Predicting Extent of Arctic and Antarctic Sea Ice Cover

Sidney La Fontaine, Esther Anemeje, Dylan Sapienza

Abstract

Global warming is a worldwide crisis that continues to grow more and more severe. In an attempt to understand its global effects, we used regression models and neural networks to predict the extent of the Arctic and Antarctic sea ice cover, the area of the ocean where there is at least some sea ice, from important climate features including the average monthly global temperature, CO_2 levels, and the number of natural disasters that occurred on the day. Hopefully, this project can be a reliable predictor of future effects of climate change and be a warning for the dangerous path humankind is currently on.

Introduction

Global warming is a problem that is becoming worse every day. It is going to have more and more drastic effects on humankind unless it is addressed. Our project revolves around demonstrating the effect global warming, through important climate features, is having on the extent of the Arctic and Antarctic sea ice cover, which is a measurement of the amount of the ocean with at least some sea ice. The extent of the Arctic and Antarctic sea ice cover is determined by the area of the ocean made up of at least 15% sea ice.

In general, the goal of our project is to illustrate the severity of our current and future climate crisis. We will be using a number of important climate features including the average monthly global temperature from Berkeley Earth, CO_2 levels from the Scripps Institute of Oceanography, and the number of natural disasters that occurred on that specific date according to the Center for Research on the Epidemiology of Disasters to predict the extent of the Arctic and Antarctic sea ice cover.

Technical Approach

We trained several machine learning models to predict the extent of the Arctic and Antarctic sea ice cover. We used regression models to implement this task. First, we used Lasso regression to predict our target variable because Lasso would eliminate the effects of variables that are not relevant in the computation by setting their coefficients to 0. We also used Ridge regression to assess the effects of not doing the feature selection present in Lasso regression but simply penalizing features deemed less irrelevant.

We then implemented two neural networks models. The first was a Linear regression model while the second was a non-Linear regression using sigmoid for the activation function. With these, we tested to see how the size of the first and only hidden layer's neuron count affected results.

Experimental Results

Data Pre/Post-Processing

There wasn't an all-encompassing dataset containing all the climate features we needed for this project, every reputable climate dataset we found only contained one viable climate feature we needed for our project. So we had to spend quite a lot of time finding datasets containing vital features for our dataset. Specifically, we used the average monthly global temperature dataset from Berkeley Earth, the daily extent of the Arctic and Antarctic sea ice cover dataset from NSIDC (National Snow and Ice Data Center), the daily CO_2 level dataset from the Scripps Institute of Oceanography, and the natural disaster dataset from EM-DAT (Emergency Events Database).

To put together our complete dataset we had to determine the set of days that we had a value for every feature for. Then we randomly picked 2,205 days from that set to be the dates for our dataset. Next, we joined the average monthly global temperature dataset, the daily extent of the Arctic and Antarctic sea ice cover dataset, and the daily CO_2 level dataset on those randomly chosen dates. The last step to complete the dataset was to determine the number of natural disasters that occurred on each of those randomly chosen

dates according to the EM-DAT dataset. This last step was quite challenging due to how inexact the start and end dates of natural disasters are, the number of dates we had to determine the number of natural disasters that occurred on those dates, and the size of the EM-DAT database. After adding the number of natural disasters feature our dataset was complete.

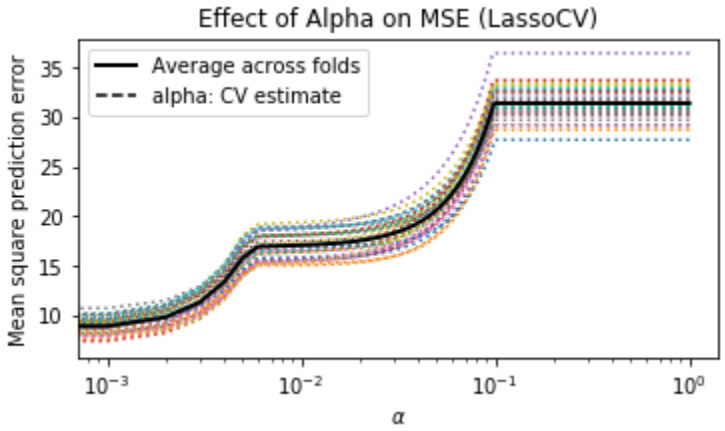
Lasso and Ridge Regression

For the Lasso and Ridge regression implementations, we decided to use LassoCV and RidgeCV since they have built-in cross-validation functionality while making use of regularization. First, we converted the dates to ordinal values for compatibility. We then split the data into training and testing sets with a 75%-25% split. We then used RepeatedKFold as our cross-validator for the models with 10 splits and 3 repetitions. We set our target variable to be the sea_ice_extent. We performed hyperparameter tuning for both models to find the optimal alpha. The list of alphas used by our models included every number from 0 to 1 using increments of 0.001. The models then performed cross-validation using all the values for alpha and selected the model with the lowest R^2 value.

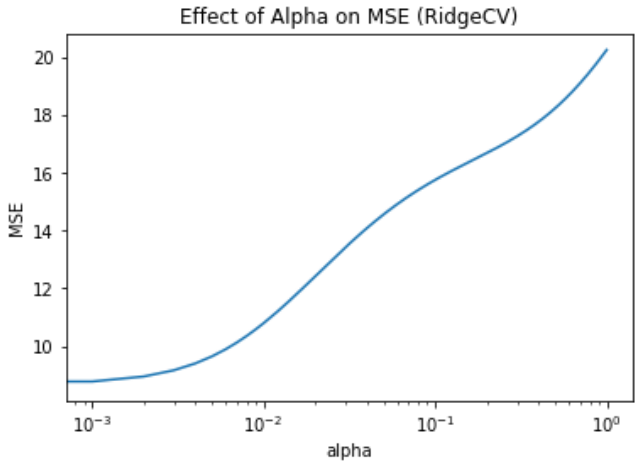
The models had nearly identical results. This is likely due to the fact that there were not many features and there was not a lot of collinearity present. Both selected 0 as the optimal value for alpha. They both had 0.72662 as their R^2 value for the training set and 0.71475 for the test set. We also calculated the Mean Absolute Percentage Error to use as a measure of accuracy. This value represents the average of the absolute percentage errors of each entry in a dataset, showing, on average, how accurate the forecasted quantities were in comparison with the actual quantities Using this value, both models had an accuracy of 72.5%. This suggests that the models were pretty accurate but it would have been interesting to see how the models would have differed if they had more features to use for prediction.

Feature	Date	Average Temp	CO2	Number of Natural Disasters
Coefficient	0.005444	0.958163	-1.088449	-0.002879

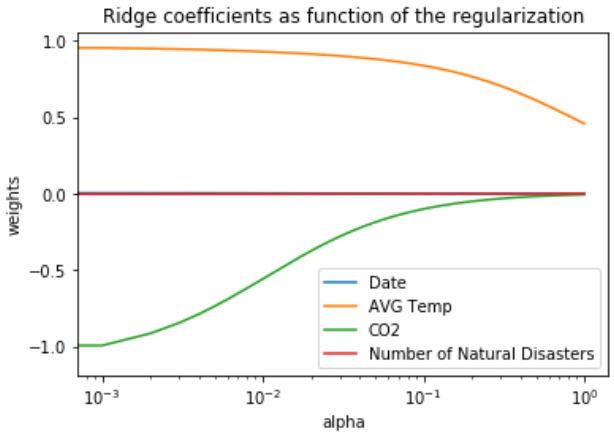
This table shows the coefficients that the models assigned to each feature. Average temperature is the only feature that has a strong positive correlation with the sea_ice_extent while CO2 has a strong negative correlation. The other 2 features barely have any weight in computing the predicted sea_ice_extent.



This graph shows how the Mean Squared Error (MSE) varied with the alpha values across the folds of the LassoCV regression.



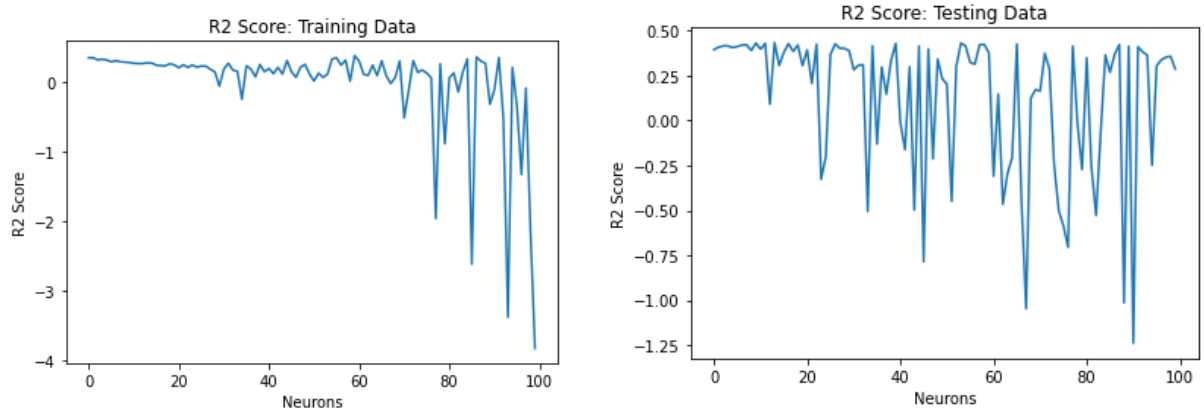
This graph shows how the Mean Squared Error (MSE) varied with the alpha values for the RidgeCV regression.



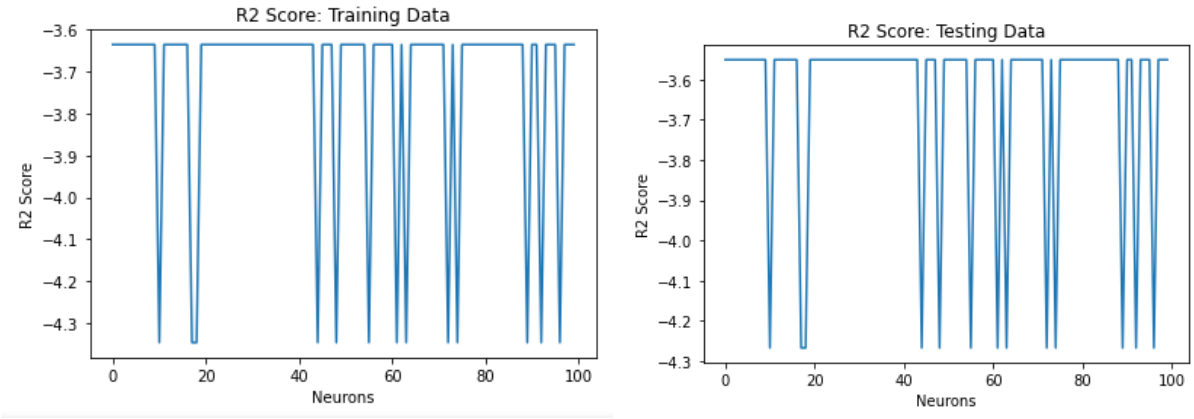
This graph shows how the weights of the features varied with the alpha values for the RidgeCV regression.

Neural Network

For the neural network model we followed similar implementation ideas compared to the LassoCV and RidgeCV. We kept the 75%-25% split of the data as well as having the target variable be the sea_ice_extent. For Neural Networks we implemented both a linear and non-linear fashion. Where the non-linear model used sigmoid as an activation function on the final layer. Both models were trained using 10 epochs where each epoch went through all the data. The optimizers in both models were Adam. In order to find the optimal model we used 1 hidden layer and increased the neuron count from 1 to 100 checking the Mean Squared Error, R^2 value, and Mean Absolute Percentage Error. Similar to our Lasso and Ridge implementation we found the model that had the lowest R^2 value.



Linear Models (3-X-1 Architecture)



Non-Linear Models (3-X-1 Architecture)

The results for the Linear models show the lowest R^2 value for the training set (using 58 Neurons) as .0087 and the lowest R^2 value of the test set as .1228 (68 Neurons). The non-Linear model got an R^2 value for the training set using at -3.635 (1 Neuron) and -3.55 (1 Neuron) for the testing set. Using Mean Absolute Percentage Error to get accuracy, we got 73.8% using the Linear Model. However, we got an increased accuracy of 87.6% using the Non-Linear Model even though it had worse R^2 values.

Participants Contribution

Group members: Sidney La Fontaine, Esther Anemeje, Dylan Sapienza

We worked very collaboratively throughout this project, we all assisted each other throughout each step of the project. Sidney La Fontaine was largely responsible for finding, pre-processing, and post-processing the dataset. Esther Anemeje was largely responsible for all linear regression models and results. Dylan Sapienza was largely responsible for all neural network models and results. Although we want to emphasize that we all helped out with every part of the project.

References

Berkeley Earth, *Climate Change: Earth Surface Temperature Data*, 2015, Berkeley Earth, Version 2, https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv

National Snow and Ice Data Center, *Daily Sea Ice Extent Data*, 2018, National Snow and Ice Data Center, Version 3, https://www.kaggle.com/nsidcorg/daily-sea-ice-extent-data?ref=hackernoon.com

Walker S. J., R. F. Keeling, and S. C. Piper, *Daily CO_2 Data*, 2016, Scripps Institute of Oceanography, http://scrippsco2.ucsd.edu/data/atmospheric_co2/

Center for Research on the Epidemiology of Disasters, *Emergency Events Database*, 2021,
Center for Research on the Epidemiology of Disasters, https://www.emdat.be/