University of Dayton
Department of Aerospace and Mechanical Engineering
MEE 460 – Engineering Analysis

Development of a Carbon Emissions Model for the United States Considering Population Characteristics

Shane Kosir & Andrew Zarlinski

PROJECT 3 FINAL REPORT

1.0 INTRODUCTION/ BACKGROUND

1.1 Problem Statement

With carbon emissions being at an all-time high around the world, the green energy initiative has become the foreground of focus in many countries to help minimize damage to ecosystems and the climate. Much of this damage comes from the daily lives of citizens and the technology used within them. With the global population expected to reach 9.7 billion by 2050 [1], the importance of reducing our environmental footprint is critical. This project was initiated by the US Department of Energy to develop a model to identify which specific population characteristics correlate with carbon emissions in the United States. The goal is to use this model to determine which population characteristics provide the largest opportunities for carbon emission reductions.

1.2 Background

A neural network model will be created that can predict US carbon emissions as a function of population characteristics. Population characteristics will be dependent on available data but can include diet, age, per capita GDP, and other factors. The model will allow for carbon emission predictions for different scenarios, for instance, low-meat consumption and increased renewable energy. Because of the nature of neural networks, this model will be interpolative rather than extrapolative, meaning that predictions will be made within the range of data the model is trained on. It follows that the model will be useful for predicting shifts in carbon emissions based on current trends but cannot account for future innovations such as cheaper renewable energy or increases in sustainable agricultural practices. The model can also be expanded to predict ecological impacts in specific sectors such as forestland or fishing grounds. These could provide comprehensive information by determining which adaptations will be most beneficial for specific sectors.

2.0 TECHNICAL APPROACH

Figure 1 depicts a flowchart of the approach taken to generate a model for predicting carbon emissions in the United States. The bullets that follow provide further details about the various steps of model generation. The flowchart is a modified rendition of a flowchart found in the literature [2].

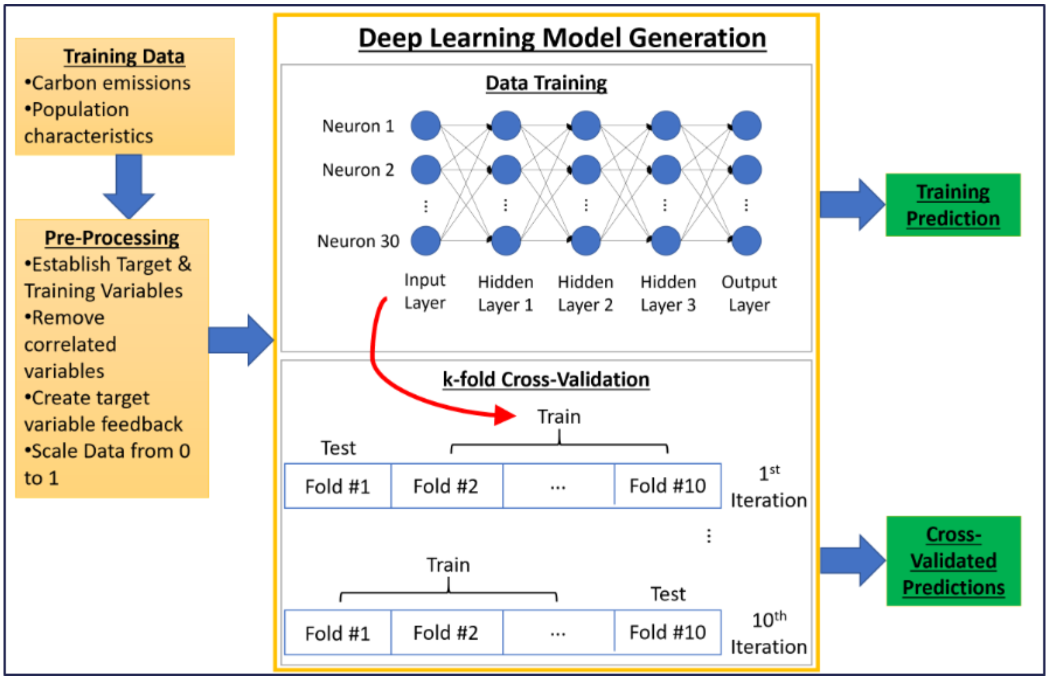


Figure 1. Model Generation Flowchart.

1. Model training data

Time series data, which is data that varies with time, will be used to train the model. The training variable will come from the Global Footprint Network, which maintains a dataset containing the United States' ecological deficits and reserves in specific sectors from 1961 to 2014 [3]. Sectors include carbon emissions, cropland, grazing land, forestland, and fishing grounds. The total carbon emissions from consumption, measured in hectares of forest required to sequester the carbon, will be used as the target variable for this study. Consumption is defined as production and imports minus exports, as can be seen in Figure 2. Features to train the model will come from the World Bank's "Health Nutrition and Population Statistics" dataset [4], which provides data on nutrition, health, and population dynamics in the United States from 1960 to 2015. The data will be downselected from 1961 to 2014 to match the time span of the Global Footprint Network dataset.

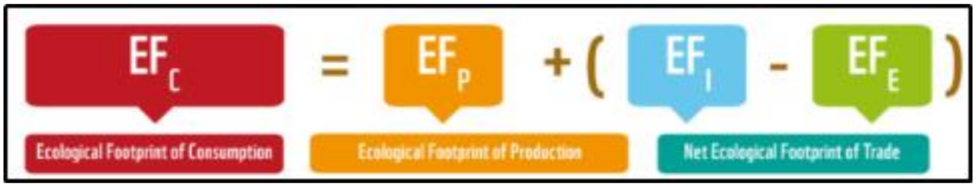

$EF_C = EF_P + (EF_I - EF_E)$
Ecological Footprint of Consumption
Ecological Footprint of Production
Net Ecological Footprint of Trade

Figure 2. Definition of the ecological footprint of consumption.

2. Data preprocessing

Pearson correlations are a method for visualizing linear correlations between features in a dataset. A Pearson correlation was used to remove correlated features from the World Bank training data, as correlated variables can make a neural network unstable and inaccurate [5]. A Pearson correlation with all of the initial features from the World bank dataset can be seen in Figure 3. Strong linear correlations between variables are indicated by larger circles. Blue circles represent positive correlations, whereas red circles indicate negative correlations. Before a Pearson correlation cutoff of -0.9 < x < 0.9 was used, 37 features were considered.

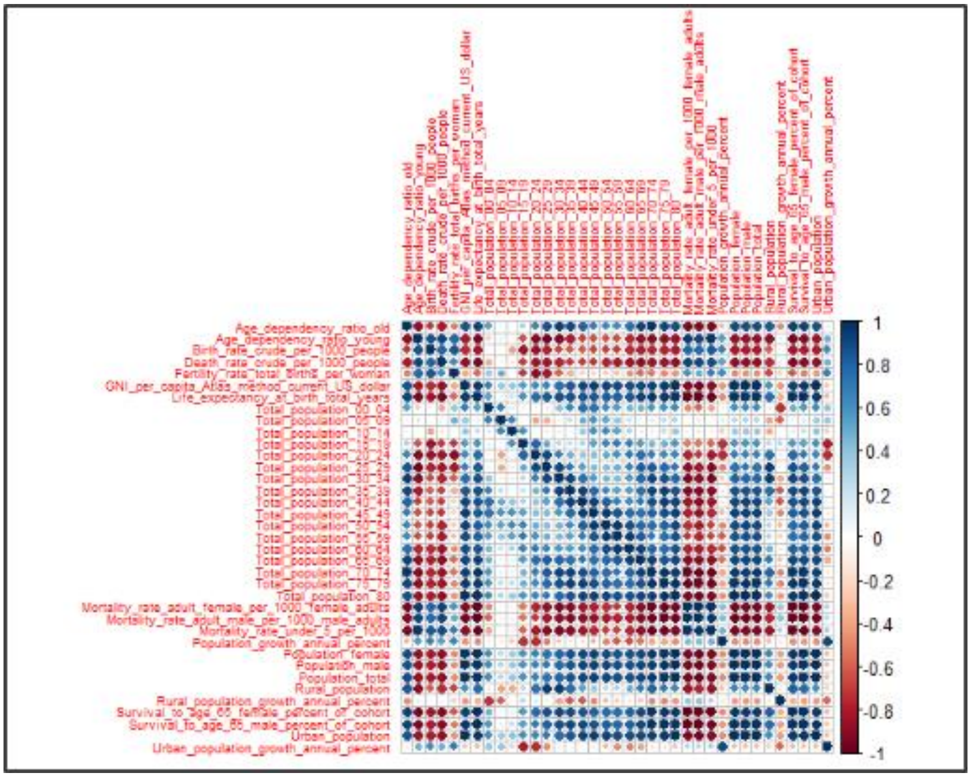


Figure 3. Pearson correlation without highly correlated variables removed.

Figure 4 depicts the Pearson correlation after correlated variables were removed. The number of features was reduced from 37 to 19. Not all variables with a correlation greater than 0.9 were removed. For instance, the population above 80 was kept despite its high correlation with GNI (gross national income) because there is no physical reason for them to be linearly correlated. Another example of this would be that the death rate and the age of men on the Titanic were linearly correlated. Yet, men of age 25 were not dying because of their age but because they were allowing women and children to board lifeboats first. It is also important to note that the rural population is still captured by the data because the total population is the sum of the rural and urban populations.

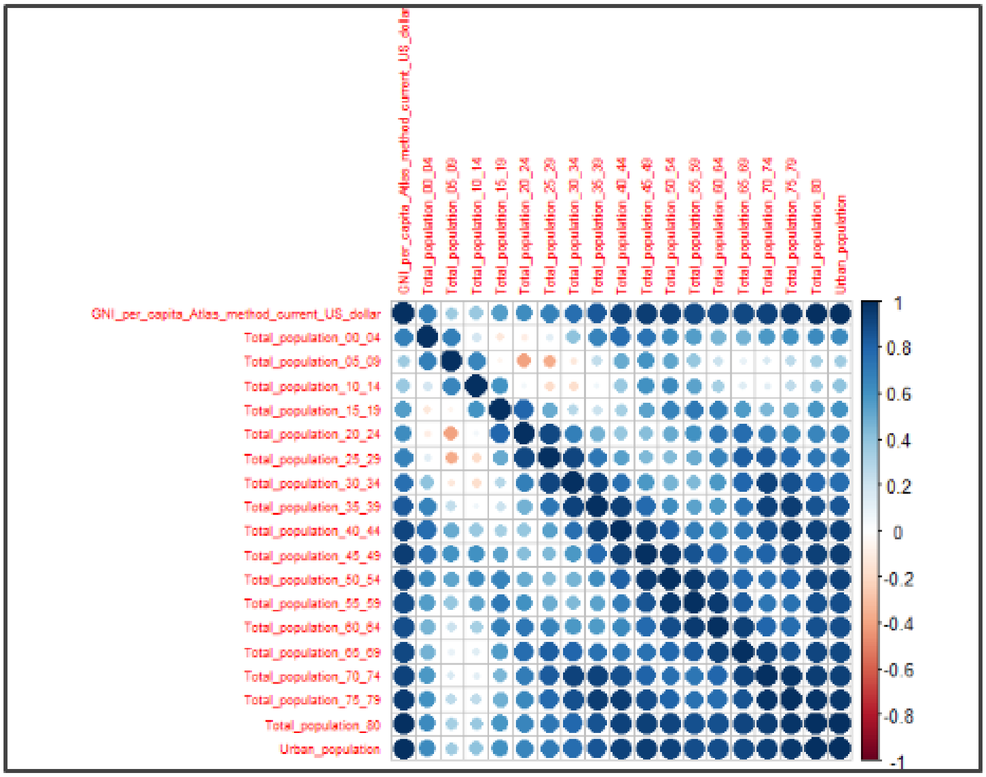


Figure 4. Pearson correlation with some highly correlated variables removed.

3. Spectral analysis and target variable feedback

Spectral analysis was performed on the carbon emission data to illustrate periodicities in the target variable. The result of the spectral analysis can be seen in Figure 5. The x-axis represents the frequency over which carbon emissions changed, while the y-axis represents the spectral density. Spectral density represents the strength of the variations at the frequencies designated along the x-axis. It follows that frequencies of down to 0.04 - periods of 27 years - are significant and should be considered when feedback is generated.

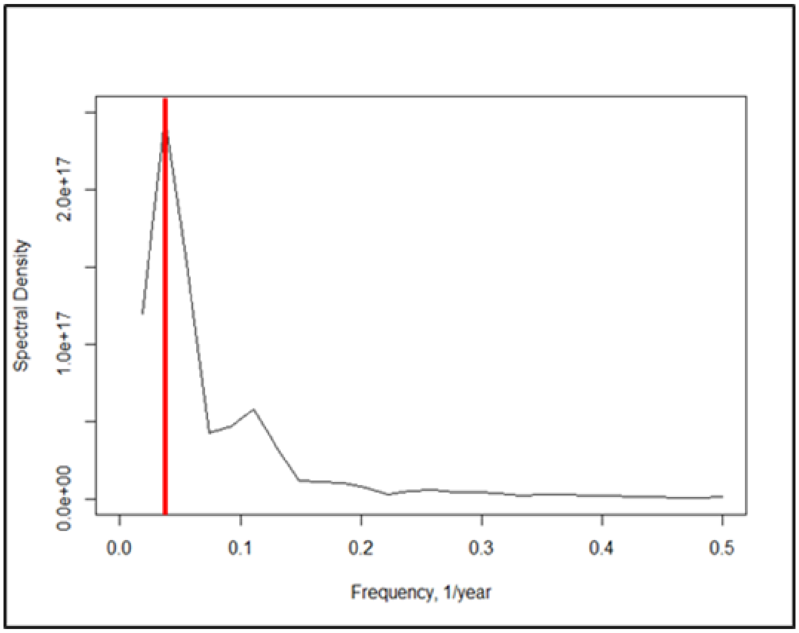
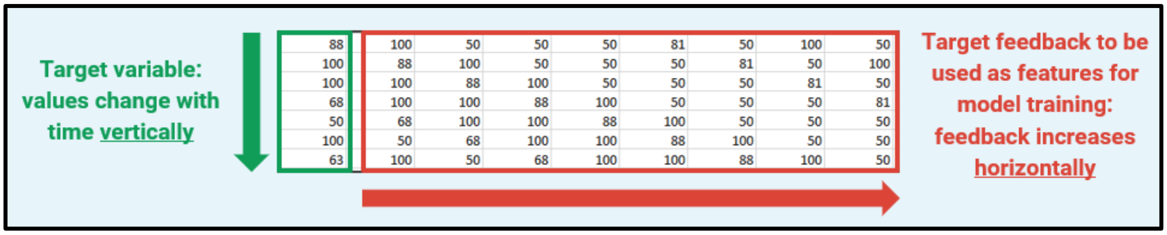


Figure 5. Spectral analysis for carbon emissions in the United States.

Target variable feedback allows neural networks to handle time series data. An example of target variable feedback can be seen in Figure 6. Note that these are not actual carbon emission values. The original target variable, which is represented in green, is offset by one year for every column in red. Each column in the red then serves as a feature for model generation.


Target variable: values change with time vertically
Target feedback to be used as features for model training: feedback increases horizontally

88	100	50	50	50	81	50	100	50
100	88	100	50	50	50	81	50	100
100	100	88	100	50	50	50	81	50
68	100	100	88	100	50	50	50	81
50	68	100	100	88	100	50	50	50
100	50	68	100	100	88	100	50	50
63	100	50	68	100	100	88	100	50

Figure 6. Example feedback for time series data. Green represents the original time target variable, while red represents feedback, with each column being used as a separate feature for model generation.

4. Neural network model generation

A multilayer feedforward neural network with backpropagation was generated using the deep learning platform H_2O [6]. Feedback was varied between nine to 27 years to determine which period produced the best R-squared and MAE (mean absolute error). Additionally, neural network parameters such as the number of hidden layers, neurons per hidden layer, and epochs were varied to produce optimal fits. 5 k-fold cross-validation was performed as an indicator of the model's generalizability to new data.

Design parameters and constraints:

- Prediction with <5% error
- The capability of handling shifts in population characteristics
- Generalizability to new data for predictions
- The capability of being expanded to new target variables (i.e. forestland)
- The capability of incorporating new features (i.e. meat consumption)

3.0 TESTING MATRIX

Testing scenarios for this work can be seen in Table 1.

Table 1: Model Scenarios

Parameter	Range	Justification
1961-2014 validation	Predictions made within Global Footprint Network and World Bank datasets.	These predictions were made to generate an actual vs. predicted plot and obtain R-squared and MAE metrics to determine the accuracy of the model.
2014 with decreasing urban population	The urban population decreased within the range of the original dataset while all other features remained at 2014 levels.	This scenario simulates migration from urban city centers to rural areas to determine if any effect on carbon emissions occurs.

4.0 RESULTS AND DISCUSSION

Figure 7 depicts actual carbon emissions plotted against predicted carbon emissions. Blue circles represent predictions made by the model exported from H_2O, while orange circles represent predictions from k-fold cross-validation. Three hidden layers with 100 neurons each and 5000 epochs were used for model generation. The k-fold predictions are accurate within 1.57% carbon emissions or 19,133,184 hectares of forest required to sequester the carbon. This indicates that the model will likely be generalizable to new data for predictions.

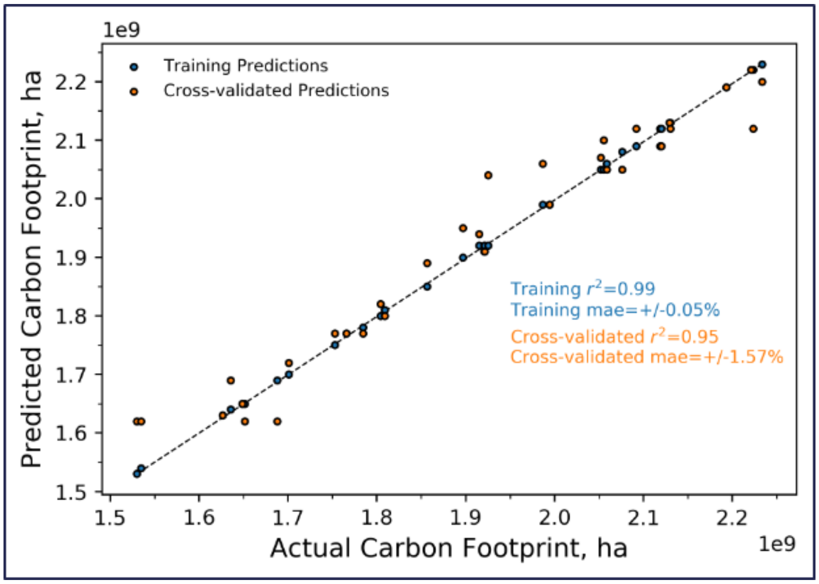


Figure 7. Actual vs. predicted carbon emissions plotted against one another. Blue represents predictions made by the model exported by H_2O, while orange represents predictions from k-fold cross-validation.

Figure 8 depicts the effect of urban population on carbon emissions. The model predicts a 1.2% decrease in carbon emissions with a 48.5% decrease in the urban population. It should be noted that it is assumed that the total population, GNI, and age distribution remain the same during the urban population decrease. It follows that the rural population increases for this analysis.

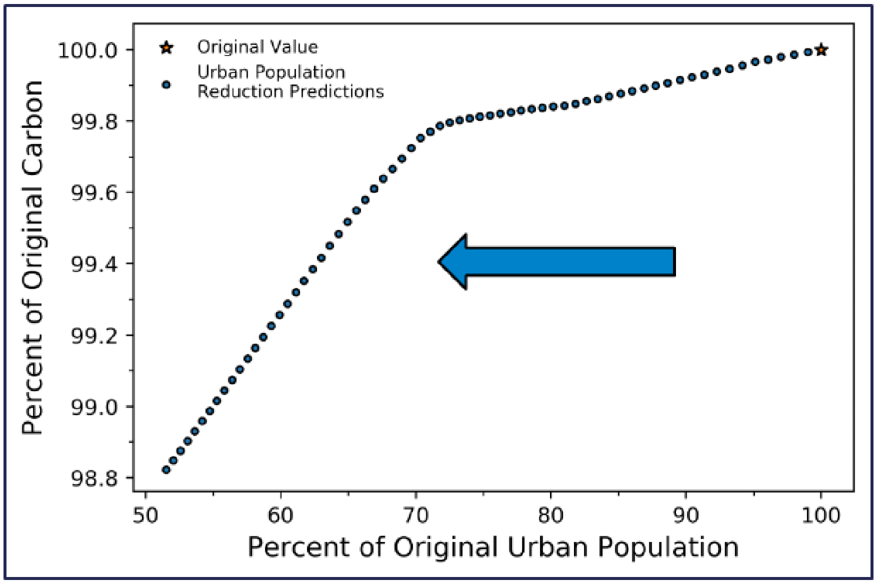


Figure 8. Carbon emission as a function of the urban population. The 2014 urban population is represented by the orange star.

The effect urban emigration has on carbon is relatively small at 1.2%. For reference, US forest biocapacity is 25.9% of current carbon emissions, putting the US in a significant carbon sequestration deficit. Figure 9 depicts random forest feature importances from the neural network model. It is likely that varying features with higher importance, such as population 25-29 or GNI, would have a larger effect than urban emigration.

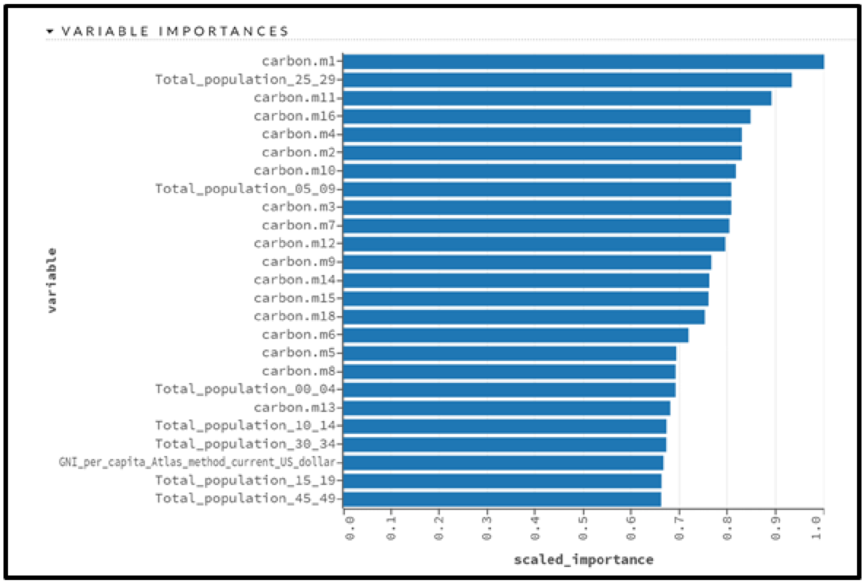


Figure 9. Random forest feature importance values from the neural network.

When building a neural network, it is important to check against overfitting, which is depicted in Figure 10. Overfitting occurs when a model fits the training data perfectly (small training error) but does not apply to new data (high validation error). k-fold cross-validation is used to identify overfitting in neural networks. In general, early stopping, dropout, or parameter regularization can be used to reduce overfitting.

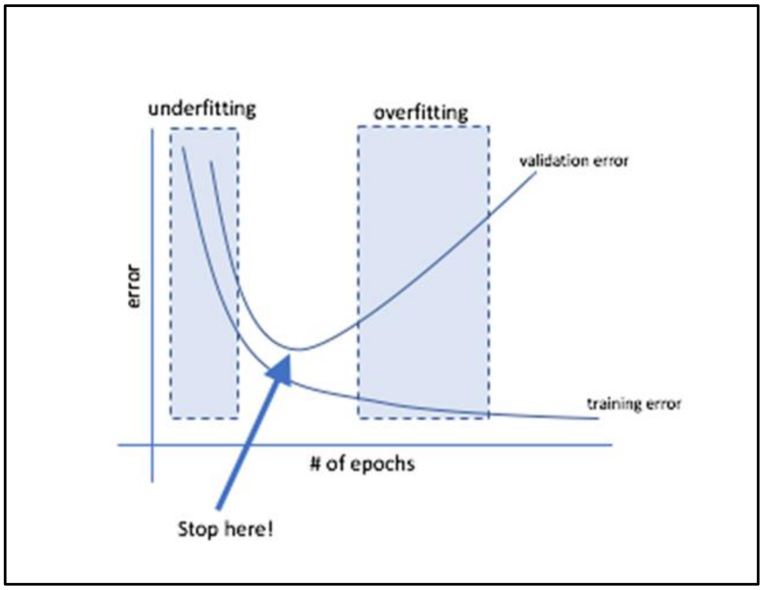


Figure 10. Depiction of model overfitting data.

Five k-fold cross-validation plots are shown in Figure 11. In general, overfitting was not experienced. However, the high variance in validation error and distance between training and validation error indicates that more features are likely required to capture the behavior of carbon emissions and reduce cross-validation error more fully. It should be noted that the deviance metrics are squared to increase training gradients.

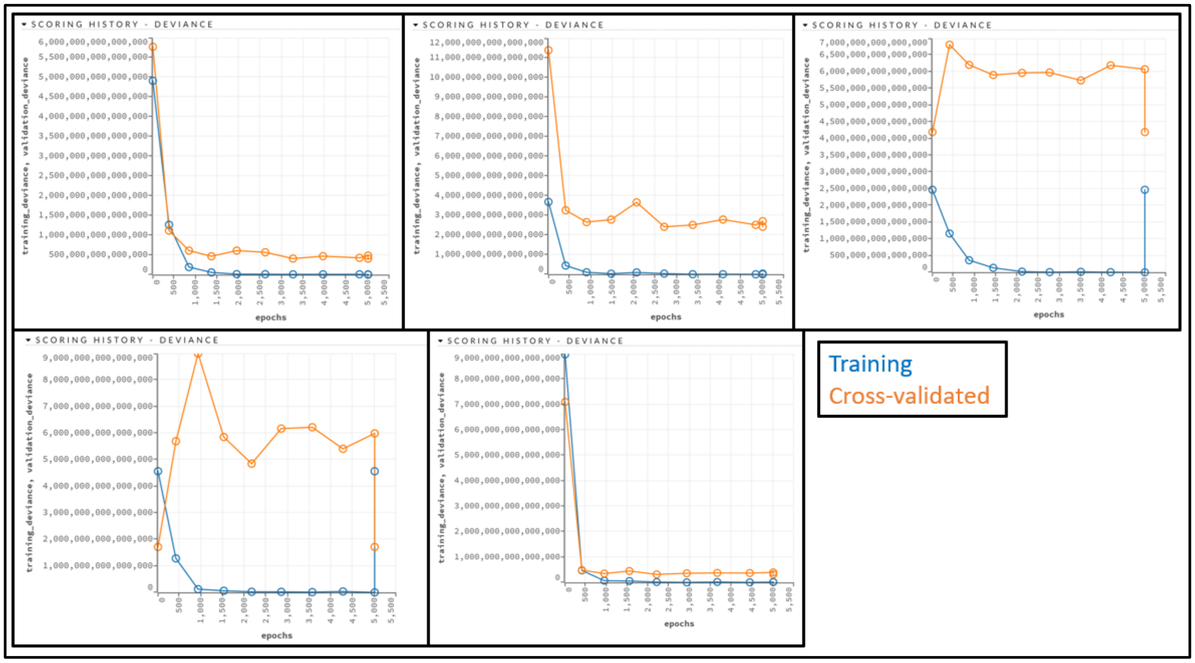


Figure 11. Model training and cross-validated data from k-fold validation.

Future directions for this project include the addition of other features such as energy consumption by sector or suburban population. Additionally, various scenarios can be explored, such as increasing renewables or decreased meat consumption.

5.0 REFERENCES

[1] Becker R. World Population Expected to Reach 9.7 Billion by 2050. Natl Geogr Mag 2015. https://news.nationalgeographic.com/2015/07/world-population-expected-to-reach-9-7-billion-by-2050/ (accessed September 4, 2019).

[2] Kosir S, Hill J, Roell T, Sheppard N. Cummins Chemical Analysis Semester 1. Dayton: 2018.

[3] Data and Methodology. Glob Footpr Netw 2014. https://www.footprintnetwork.org/resources/data/ (accessed September 4, 2019).

[4] Health Nutrition And Population Statistics. World Bank 2015. https://datacatalog.worldbank.org/dataset/health-nutrition-and-population-statistics (accessed September 4, 2019).

[5] Chatterjee S. Good Data and Machine Learning. Towar Data Sci 2017. https://towardsdatascience.com/data-correlation-can-make-or-break-your-machine-learning-project-82ee11039cc9.

[6] Using Flow - H2O’s Web UI. H2OAi 2019. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/flow.html.