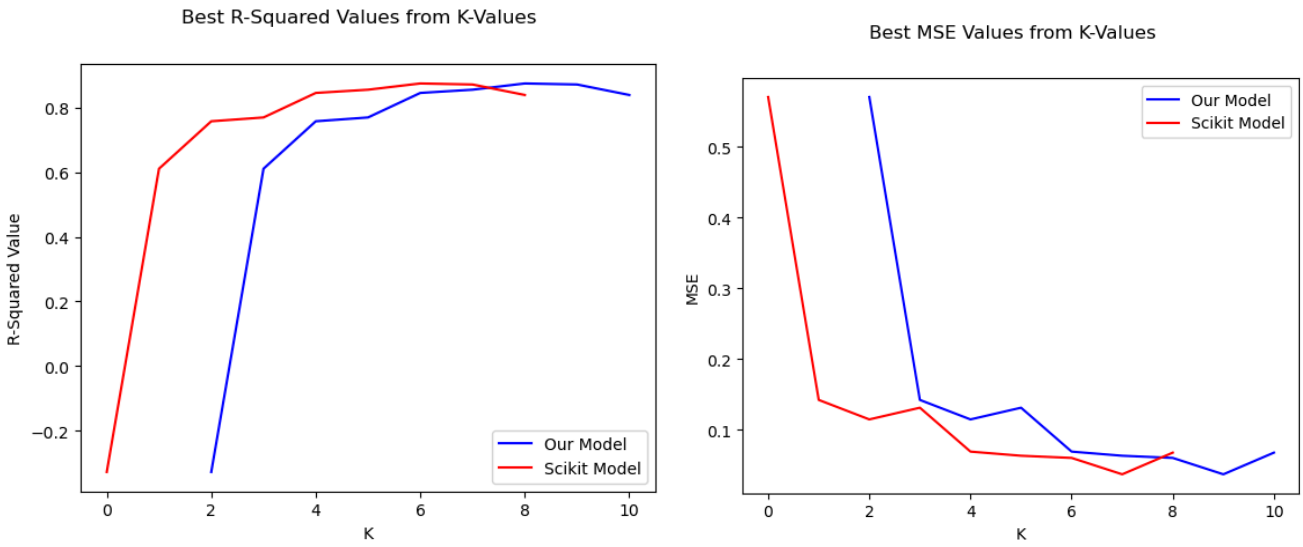
The biggest challenges we faced was setting up the data to be in a format that was usable for our machine learning model. In class we used pre-processed data that already came in a format easily readable to the

model. Conversely, the data we used for our model was a public dataset and required an intensive amount of cleaning and re-coding in order to be usable by our model.

Results

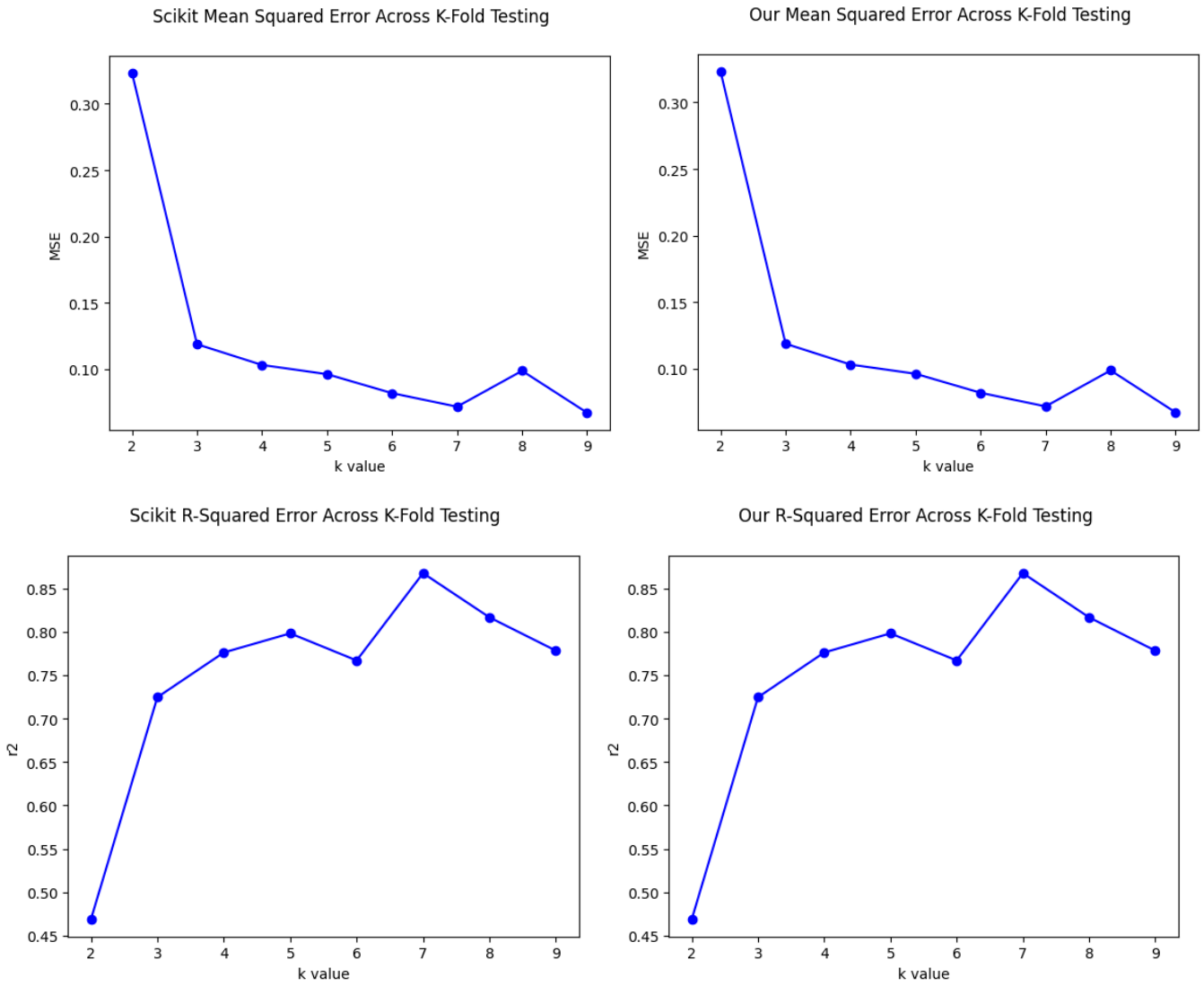
Dataset 1

The accuracy for our model far exceeded the expectations we laid out in our proposal. There was a general trend of increasing r-squared and decreasing MSE the more k-folds we used, as is to be expected when using k-folds. The highest r-squared when running it for dataset 1 was 87.4% (technically 0.8744148913884162) and our lowest MSE was 0.03% (technically 0.03681877429759664). This is in comparison to the sci-kit model whose highest r-squared was also 87.3% (technically 0.8744148913884302) and lowest MSE was also 0.03% (technically 0.03681877429755837). Here's the trend for r-squared and MSE score across various ks.



Figures for Dataset 1, trained on country average temperature and country region

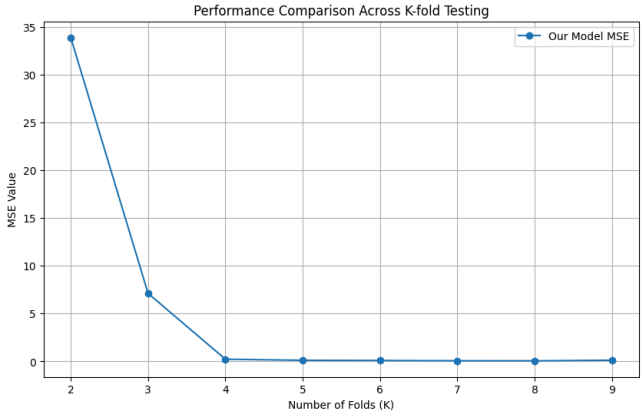
Dataset 2

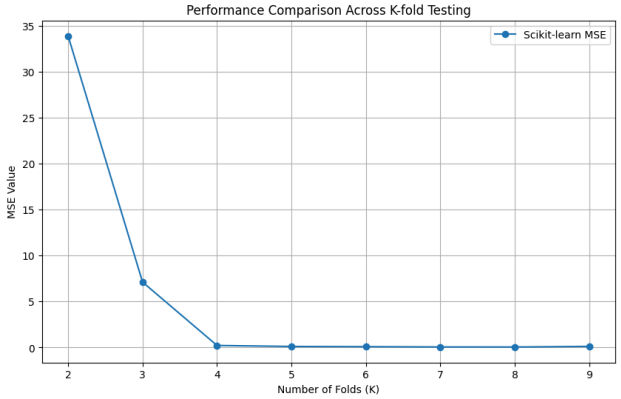


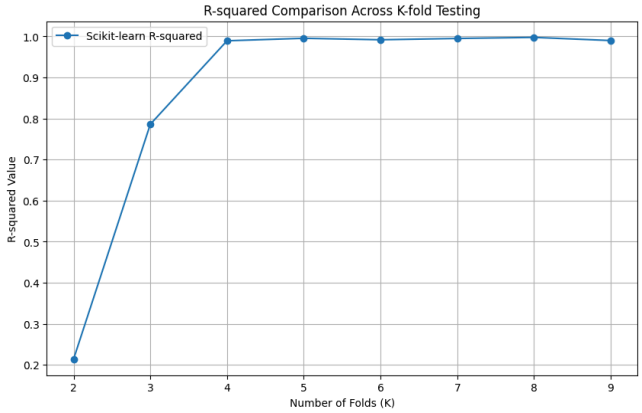
Figures for dataset 2, trained on country average temperatures, country region, and country income level.

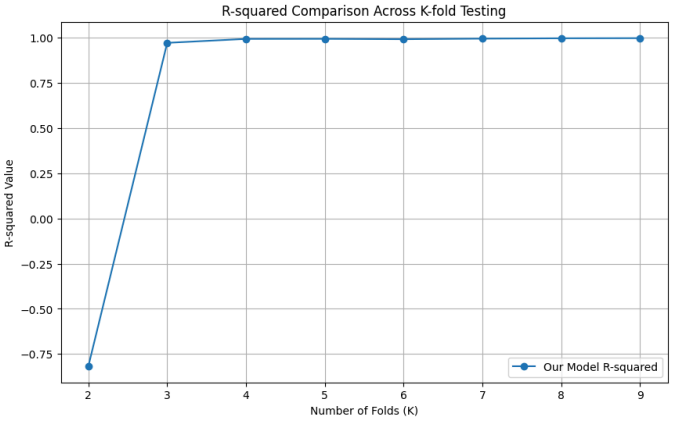
Dataset 2 was trained on country average temperatures, country region, and country income level. When compared to the scikit model, our model achieved equal mean squared error and r-squared scores for all k values from k = 2 to k = 9. The general MSE and r-squared score trends indicate a similar pattern to our results on the first dataset, with MSE decreasing and r2 increasing as we increased the number of folds. These results make sense, since more folds increases the amount of training data and reduces variability in the data. Dataset 2 achieved a maximum r2 score of 0.8675321689916956 where k = 7 and corresponding minimum MSE score of 0.07194561075577775 at k = 7.

Dataset 3









Figures for dataset 3, trained with the country average temperatures, country region, country income level, as well as country carbon emission level per capita.

For R-squared values across different values of K, both the Sci-kit model and our model achieved nearly the same results, as shown below.

Sci-kit:
[-0.6867524602919112, 0.8632630983581269, 0.6372738104707478, 0.6932691695149036, 0.9492342814034673, 0.6593520276818577, 0.9091508274014752, 0.8744192575695767]
Our model:
[-0.6867524206990372, 0.8632630985844362, 0.6372738105479858, 0.693269169814796, 0.9492342815064087, 0.6593520278442563, 0.9091508273993071, 0.8744192574633343]

For MSE values across different values of K, both the Sci-kit model and our model also achieved nearly the same results, as shown below.

Sci-kit:
[450.0533409047021, 11.077731839022292, 12.186155575166563, 35.06376340997255, 2.244797478981679, 12.893526851120876, 5.2938730827963605, 42.51190651693941]
Our model:
[450.0533296641848, 11.077731943849637, 12.186155579300017, 35.06376333941954, 2.244797478668192, 12.893526842487077, 5.293873083079649, 42.511906517983675]

Overall, both the Sci-kit model and our custom model performed comparably well on this dataset, achieving similar performance in terms of both R-squared and MSE values. The initial negative R-squared values indicate that both our model and sci-kit's performed suboptimally at first – that the data fit worse than a horizontal line – though eventually both models improve at the same rate.

Discussion

The model that had the highest R-squared value and the lowest MSE score was the model trained on dataset 1, with just country sub-region and country average temperature. This could be because, as we suggested earlier, these are the variables most obviously correlated with country average temperature in 2022. The model trained on dataset 2, with counxtry sub-region, country average temperature, and country income level performed almost as well, but this could be because the additional variable of country income level had nothing to do with country average temperature and thus the model was still able to make predictions with just the original two variables. The third model, trained on dataset 3 with country subregion, average temperatures, country income, and carbon emissions level. The MSE scores and r-squared negative values for the earlier k-fold values from both the scikit's and our model would suggest that the extra layer of carbon emissions level negatively affects the model's ability to predict country average temperature in 2022. This could be because, while carbon emissions level affects global temperatures, it might not have such an effect on a country level. Recently in the news, just because countries like America and China produce more carbon emissions, we see far more devastating climate disasters and climate refugees from countries like Haiti that have significantly lower carbon emission levels.