Graduate Data Scientist Case Study

Alyssa Yao
Faculty of Business and Economics, The University of Melbourne

May 10th 2024

In this study, I built machine learning and deep learning models to predict different indices of air pollution. I leveraged Multiple Linear Regression, Support Vector Machine, Random Forest, and Multi-layer Perceptron Models, ensuring best model performance. My analysis revealed an optimal model in predicting air pollution indices. Despite achieving commendable results, there remains room for improvement, particularly in computational complexity and KNN imputation. My findings underscore the potential of predicting one specific index using other information, the possible future application is to effectively reduce the use of expensive sensors.

1 Pre-Process

After importing the data, we first split the time into Year, Month, Day, and Hour. We then remove the last two columns and the last 114 rows because they are empty.

2 Exploratory Data Analysis

This part contains EDA of 5 selected columns, steps are divided into: cleaning data, revealing data features, visualising data, and analysing their correlations.

2.1 Choices of Variables

After pre-process, CO(GT), NMHC(GT), C6H6(GT), NOx(GT) and NO2(GT) still have missing values, thus, we choose these 5 columns to conduct EDA.

2.2 Data Cleaning

Since we have already removed some missing values, in this step we mainly focus on imputing missing values and deleting outliers.
The KNN imputation method is used but not simple mean filling as it can improve the accuracy of filled values and the quality of the data. 7072 missing values are imputed.
In terms of outliers, 367 outliers are deleted under the deleting criteria z-score$>$3.

2.3 Summary Statistics

After taking a look at the mean, standard deviation, min, max and median of these data, we have a preliminary understanding of them.

Variable	Mean	STD	MIN	50%	MAX
CO(GT)	2.0057	1.2068	0.1	1.76	6.4
NMHC(GT)	231.4038	173.5125	11	175.8	809.4
C6H6(GT)	9.7278	6.3437	0.3	8.4	32.6
NOx(GT)	222.8695	171.9457	2	172	858
NO2(GT)	109.5823	44.1511	2	107	254

Table 1: Summary Statistics

As CO(GT) and C6H6(GT) have lower values than other 3, in the following analysis, we plot them separately for better understanding. We can also notice they have different scales, thus, data standardisation is necessary in future steps.

2.4 Data Visualisations

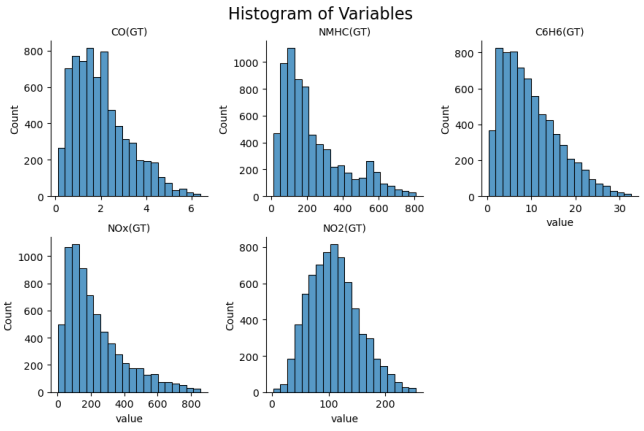

Figure 1: Histogram

From Figure1, the distributions of NO2(GT) is approximately normal, while other 4 are obviously right-skewed.

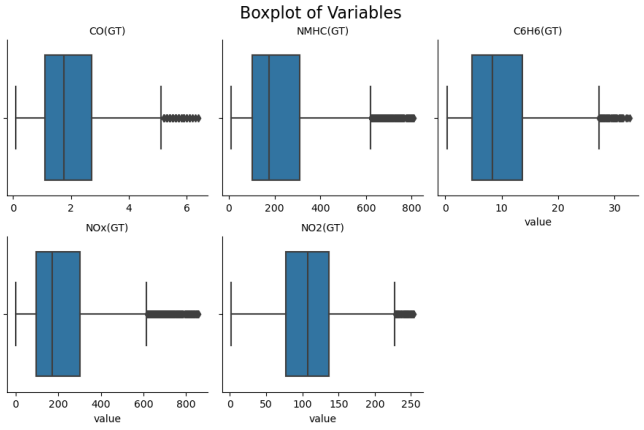

Figure 2: Boxplot

Figure2 gives the same result as Figure1 - with the exception of NO2(GT), all variables have left-leaning medians and boxes.

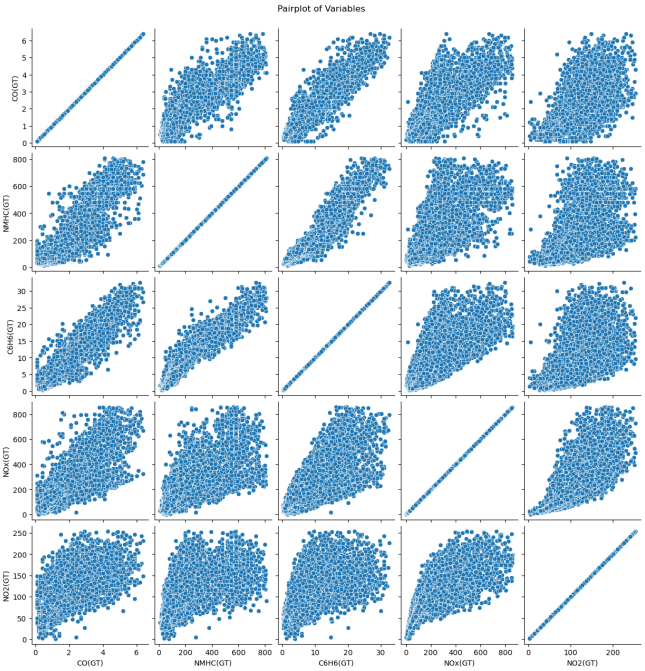

Figure 3: Scatter Plot

Figure3 can show positive relationships between each variable - as they tend to increase together. However, the points are not very dense, indicating that this is not a very strong relationship.

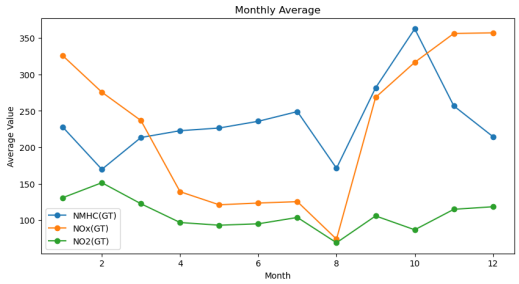

Figure 4: Monthly Plot

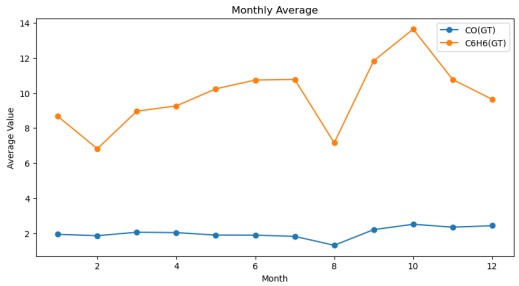

Figure 5: Monthly Plot

From Figure4 & 5, all pollution decreased in August, but increased significantly in October (except NO2).

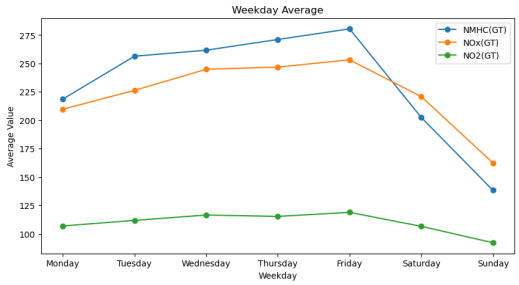

Figure 6: Weekday Plot

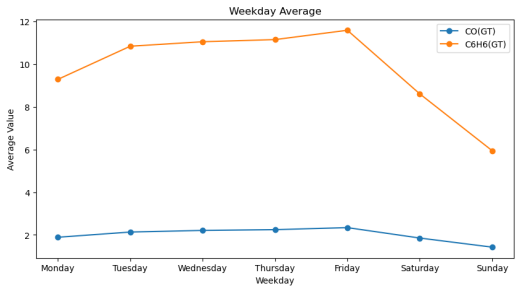

Figure 7: Weekday Plot

From Figure6 & 7, we can see pollution is significantly lower on weekends than on weekdays.

2.5 Correlation Analysis

As we find positive relationships in step3, we conduct further analysis in correlation and multicollinearity.

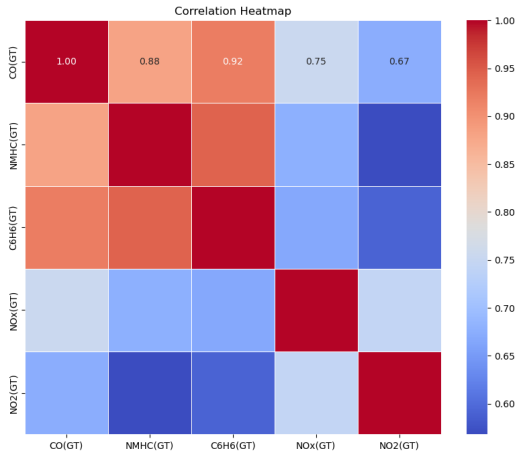

Figure 8: Correlation Heatmap

From Figure8, CO(GT), NMHC(GT), and C6H6(GT) have high correlations with each other.

Variable	VIF
CO(GT)	32.69
NMHC(GT)	26.33
C6H6(GT)	46.00
NOx(GT)	7.69
NO2(GT)	8.45

Table 2: Variance Inflation Factor

This can also be observed in VIF - these 3 variables have VIF >10.

3 Split Data

As we need to predict all 13 columns, we generate 13 dataframes (each with 1 response variable and 13 features), they are stored in a list.
Then we split the data into train and test datasets, with proportion 8:2.
In this step, we only use Month from all time variables, as per discussed, months have some effect on pollution, while years, days and hours have less.
Since the scale of the variables is different, we will standardise them first.

4 Multiple Linear Regression (MLR)

4.1 Model Assumptions and Characteristics

Model Assumptions
Linearity: The relationship between the variables is linear.
Errors: The errors should be independent and normally distributed.
Homoscedasticity: The variance of the errors should be constant.
Multicollinearity: The independent variables should not be highly correlated with each other.
Outperform
When the Relationship is Linear
Advantage
Interpretability: MLR provides coefficients for each predictor, thus, it is easier to interpret.
Disadvantage
Sensitivity to Assumptions: MLR's performance is sensitive to the violation of model assumptions.

4.2 Model Fitting

We use LinearRegression from sklearn to fit models. As our goal is to predict all columns, we use a for loop to fit 13 linear models (as per described in data spliting process), and then store our models in a list.

4.3 Tuning Parameters

The Backward Step-wise Selection is used to select predictors. However, as predictors are highly related with each other, MSE will decrease when we delete any predictor, thus, we keep all of them, setting n_features_to_select to be 13.

5 Support Vector Machine (SVM)

5.1 Model Assumptions and Characteristics

Model Assumptions
SVM does not make strong assumptions about the distribution of data. However, SVM performance can be affected by the choice of kernel and hyperparameters.
Outperform
Binary Classification: SVM is particularly effective for binary classification tasks.
Advantage
Flexibility: SVM allows for both linear and non-linear decision boundary within data through the use of different kernels.
Disadvantage
Computational Complexity: It does not perform well when the data set is too large otherwise it would be very computational expensive.

5.2 Model Fitting

We use SVR from sklearn.svm to fit our models. Same as MLR, we store 13 models in a list.

5.3 Tuning Parameters

The choice of kernel will significantly affect SVM model's performance, thus, we fit different SVM models with linear, polynomial, RBF, and sigmoid kernel.

We will keep them until the model selection step - to see which one performs better.

6 Random Forest (RF)

6.1 Model Assumptions and Characteristics

Model Assumptions
Random Forest does not make strong assumptions about the distribution of data.
Outperform
Random Forest performs well on large datasets with many features and instances.
Advantage
Accuracy: Random Forest tends to achieve high accuracy in classification and regression tasks.
Disadvantage
Interpretability: Random Forest models can be challenging to interpret due to the ensemble nature of the model.

6.2 Model Fitting

We use RandomForestRegressor from sklearn.ensemble to fit our models. Same as before, we store 13 models in a list.

6.3 Tuning Parameters

In random forest, we simply set number of tress to be 500 as it might not affect too much on our model, and we try two maximum features, 0.2 and square root.

7 Multi-Layer Perceptron (MLP)

7.1 Model Assumptions and Characteristics

Model Assumption
MLP does not make strong assumptions about the distribution of data.
Outperform
Large Datasets: MLP can handle large datasets with many features, given sufficient computational resources.
Advantage
Feature Learning: MLP can automatically learn useful representations from raw data, reducing the need for manual feature engineering.
Disadvantage
Overfitting: MLP models are prone to overfitting, especially when the model architecture is too complex or the dataset is small.

7.2 Model Fitting

We use MLPRegressor from sklearn.neural_network to fit our models. Same as before, we store 13 models in a list.

7.3 Tuning Parameters

In this step, we tune parameters when fitting models. We mainly focus on hidden layer size and alpha. We define 9 parameter combinations, that are,

Combination	Layer	Alpha
1	(50,)	0.0001
2	(50,)	0.001
3	(50,)	0.01
4	(100,)	0.0001
5	(100,)	0.001
6	(100,)	0.01
7	(50,50)	0.0001
8	(50,50)	0.001
9	(50,50)	0.01

Table 3: Parameter Combination

We fit them separately to see whether the optimisation converges when the maximum iterations is reached.Due to limited computing power, we set maximum iterations to be 1000.
After fitting all 9 models, parameter combination 7, 8, and 9 have more times of convergence than other 6 combinations. Thus we can conclude hidden layer size (50,50) is more suitable for our dataset and model, which means there are two hidden layers in the neural network, with the first hidden layer containing 50 neurons and the second hidden layer also containing 50 neurons. We keep all 3 models with (50,50) until model selection.

8 Summary

In this section, we summarise out finding (best model) with selection criteria. Limitations and future improvements will be disscussed as well.

8.1 Model Selection

Selection Criteria
Root Mean Squared Error(RMSE): RMSE provides a clear indication of how well the model fits the data. Lower RMSE values indicate better model performance in terms of prediction accuracy.
Adjusted R2: R2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Adjusted R2 provides a balance between model fit and complexity as it adjusts for

the number of predictors in the model and penalises the addition of unnecessary variables.
BIC: BIC balances the goodness of fit of the model with the number of parameters, penalising complex models to avoid overfitting. Lower BIC values indicate better model fit while considering model complexity.

Model	Test RMSE	Train RMSE	adjr2	BIC
Linear	39.58	38.63	0.9076	6844
SVM Linear	41.39	40.72	0.9006	6937
SVM Poly	74.84	72.90	0.7557	8445
SVM RBF	63.09	60.79	0.8592	7029
SVM SIG	147.25	142.90	Negative	14189
RF Linear	26.96	9.77	0.9516	5879
RF SQRT	25.59	9.24	0.9584	5528
MLP a=0.0001	21.89	19.21	0.9710	3807
MLP a=0.001	22.16	19.08	0.9703	3776
MLP a=0.01	22.19	19.38	0.9711	3717

Table 4: Model Performance

From the above result, we can choose MLP with hidden layer size (50,50) and alpha=0.01 as our optimal model, as it has second-lowest Test MSE, highest adjusted R square, and lowest BIC.
Among other 3 types of models, random forest performs well, however, the Test RMSE is nearly three times the Train RMSE, implying the potential problem of overfitting. Maximum features in this situation have little to none effect on our model, as we can see the two RF models do not differ a lot in each indice.
It is worth mentioning MLR performs better than any SVM model, this may be caused by data size and relationships between variables. SVM Model with linear kernel performs best among all kernel types, while sigmoid kernel is obviously a worst choice.

8.2 Limitations and future improvements

MLP computational complexity: The running time of MLP models is significantly longer than other models, and further tuning of parameters may improve this issue.
KNN Imputation: For NMHC(GT), we impute almost 90% of the data as they are missing. Although we use KNN to decrease the noise brought by imputation, it is still a problem. Besides, we use the default K (5), we could use cross validation to determine K for better imputation.
Data Split: We split the data by assigning 80% data to training set and 20% to the testing set. However, if we change the weights we assign to the training and testing set, the error rate may change. For example, if we give 90% weight for training data set and the rest 10% for test data set, the training error rate may have two different changes. For one, the training error rate may decrease a lot since there are more data for learning so the model will fit very well. However, it also causes the problem of overfitting and the test error rate might be higher. For the other, the training error rate may increase because more addition of the data may introduce a significant amount of noise or mislabeled data making the model struggle with distinguish the correct and incorrect ones.
Random Forest Overfitting: Although according to our selection criteria, MLP will be chosen regardless of RF's overfit, we still need further evidence to conclude it is overfitted.