Predicting Continuous CO and NOx Values Using Various Regression Models

2047088

April 2023

Abstract

A database of 9 descriptors and 2 emission measurements (CO and NOx) was explored with standard exploratory statistics and were used to train regression models in order to investigate the viability of predicting CO and NOx with this data. Exploratory statistics were used to discover the outliers in the data which were subsequently removed in order to clean the data. The cleaned data set was then used to build regression models for predicting CO and NOx values. Various models were trialed with the random forest regression model significantly out scoring the other models when using the metrics: R-squared score, mean average error (MAE) and mean squared error (MSE). This model achieved an R-squared score of 0.81 for CO and 0.88 for NOx, a MAE of 0.30 for CO and 1.95 for NOx and a MSE of 0.22 for CO and 8.57 for NOx.

1 Introduction

Indirect greenhouse gases contribute to the greenhouse effect by means such as ozone formation and production of direct greenhouse gases (1). Carbon monoxide and nitrogen oxides are two examples of common indirect greenhouse gases (2). Due to the recognisd negative effects these gases have on global warming; there are many examples of implementation of international agreements and domestic policies in an attempt to cut production of these harmful gases. Such polices include the Kyoto Protocol which commits industrialized countries to reduce and limit their greenhouse gas emissions as well as the National Emission Ceilings Regulations 2018 and the Long Range Transboundary Air Pollution (CLRTAP) (3)(4).

It is therefore in the energy production industry's interest to reduce the amount of NOx and CO produced per year in order to meet requirements set by government regulators. Hence it is valuable to analyse the emission data produced in their gas fired power plant environments to optimise for the reduction of these emissions in the combustion process as well as produce models to allow for forward planning and investment into gas fired plants.

For this reason an operator of a gas turbine power plant has asked for a consultation regarding the plausibility of making accurate predictive models of the CO and NOx exhausts released from their turbine. They have provided a database containing measurements from an in-operando turbine (Appendix 1) with no intention of collecting more data.

1.1 Exploratory Methods

Using exploratory statistics helps uncover issues within the data such as outliers and missing values. It can also be used to investigate the data's distributions to prevent biases forming from skewed features as well as being an invaluable tool for identify trends and relationships between the variables.

A vital first step to any data analysis process is to cleanup the raw data. Cleaning up involves the imputation of data due to missing values or the removal of data by identifying and removing outliers.

Identifying outliers is a difficult part of cleaning the data due to this process being ultimately subjective. Using graph-

ical inspections of the data's distribution is often used to identify outliers. Methods such as box plots, violin plots and histograms are able to give a good suggestion of which data points could be considered outliers (5). A histogram divides the range of values that is being measured into a number of bins. The number of data points which fall into each of these bins is summed. This information is then displayed like a bar chart where each bar represents a bin and the height of the bar represents the amount of data within that bin. These plots are useful for identifying the distribution of the data and therefore get insights into the modality or skewness. Histograms can often be helpful for discovering possible outliers by comparing the heights of the bars (5). Box plots are another type of distribution plot however these plots mainly focus on the statistical values of the data such as the median, the first and third quartiles and the minimum and maximum values. This plot contains a box which shows the interquartile range, a line inside the box over the median value and then there are two whiskers on either side of the box which extend to the maximum and minimum values, traditionally calculated as 1.5 times the interquartile range at each end of the box (5). Any values beyond the whiskers are considered to be outliers in the data. Box plots are good visualisations to see a distribution's skew and variance which would be indicated by the positions of the features discussed above. For example, if a median value is significantly closer to the lower interquartile range than the higher range than the data is skewed towards the lower values. When using box plots to investigate outliers it is important to consider the type of distribution of the data. Box plots work best when the data is uni-modal and roughly uniformly distributed over a median value. The third type of distribution plot discussed is a violin plot which make for a good middle ground between the histogram and the box plot. This is because it is able to show both the actual shape of the distribution through it's density curve; similar to how a histogram displays the data. As well as displaying the median, quartiles and outliers (5). It is best practise to use a range of distribution plots as this will facilitate the best understanding of the data.

Correlation is the statistical measure of the linear relationship between two variables. This makes it a useful tool in discovering patterns and insights in the data. Correlation values range from -1 to 1 with 1 representing a perfectly positive correlation and -1 representing a perfectly negative correlation. Correlation can be used to identify the main variables which affect a specific response variable allowing for an emphasis of the most important data in a data set. Correlation can also be used to eliminate statistically insignificant features from a model in order to reduce any over fitting. Due to the fact that correlation does not mean causation, it is therefore important to not rely too heavily on these values.

It can be difficult to identify patterns in highly dimensional data which is a problem when exploring the relationships within large data sets with many features. Principle Component analysis (PCA) is a commonly used technique on high dimensional data to reduce the dimensionality of the problem. It does this by creating a new data set of orthogonal variables called principle components (6). The principle components are created to preserve as much of the data's variability as possible and are ordered from the most variance explained to the least. PCA's main uses are in data reduction as it reduces the size of the data set while keeping the most important contributors in the data. As well as in data visualisation where visualising data with high dimensions becomes an increasingly difficulty challenge and therefore needs to be reduced to a dimension that can be plot in the standard lower dimensional graphs.

1.2 Regression Models

Regression models are a type of statistical prediction model which relate the independent variables in the data to the dependent variable(s). This relationship can then be used to make predictions of the value of the response variable for new input data. There are many different types of regression model, all of which are able to capture different types of relationships within the data. Each model has it's own strengths and weaknesses and there-

fore it is valuable to train and test different regression models on a single data set in order to end up with the best model.

1.2.1 Linear Regression

Linear regression is a regression model in the form shown in equation 1 where a range of input variables are accompanied by their own weightings which sum together in order to produce a predictive result (7).

$$y = b0 + b1x1 + b2x2 \quad (1)$$

The standard method of calculating the weights for each input term is using the least squares method which aims to reduce the discrepancies between the observed values of the dependent variable and the predicted values obtained from the independent variables (8). Simple linear regression is the relationship between the dependent variable and one independent variable, achieved through the use of the least squares method to calculate the best fitting straight line through the data.

$$y = b0 + b1x1 \quad (2)$$

Due to the limitation of one independent variable, the applications of simple linear regression falls short to many other models. Especially due to the fact that their is a similar, more powerful and just as simple-to-implement method such as Multiple linear regression. Multiple linear regression is a linear regression with more than one independent variable.

$$y = b0 + b1x1 + b2x2 + b3x3 + ... \quad (3)$$

For multiple linear regression the least squares method aims to produce the best fitting hyperplane through the data.

1.2.2 Regression Tree

A decision tree is a supervised learning method which can be used for both classification and regression. This method contains a hierarchical tree structure consisting of a root node connected to decision and leaf nodes through branches. The root node represents the whole data set. From here the data is split through the different branches depending on the conditions set out at each of the nodes. The branches either lead to decision nodes which are more conditional statements or to leaf nodes which do not have any further splitting (9). The structure of this method is shown in Appendix 2. A regression tree is simply the type of decision tree used for regression problems, where the goal is to predict a continuous numerical value rather than a class.

1.2.3 Random Forest Regression

Random Forest Regression is an extension to the regression tree and is an example of an ensemble method. This means that it creates multiple models and combines them in order to improve the overall model's results. This method combines a large number of regression trees together where each tree is trained on a different subset of the training data, subsequently reducing the over fitting of the model. When producing a prediction for new data points the random forest regression takes the average prediction from all the regression trees. This has the effect of reducing the noise and error caused by the outliers and therefore increasing the model's accuracy.

Random forest regression is beginning to be recognised for its high accuracy compared to other learning models. This high accuracy is due to it's "wisdom of the crowds" approach (10) (11).

1.2.4 Regression Neural Network

Neural networks are an unsupervised machine learning method which consist of layers of interconnected nodes loosely modelled on the neuron connections in our brains (12). Each connection between the neurons in the artificial network has an associated weight which is responsible for representing the significance of a certain connection (13). Data is passed from the input layer, through the hidden layers and to the output layer. Each layer applies a function on the data before sending it the the next nodes. There are many different functions and therefore different types of layers. In dense layers each node receives the data from all the nodes in the previous layer, this helps the model understand the complex non-linear relationships between the layers by adjusting the

many weights and biases during the training stage of the model. This is a very expensive yet important part of a neural network. It is expensive due to the number of connections, weights, and biases that it creates and computes. Another common type of layer is a convolutional layer which is used to extract data from strings of data such as images and are largely used for classification.

Regression neural networks are designed to output a continuous numerical value based one or more input features. These models are typically trained using back propagation which adjusts the weights and biases of each node in each layer depending on the score of statistical values such as the mean squared error or mean absolute error (14). This adjustment is done to optimise the accuracy of the predicted values to the output values. This allows the model to know when it is improving as these metric values would be reducing, with the aim to minimise these values to the greatest extent possible.

1.2.5 Evaluating Regression Models

Regression models can be analysed using a handful of statistical metrics such as mean squared error, mean absolute error and R-squared score to compare their predictive performance and differentiate between the success of different models.

The mean squared error (MSE) is a measure of how close a regression line is to the test data that it is being fitted for. It is calculated by summing the squared difference between the predicted values and the actual values. It therefore gives an indication as to how well the regression is representing the data. A low MSE shows that the the predicted data points are close to the actual data points and therefore when building a regression model the aim should be to reduce MSE in order to improve the accuracy of the model.

The mean absolute error (MAE) is similar to MSE in that it a measure of the distance from the actual values to the predicted, the difference between the two is that MAE is the absolute value between the two points and MSE is the squared value. MAE is another metric for measuring the accuracy of a model. MAE fails to punish the large errors in predictions compared to the MSE Metric. MAE is therefore a much kinder metric to outliers as it only measures the absolute distance of the error instead of squaring it like the MSE does. This means that for MAE the large prediction errors are less significant to the overall value compared to MSE.

The better the predicting power of the model, the lower these indicators will be when using it on new data. Therefore the goal of a model is to reduce these values as much as possible. Both of these metrics can be used to get an understanding of the errors of the prediction model.

R-squared score (R2 score) is a metric that measures the proportion of the variance for a response variable that is explained by the input variables (15). The R2 score is a number between 0 and 1 where 1 represents that the model perfectly fits the data and 0 represents that the model is unable to explain any of the variance in the response variable. The R2 score is therefore another metric for evaluating the accuracy of the model. However it should not be used as a sole metric for accuracy as R2 score values can be inflated due to over fitting and therefore does not always give an accurate representation of how well the model is at predicting on new data.

2 Analysis and Discussion

2.1 Exploration of the Data

The first step in the analysis was to investigate the shapes and distributions of the variables in the data set. This was done using the pandas library in Python. The function info() showed that there was no missing data in the data set. The describe() function was then used which gave a breakdown of the statistical metrics of the data. When looking at the statistical breakdown it was clear to see that there were some variables which had data points significantly far away from the mean. The most prominent example of this was for the CO variable which had a mean value of 2.03 and a standard deviation of 1.77 but with a maximum value of 34.8 which is approximately 19.4 stan-

dard deviations away. The other notable variables which had this same relationship were NOx and TA. The variables were then plotted as box plots (figure 1) in order to have the visual conformation that these data points were abnormal.

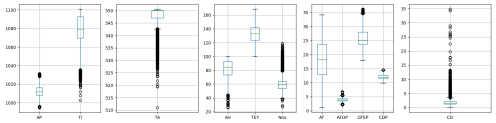


Figure 1: Box plots showing the data distributions of the variables in the data set.

The box plots in figure 1 show that there are a significant amount of outlying data points in the variables, especially in CO, NOx and TA. Violin plots (figure 2) were then used to get a more detailed idea of how the data was spread over the box plot.

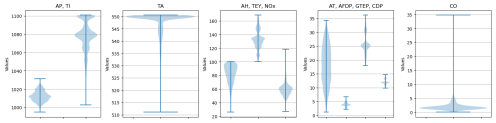


Figure 2: Violin plots showing the data distributions of the variables in the data set.

The violin plots in figure 2 more clearly show the types of distribution for the various variables which helps to identify which variables are significantly skewed by the possible outliers. The violin plot for CO is an example of where the density of the plot at the end of the violin opposite to the main bulk is negligible and therefore the values can be considered as outliers. From the violin plots (figure 2) the three variables exhibiting the worst forms of negligible densities were identified.

Histograms were then plotted (figure 3) for these three worst offending variables.

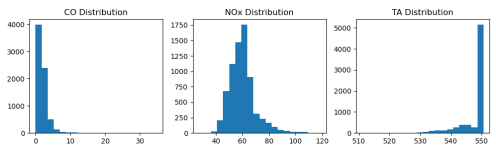


Figure 3: Histograms showing the data distributions of the variables CO, NOx and TA in the data set.

The histograms are shown in figure 3, in these plots the horizontal-axis spans across all the data points for the given variable and therefore the space where there appears to be an absence means that there are a negligible number of data points lying there. This also shows how far away these points actually lie to most of the data. This is especially visible in the CO histogram where the visible bins occupy approximately a fifth of the axis.

Plotting these graphs (figure 1,2,3) led to the conclusion that there were outliers in the data set and therefore the data should be cleaned before being used as test and training sets for any models.

The data was cleaned up by removing the data points which were three or more standard deviations away from the mean as this is a very standard classification of an "outlier". A function was produced to do this operation and it was applied to every variable in the data set.

The result can be seen when comparing the old and new CO histograms (figure 4).

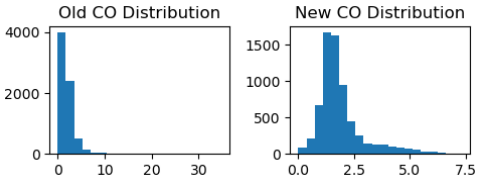


Figure 4: Histograms showing the effect of the data cleaning on the CO variable.

The next thing explored was the linear correlations between the variables. A correlation matrix was plotted (figure 5) to give an overview of all the correlations in the data set.

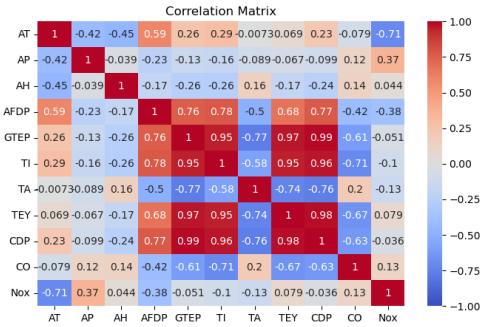


Figure 5: Plot of matrix correlations between all the variables in the data set.

The middle of this matrix plot (figure 5) shows that there is a high level of multilinearity between the variables. This is shown by the large portion of variables

having significantly high correlation coefficients with each other, represented by the bright red and blue colouring.

The most important correlation coefficients are between the input variables and the response variables. Therefore the correlation coefficients for CO and NOx were plotted (figure 6) so that the relationships between the variables and the response variables could be easily visualised and analysed.

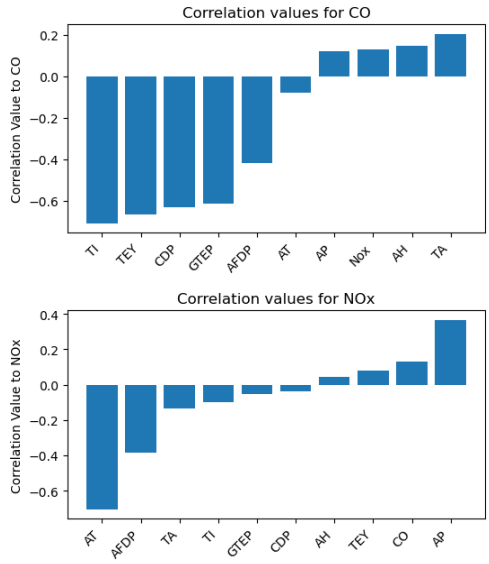


Figure 6: Plot of correlation coefficients for a, CO and b, NOx

From the CO plot (figure 6,a) is can be seen that there are are few relatively good correlations where the coefficients are approximately -0.7. However when looking back to the coefficient matrix it can be seen that these high correlation coefficients are all largely correlated with each other. This multicolinearity may effect some types of models such as a linear regression model as these values can not be considered independent from one other.

The NOx plot (figure 6,b) shows a far less correlated relationship between NOx and the variables with just one variable close to the -0.7 mark. This plot when compared to the CO plot shows that the NOx variable will likely be a more difficult value to accurately predict due to the weaker relationship with the input data.

The correlation coefficient between the two response variables is 0.13 meaning that the relationship between the CO and NOx is insignificant. This value highlights the need for two separate predictive models, one for each response variable. This also means that a model type that works the best for one of them may not work for the other and therefore a few models should be tested for each in order to create the best predictive models for the respective variables.

PCA was then used to pick up the relationships in the overall data which would not have been picked up by the correlation coefficients. This data set is highly dimensional with 9 input variables. The data set was first standardised using sklearn's StandardScaler() method to remove the mean and scale to unit variance. A PCA was then performed on the standardised data followed by a plot of the amount of variance explained by each of the principle components (figure 7). This was done to see how effective the PCA was at capturing the variance of the data.

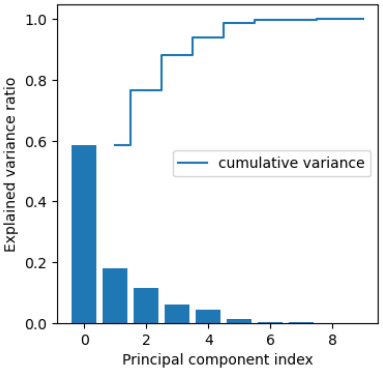


Figure 7: Bar chart and line graph showing the variance explained for each principle component.

This plot (figure 7) shows that the first principle component explains approximately 60% of the variance in the data which by itself is not a great representation. However when the second principle component is added these two add to make a reasonable data set which manages to explain 80% of the variance. These two principle components can be plotted against one another (figure 8) in order to explore the relationships or clustering effects in the data.

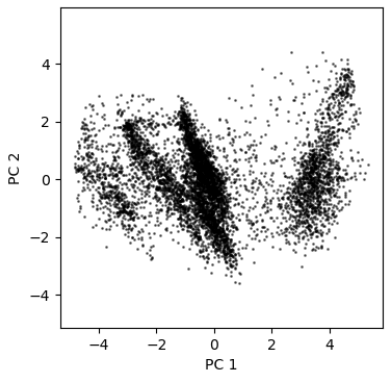


Figure 8: A scatter plot of the data Principle Component 2 vs Principle Component 1.

There is clearly a clustering effect to this data seen from the regions of high density relative to the overall scatter (figure 8). There appear to be four or five separate clusters when plotted in 2D with the first two principle components. The clusters also appear to have a linear relationship seen from the stretched straight line format that they are in.

Adding in the third principle component adds an extra 10% explained variance to the data set. This new data set can be visualised with a 3D plot (figure 9) however it is only improving on the previous plot by 12.5%.

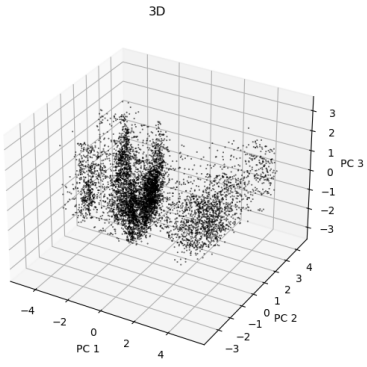


Figure 9: A 3D scatter plot of the data Principle Component 3 vs Principle Component 2 vs Principle Component 1.

As expected, the impact to the clustering is quite insignificant since the third principle component does not add much more to the explained variance. Given plotting in 3D makes the data less visually understandable, from now on the main exploration will be in 2D using the first two principle components and therefore will account for 80% of the variance in the data.

The main variables of interest are CO and NOx, the next step was to find out how these variables were related to the principle components. This was done by classify the CO and NOx data based on value bins and then applying that class data onto the plot as a colour map. Since there are roughly four clusters the data was split into four bins. Two plots were created, one showing the relationship for CO (figure 10) and the other for NOx (figure 11).

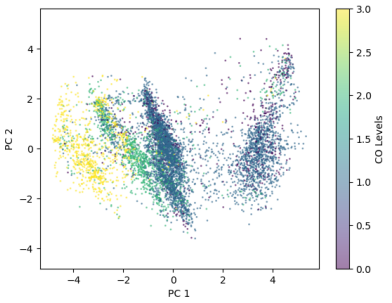


Figure 10: PCA of the first two principle components coloured with a map of the CO values.

From the CO plot (figure 10) it is apparent that the first principle component is the component responsible for the CO variable. This is seen by comparing the colours of the clusters along the PC1 axis. Here it is observed that the CO value changes from yellow to purple the further along the axis representing a decrease in the CO value when the PC1 value increases.

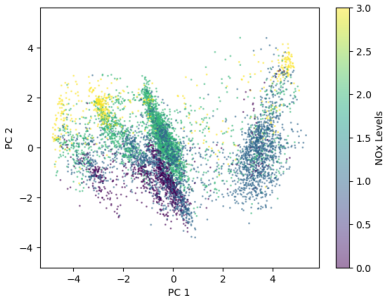


Figure 11: PCA of the first two principle components coloured with a map of the NOx values.

However for the NOx variable (figure 11), it can be seen that the determining principle component is actually the second. Seen by the colour change along the PC2 axis, where an increase in the PC2 value increases the NOx value.

Looking into the weights of each of the principle components it can be seen that

the main contributors to the first principle component match up with the highly correlated factors for CO and the main contributors for the second principle component match with the highly correlated factors of NOx.

Through this exploration the main contributing variables to each response variables have been realised and an overlying clustering within the data has been discovered. Through looking at the distributions and a few statistical metrics, the data was cleaned by considering outliers to be more than three standard deviations away from the mean. This reduced the data set by approximately 1000 entries, however this was not too big of a concern as the data set still contained almost 7000 data points.

2.2 Creating the Predictive Models for CO and NOx

Four types of regression models were trialed for both CO and NOx. The model types included a multiple linear regression, a regression tree, a regression neural network and a random forest regression. Three metrics were calculated for each model, the three metrics being R-squared score (R2), the mean squared error (MSE) and the mean absolute error (MAE) which enable the models to be compared to one another. Despite calculating and displaying the r2 score, this will not be the main metric used when discussing the accuracy of the results as the r2 score has many limitations which can prevent it from being an accurate description of the accuracy of the models for this data set. This is because the data that we are working with is multi-linear. The models are also being created for predictive purposes however the r2 score is not able to detect over fitting and therefore a highly r2 score may still mean a bad predictive model. For that reason the discussion will mostly focus on the MAE and MSE scores and use r2 will be to compare the different models to each other. The metrics MAE and MSE were then also compared to the standard deviation of the data in order to evaluate how accurate the predictions are. The standard deviation of the CO data is 1.063 and for NOx is 8.36.

The data needed to be set up into a format that could be put into the regression models. This was done by first splitting up the data into the feature data and the response data. The feature data labelled X included all of the variables except CO and NOx as these are the two response variables. The feature data X was then standardised using the StandardScaler() function. The y feature in the model was the only input data that was changed between the NOx and the CO models, i.e. y is the CO values for the CO models and NOx values for the NOx models. The X features and the y data were split into testing and training sets using sklearn's TestTrainSplit() function. The size of the testing set was chosen to be 20% of the overall data. This value was chosen as 20% is still able to be a significant subset of the data to test the data on, especially when considering the number of data points. However we want the training set to have as many training points as possible. A random seed was selected in order to keep the results reproducible which helps to accurately determine if a model is improving when changing it's parameters.

The simplest type of model trialed was the multiple linear regression. This was performed using sklearn's linear_model.LinearRegression() function. A multiple linear regression was considered as a model as it is simple but still able to take into account the high dimensionality of the data. The regression model for the CO response variable achieved a MSE of 0.409 and a MAE of 0.458 and an r2 score of 0.65. Since the two values are quite similar it shows that the model is not being significantly effected by large prediction errors between the prediction value and the actual value. It instead shows that the error comes from small deviations from the actual value. Since the CO data has a standard deviation of 1.06 that means that the average error of the predicted value is approximately half a standard deviation away.

The NOx model achieved a MSE of 27.2 and a MAE of 3.79, the large difference in values shows that the model is making very poor predictions for a small subset of the test set despite not being too bad with the overall predictions. Since

the standard deviation is 8.36 the average prediction is always within a half standard deviation from the actual value. This model achieved an r2 score of 0.62.

It is clear that the linear regression model performed worse for predicting the NOx values compared to the CO values. The poor performance for linear regression on the NOx data was predicted as this variable only had a small number of significant correlations in the input data.

The next model explored was the regression tree model. This was implemented using the sklearn's DecisionTreeRegression() function. For the CO predictions this model managed a MSE score of 0.45 and an MAE of 0.42. Once again the CO data seems to be less prone to large prediction errors and manages to make lots of similar sized small errors. The model gets an r2 score of 0.61 which is less than that of the linear model. The changes in the metrics are very small and therefore there is no significant difference between the results of these models. This model should only be considered over the linear regression when prioritising a lower MAE and therefore when assuming that the data is perfectly clean with no outliers.

For the NOx model the regression tree performs significantly better than the linear regression. The MSE score is 15.5 which is almost twice as good as the linear regression and a MAE error of 2.44 which is approximately 15% better. Since this model has a much higher MSE than MAE it means it is making a few significantly wrong predictions but quite good at predicting most of the data. The r2 score for this model is 0.78 which is quite a dramatic increase from the 0.62 before. This model is now on average predicting within a third of a standard deviation from the actual value.

The regression tree model was able to give a significant increase in predictive power over the linear regression for the NOx variable, however there was a negligible change for the CO variable. The increase in the NOx variable is likely due to how poor the linear regression was at predicting due to the lack of correlation to NOx within the input data. However for CO which did have a sufficient amount of correlation, the regression tree did not have quite as much of an impact. Another consideration into the underwhelming difference for the CO model could be due to the effects that the data distribution has on a decision tree model. This is because the regression tree model can suffer when the data is significantly imbalanced or skewed which happens to be somewhat true for the CO data which exhibits a beta distribution shape.

Due to the success of the regression tree the next model that was trialed was a random forest regressor which is a number of regression trees combined together in order to improve the accuracy of the predictions and reduce the over fitting of the model. The RandomForestRegressor() function was used with a random state set for repeatability for when the number of estimators was being changed. The number of estimators is essentially the number of regression trees in the model. This parameter was set to the value at which the accuracy metrics began to converge. The expectation was that this model would perform even better than the regression tree. Random forest was now able to significantly increase the accuracy scores for the CO variable. The MSE score was 0.22 which is approximately half what the other two models were getting. A MAE of 0.31 which is 25% better than the two previous as well as a large 0.81 r2 score. This model is significantly better than the previous two. And with a larger decrease in the MSE compared to the MAE it means that the model is predicting with less significant prediction errors.

The results for the NOx data set were just as impressive with a MSE score of 8.57 which is almost twice as good as the regression tree and more than three times as good as the linear regression. The same goes for the MAE with the smallest value so far of 1.95 and an r2 score of 0.88. This value of MAE means that the predicted values are within one quarter of a standard deviation from the actual values.

The random forest regression is able to significantly make the best predictions out of all the models so far. Due to the success of the regression tree on the data this was not much of a surprise as this model is essentially always stronger due to eliminating lots of the over fitting and outlier problems that effect its predictive

power.

The last model experimented with was a regression neural network. The construction of the neural networks was mostly down to trial and error for which layers would be added, changed or removed based on the results for a given trial. The only types of layers used were dense layers and drop out layers. From research conducted it seemed that a combination of these layers was fit for regression purposes, however the exact order and parameters would need to be tinkered with. The neural network was built using Tensorflow's keras package. For both the NOx and CO models the adam optimiser was used and the loss function was set to minimise MSE. The models also made use of the EarlyStopping function which acts as a call back to stop the model from over-fitting, this meant that the number of epochs chosen did not really matter unless the amount was too small. The call back function required a delta value, the delta value is a number which tells the loss minimiser how small the change in the loss needs to be to trigger the call back. There is also a patience parameter which allows for the delta value to be hit but for the model to keep training as long as there was a significant decrease in the loss within the patience width. The patience can be thought of the number of epochs for the delta value condition to be met so that the model stops learning. These two parameters were altered for the two models until the loss curve started to converge so that the model did not over fit. The number of epochs chosen for both was 100 however the call back was activated before reaching this point for both models.

The CO model structure is shown in figure 12 and displays the various layer types and shape of the neural network through the layers.

Layer (type)	Output Shape	Param #
dense_176 (Dense)	(None, 128)	1280
dropout_59 (Dropout)	(None, 128)	0
dense_177 (Dense)	(None, 128)	16512
dropout_60 (Dropout)	(None, 128)	0
dense_178 (Dense)	(None, 128)	16512
dense_179 (Dense)	(None, 64)	8256
dense_180 (Dense)	(None, 1)	65

Total params: 42,625
Trainable params: 42,625
Non-trainable params: 0

Figure 12: Structure of the Regression Neural Network for Predicting CO

The training and validation loss and MAE were calculated after each epoch during the training of the neural network and can be seen in figure 13 and 14.

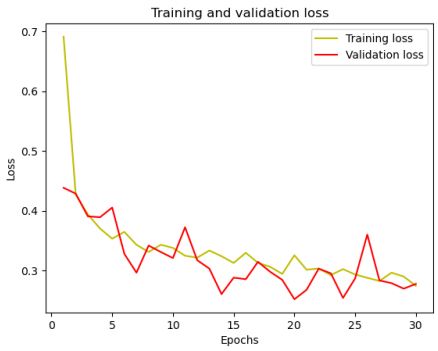


Figure 13: A Loss vs Epoch graph for the CO Neural Network

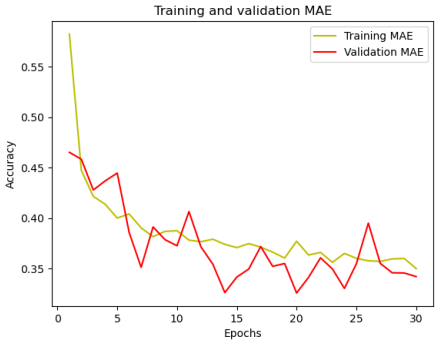


Figure 14: A MAE vs Epoch graph for the CO Neural Network

These plots (figure 13,14) were used to reduce the over fitting of the model by tweaking the call back conditions so that the training of the neural network would stop when the loss started to converge. This model resulted in a MSE of 0.28 and a MAE of 0.36 which is significantly better than the linear regression and regres-

sion tree however failed to beat the random forest regression model by 0.06 on both metrics. The slight lack in prediction power is shown by its r2 score of 0.76 compared to the random forest's of 0.81. Due to the infinite number of layer combinations that the neural network could have it cannot be said that this method does not perform as well as the random forest regression, it is just that the best neural network model that I was able to produce does not quite outperform it.

It was a similar result for the NOx Neural Network with the metrics competing with but not besting those of the regression forest. The network structure of the NOx model is shown in figure 15.

Layer (type)	Output Shape	Param #
dense_214 (Dense)	(None, 128)	1280
dropout_73 (Dropout)	(None, 128)	0
dense_215 (Dense)	(None, 128)	16512
dropout_74 (Dropout)	(None, 128)	0
dense_216 (Dense)	(None, 128)	16512
dense_217 (Dense)	(None, 128)	16512
dense_218 (Dense)	(None, 64)	8256
dense_219 (Dense)	(None, 1)	65

Total params: 59,137
Trainable params: 59,137
Non-trainable params: 0

Figure 15: Structure of the Regression Neural Network for Predicting NOx

The structure is almost identical however it has one more dense 128 layer than before. Once again the call back parameters were adjusted to minimise the over fitting of the model by attempting to cut the learning short once the loss started to converge. The loss and MAE vs epoch plots are shown in figure 16 and 17 respectively.

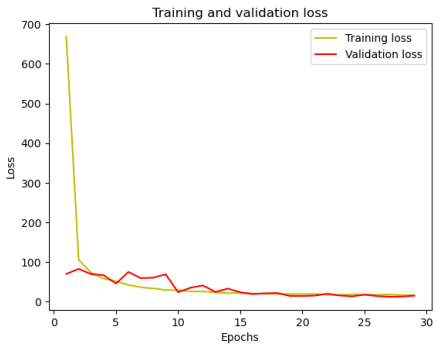


Figure 16: A Loss vs Epoch graph for the NOx Neural Network

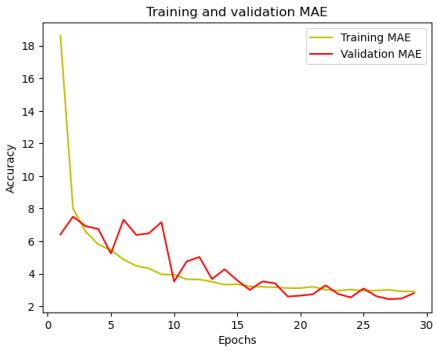


Figure 17: A MAE vs Epoch graph for the NOx Neural Network

As can be seen in the loss plot (figure 16), the parameters to stop the over fitting were not as successful with the model training significantly past the approximate convergence of the loss line. This effect is no so bad for the MAE plot (figure 17) but compared to the CO model these plots are relatively over fitted. For this reason the regression neural network only performed better than the linear regression for the NOx, rather than performing better than both the linear regression and the regression tree like the CO network did. This model was able to achieve a MSE of 16.2 and an MAE of 2.93 which are far off of what the regression forest was able to do. However from the loss vs epoch graphs there is a route to explore where this model went wrong. By fixing the over fitting of this model it may be able to perform on a similar level to that of the regression forest. However for the model actually produced it was not able to achieve a greater score than even the regression tree.

The metrics for all the models have been put into a table in order to conveniently see the differences and make a decision on which models should be used for predicting each variable. These tables are shown in figure 18 and 19.

CO Metrics			
Model	MSE	MAE	R2
NN	0.279503	0.362353	0.761687
LR	0.409288	0.457536	0.651028
RT	0.412073	0.404989	0.648653
RF	0.217908	0.301386	0.814204

Figure 18: Table of Metrics for CO Models

Nox Metrics			
Model	MSE	MAE	R2
NN	16.16053	2.936454	0.772886
LR	27.17469	3.790968	0.618097
RT	15.6408	2.449467	0.78019
RF	8.57051	1.94558	0.879553

Figure 19: Table of Metrics for NOx Models

The tables (figures 18, 19) clearly show that the random forest regression was the superior model for both the CO and NOx predictions. The worst results slightly varied between the response variables with the regression tree performing the worst in two categories for CO with the worst MAE value coming from the linear regression model. Whereas the linear regression was consistently the worst across the metrics for the NOx variable. This was foreshadowed by the fact the the NOx variable had little correlation with most of the data and therefore the linear regression model was not able to form accurate predictions. In comparison to the CO variable which had a significant amount of correlation to the data and therefore the linear regression model was able to compete with the regression tree for the r2 score and the MSE.

3 Conclusion

In conclusion the random forest regression model was the best model for predicting both the CO and the NOx values. This model was able to achieve an R-squared score of 0.814 and 0.879 for the response variables respectively with MAE values less than a third of and a quarter of their respective standard deviations respectively. The worst model for the CO variable was the regression tree, whereas the worst for the NOx variable was the linear regression which is likely down to this model not being suitable due to the lack of linear correlations to NOx in the data.

References

[1] T. Statistics Finland, "Indirect greenhouse gases ." `https://www.stat.fi/meta/kas/epasuorat_kasvi_en.html`, Unknown. [Online; accessed 9-April-2023].

[2] N. A. E. Inventory, "Overview of greenhouse gases ." `https://naei.beis.gov.uk/overview/ghg-overview`, 2022. [Online; accessed 9-April-2023].

[3] U. N. C. Change, "What is the Kyoto Protocol? ." `https://unfccc.int/kyoto_protocol`, unknown. [Online; accessed 9-April-2023].

[4] N. D. for Environment Food and R. Affairs, "Emissions of air pollutants in the UK – Nitrogen oxides (NOx)." `https://www.gov.uk/government/statistics/emissions-of-air-pollutants/emissions-of-air-pollutants/-in-the-uk-nitrogen-oxides-nox`, 2023. [Online; accessed 9-April-2023].

[5] M. Y. Charito, "A Complete Guide to Box Plots ." `https://chartio.com/learn/charts/box-plot-complete-guide/`, unknown. [Online; accessed 9-April-2023].

[6] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[7] X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012.

[8] B.-N. Jiang, "On the least-squares method," *Computer Methods in Applied Mechanics and Engineering*, vol. 152, no. 1, pp. 239–257, 1998. Containing papers presented at the Symposium on Advances in Computational Mechanics.

[9] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004.

[10] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," in *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*, p. 270, Springer, 2012.

[11] Keboola, "The Ultimate Guide to Random Forest Regression ." https://www.keboola.com/blog/random-forest-regression, 2020. [Online; accessed 10-April-2023].

[12] D. J. Livingstone, D. T. Manallack, and I. V. Tetko, "Data modelling with neural networks: Advantages and limitations," *Journal of computer-aided molecular design*, p. 1, 1997.

[13] S. R, "A Walk-through of Regression Analysis Using Artificial Neural Networks in Tensorflow ." https://www.analyticsvidhya.com/blog/2021/08/a-walk-through-of-regression/analysis-using-artificial/neural-networks-in-tensorflow/, 2021. [Online; accessed 11-April-2023].

[14] G. Seif, "Understanding the 3 most common loss functions for Machine Learning Regression." https://towardsdatascience.com/understanding-the-3/most-common-loss-functions-for/machine-learning-regression/23e0ef3e14d3, 2019. [Online; accessed 11-April-2023].

[15] J. Fernando, "R-Squared: Definition, Calculation Formula, Uses, and Limitations." https://www.investopedia.com/terms/r/r-squared.asp, 2023. [Online; accessed 11-April-2023].

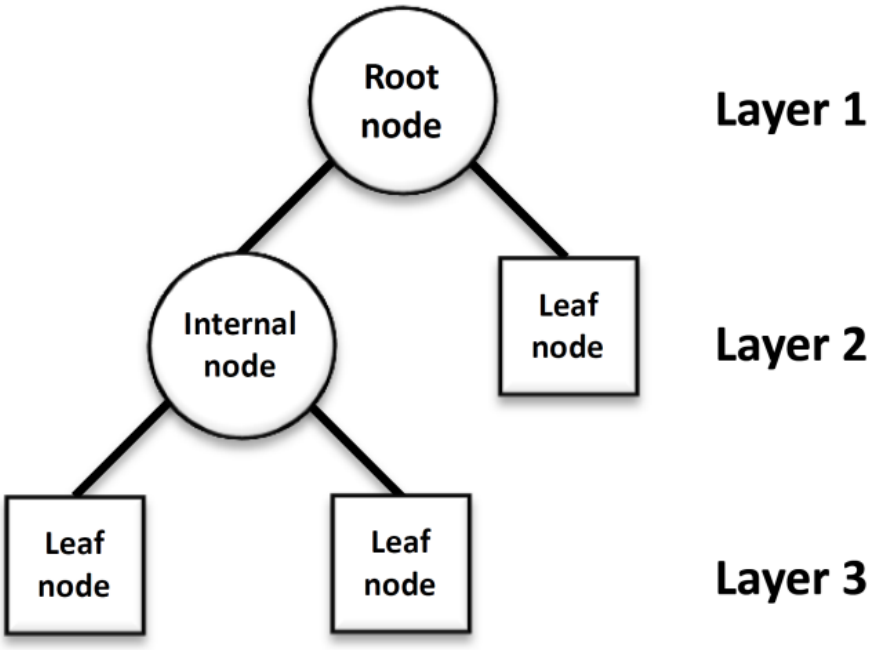
4 Appendix

Appendix 1:

Variable	Abbreviation	Unit	Comment
Ambient temperature	AT	°C	NOx emission can be reduced by working at higher temperatures
Ambient pressure	AP	mbar	
Ambient humidity	AH	%	
Air filter difference pressure	AFDP	mbar	
Gas turbine exhaust pressure	GTEP	mbar	
Turbine inlet temperature	TI	°C	More CO produced when combustion is incomplete (lower T)
Turbine after temperature	TA	°C	
Compressor discharge pressure	CDP	mbar	
Turbine energy yield	TEY	MWH	
Carbon monoxide	CO	mg/m^3	response to predict
Nitrogen oxides	NOx	mg/m^3	response to predict

Meta data for the database provided by the in-operando gas fired plant measurements.

Appendix 2:



Visualisation of the structure of the Decision Tree Method.