Tristan Curry
ECE 9603 – Assignment 1
October 20th, 2023

1. Problem Description

Carbon Dioxide (CO2) emissions are generated whenever a fossil fuel is burnt, and in Canada the large portion of fossil fuel emissions come from the burning of fossil fuels to transport cargo, people and run personal vehicles. Canada faces a unique challenge in reducing its carbon emissions as the large land amount of land the country covers creates dependency on personal vehicles for travel. By measuring the CO2 emissions from vehicles, it is possible to track just how much vehicles are contributing to the overall production of CO2. Additionally, as the rate of combustion of fossil fuels is directly related to overall output, it can allow automakers to track how fuel efficient their vehicles are, and what parameters contribute to the overall production of CO2 emissions.

2. Data

The data set used was acquired from Kaggle, an open-source dataset sharing website. The data set consists of several column of data, where each column represents a different data feature. The data sheet contains data on different makes and models of vehicles collected throughout 2023 in Canada including the engine size in liters, the number of cylinders, the transmission type, fuel type, fuel consumption both in city, on the highway and averaged, and the CO2 emissions measured in g/km. As the focus for automakers is the reduction of fuel consumption and CO2 emissions. To predict the CO2 emissions for a given input data, I used several data features in the forecasting models, these include the Engine size, the number of cylinders and the combined fuel consumption. The measured output data is the CO2 emissions.

As the formatting of the data is spread across a row in multiple columns, often separated by data that is not being used for the modelling, the data was extracted from the csv and reformatted to only include the used parameters. After reformatting the data, the data was split into two different groups, one where the data would normalize using Sklearn's preprocessing methods. To properly utilize the preprocessing methods, some data columns were removed to avoid errors with the preprocessing method. The removed data included the car make and model, the transmission type, and fuel type, as these columns were not being utilized in the training, and that they contained string values. This normalized data was used to train the Neural Network as utilizing normalized data can help the neural network determine significance between the feature data, as the non-normalized data contains values that are on different orders of magnitude, thus the network may assign more importance to the larger value data, even if the actual significance is much lower. The original data set was used as the inputs for the State vector regression and multivariate regression methods.

3. Background

To predict the CO2 emissions from the input data 3 different forecasting methods were employed. A Neural Network in the form of a Multi-Layer Perceptron was used as the first method. The key feature of an MLP is its ability to model nonlinear relationships within data by incorporating one or more hidden layers between the input and output layers. The use of an MLP is ideal for this data set as the inputs will not have a linear relationship between the input and output. The MLP uses a weighted sum for each neuron in a layer and a bias term. The output of this weighted sum is then fed into an activation function, which is typically a non-linear function such as a sigmoid function, which determines if the neuron is active or not. The output layer for regression problems,

such as the CO2 Emissions problem, is a linear combination of the neuron activations in the final layer. Additionally, it is possible to measure the error of the regression using a loss function such as mean square error. Finally, backpropagation is used to update the weights using a gradient descent method.

In addition to the MLP, a Support Vector Regression (SVR) method can be used to enable forecasting. SVR's are an extension of support vector machines, which are typically used in classification problems. The governing idea behind SVR's is to find a hyperplane that best first the data while still maximizing the margin around the values. A large difference between SVR and traditional regression methods is that SVR seeks to minimize the deviation of predictions from a specified margin, and data points within this margin will not contribute to the loss function. SVR's also employ the use of a kernel function to transform data into a higher-dimensional space, allowing for the modelling of complex patterns and trends. Some common kernel functions are linear, polynomial and radial basis functions (RBF).

The final method used for forecasting the CO2 emissions data is a Multivariate regression (MVR). MVR extends the concepts of simple linear regression to model the relationship between multiple independent variables. The equation governing the MVR is seen in (1).

$$Y=\beta0+\beta1X1+\beta2X2+...+\beta nXn+\varepsilon \qquad (1)$$

Where Y is the predicted value, Xn represents the independent variables, β0 is the constant term, β1… βn are the regression coefficients, and epsilon is the error term. MVR is a powerful tool in data analytics as it can provide insight into the influence of how predictors influence the predictions, allowing for more informed decisions.

4. Methodology

For each methodology employed in this assignment, python was used to implement the forecasting models. In addition to python, several packages were utilized to implement the models efficiently, these include Scikit, Tensorflow-Keras, Numpy, and Pyplot.

The data is split into testing and training data sets, and for the neural network, the training data is split further into validation data that is utilized to check neural networks learning throughout each epoch of training.

For the MLP model, the data was split into testing and training sets with a test size split of 0.2. A normalization layer was then created to normalize the input training data independently. The model itself was created using the normalization layer, 3 deep learning dense layers, 2 with 2048 neurons, and one with 1024 neurons. Each of the 3 Dense layers had an activation function of Rectified Linear Activation Function (ReLU). The output layer has one dense layer with 256 neurons. To compile the model, several options were chosen, for loss the measure of Mean absolute error was chosen. While the chosen optimizer was Adam, with the metric being measured is accuracy. The model was fit on the training data with 100 epochs and a validation split of the data set at 0.2. Once the model was fit, the remaining test data was used to predict and compared to the true data.

The state vector regression method utilized the SVR method in Keras, with the kernel being an RBF, a C value of 1 and using the gamma value of scale. Specifying gamma to be scale, uses the value computed by taking inverse of the number of features multiplied by the variance in the training data set. This scale value utilizes the dataset to produce a coefficient for the kernel. The values used in the SVR are the default values used when creating a SVR in Keras. Once the model has been created, the input training data and

training outputs are fed into the model to train the SVR. After the SVR has been fit to the training data, the system is told to predict the outputs from the testing input data. To obtain a score for the performance of the SVR, the score method is used, which returns the R squared value for the regression. The Mean square error and mean absolute error also calculated using the predicted and true values.

Finally, for the MVR forecasting technique is implemented using Sklearn's linear model method. The model is built using all default values, these options include calculating the intercept, copying the initial training data, use parallel computation, and whether to force the coefficient to be positive. For this regression all values are set to true. Once the data has been fit to the model using the training data and outputs, the Mean Absolute error, Mean Square error value, and R2 score are calculated, along with the slope intercept value, and the number of features that were seen during the training.

5. Results

From the MLP models training and testing it was able to predict the CO2 emissions to a poor extent. The accuracy in training and validation reached final values of 0.21% and 0.25% respectively. The testing accuracy was only 0.41%. The loss for the training, validation and testing was 5.4754, 4.3882 and 4.5826 respectively. The accuracy can was graphed and is seen in Figure 1.

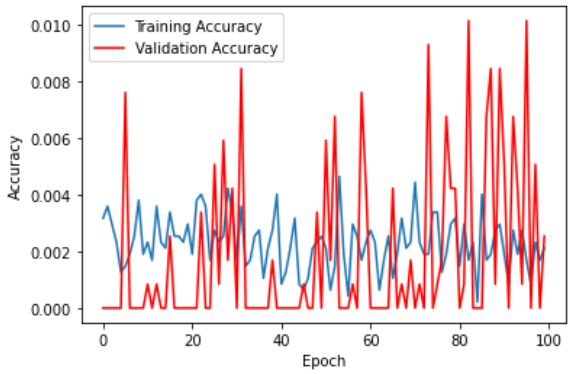

Figure 1: Training and Validation Accuracy from the MLP

The loss for the Training and Validation loss is also visible in Figure 2.

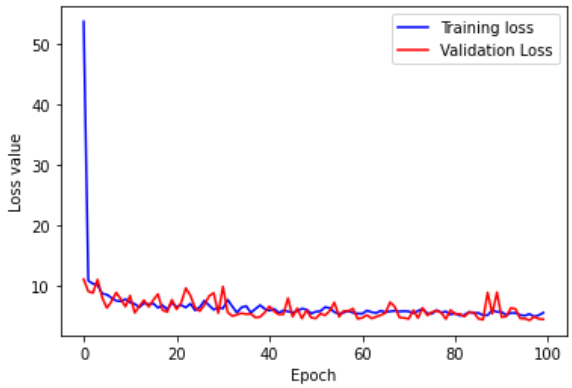

Figure 2: Loss from the Training and Validation of the MLP

These results are far below the other two models and are quite poor overall. There are several reasons that can be attributed to the poor performance, such as data that is not properly formatted for the desired regression training, the simplicity of the model, or the lack of proper parameters in the model. The data can be improved by using more robust data sets with additional parameters or the total amount of data can be increased. The models simplicity can be overcome by adding in different layers that are more proficient at regression learning such as convolutional layers, max pooling, or constructing a Recurrent Neural Network (RNN) instead of the Deep neural network that is implemented here. Finally, the parameters of the network can be changed to better support the learning progress of the network. For instance,

increasing the number of epochs, changing the optimizer, or changing the activation function for each layer to better suit the neural network and the data.

The SVR forecasting model was able to successfully fit the data to the model and achieve a coefficient of prediction (R2) value of 0.8997, with a mean sum of squares error of 336.63 and a mean absolute error of 7.5386. The R2 value shows that there is a high correlation between the data inputs and the output predictions. This shows that the model can accurately find a margin that the predictions deviate very minimally from. This outcome shows the range of data for the inputs is highly correlated and can act as accurate predictors to produce CO2 emissions in a vehicle. The Mean Square error and Mean Absolute error are both relatively low when comparing to the standard deviation between output data values to be 58.51 g/km.

The output from the MVR forecasting produced an R2 value of 0.865 with mean absolute error of 13.09 and mean sum of squares of 395.019. The output also included the coefficient value for each independent variable, these values are 5.499, 6.461, and 13.264 for the Engine size, Cylinders, and combined Fuel consumption respectively. Again, as the R2 value is approaching 1, it shows a high correlation between the features data and the output data, making the model highly accurate in predicting the test data's outputs.

Overall, the three algorithms can predict the outcome of CO2 emissions based on the input data of Cylinder number, engine size, and combined fuel consumption, however the methods are able to do this to much different degrees. The SVR and MVR methods are very good at fitting the data and determining the output from the test data. With high R2 scores and minimal error scores for major metrics, they can perform well and get reasonable results in predictions. The Neural network however has a hard time accurately predicting the training and testing data. This downfall can be rectified by recreating the neural network with an improved data set, better parameters, or different model structuring to ensure that the training data is able to improve the neural network's ability to accurately predict the test data.