Homework 3

Topics on Optimization and Machine Learning

Roberto Meroni
email: roberto.meroni@estudiantat.upc.edu

May 2024

Contents

1 Introduction

Black carbon (BC) is a major component of fine particulate matter, contributing to warming, environmental disruption, and health issues like cardiovascular and respiratory diseases. Primarily from incomplete combustion in road traffic, BC is prevalent in urban areas but challenging to monitor and unregulated by European Union Air Quality Directives.

The European Union's 2022 air quality proposal highlighted the need to monitor emerging pollutants like Black Carbon. Due to the high cost of traditional monitoring methods, there is growing interest in using measurements from cheaper sensors for other pollutants to virtually estimate black carbon levels.

The objective is to develop a proxy for Black Carbon using machine learning techniques, aiming to enhance the availability and accuracy of Black Carbon measurements, thereby supporting more informed decision-making in environmental policy and public health.

2 Understanding the data

2.1 Black Carbon correlations

Correlations with all the pollutants are found to be highly significant (p-value < 0.001).

Significantly Positive Correlation with N_CPC ($r = 0.515$): This positive correlation indicates that higher levels of Black Carbon are associated with higher levels of ultrafine particle number concentration. This relationship suggests that sources emitting Black Carbon may also be significant sources of ultrafine particles, often related to combustion processes such as vehicle emissions and industrial activities.

Moderate Positive Correlation with PM-2.5 ($r = 0.505$) and PM-1.0 ($r = 0.496$): These moderate positive correlations suggest that higher concentrations of Black Carbon are associated with higher concentrations of fine particulate matter (PM-2.5) and ultrafine particulate matter (PM-1.0). This is expected as Black Carbon is a component of particulate matter resulting from incomplete combustion.

Moderate Positive Correlation with NO_2 ($r = 0.494$) and NO_X ($r = 0.466$): These correlations indicate that higher levels of Black Carbon are associated with higher levels of nitrogen dioxide and nitrogen oxides. Both BC and NO_X are typically emitted from similar sources, such as vehicle exhaust and industrial processes, explaining their moderate positive relationship.

Weak Positive Correlation with CO ($r = 0.260$): This weak positive correlation suggests a minor relationship between Black Carbon and Carbon Monoxide. While both pollutants are products of combustion, their levels might vary due to differences in emission sources and atmospheric processing.

Moderate Negative Correlation with O_3 ($r = -0.353$): This moderate negative correlation indicates that higher levels of Black Carbon are associated with lower levels of ozone. This could be due to the fact that Black Carbon absorbs sunlight, reducing the amount available for ozone formation through photochemical reactions.

2.2 Significant correlations between other pollutants

- As expected, NO_X is highly correlated with NO ($r = 0.924$) and NO_2 ($r = 0.8704$), as NO_X includes NO and NO_2.
- *PM-1.5* and *PM-2.5* are also highly correlated ($r = 0.954, ; p$-value < 0.001), as they represent similar size fractions of particulate pollutants. A high but less strong correlation is also found between *PM-2.5* and *PM-10* ($r = 0.6658$).
- *O3* is found to be negatively correlated with nitrogen oxides (NO, NO_2, NO_X), with correlation coefficients varying from -0.439 to -0.6725. This is due to chemical reactions between nitrogen oxides and *O3*, leading to the destruction of ozone.

Component 1	**Component 2**	**Correlation (r)**
PM-2.5	PM-1.0	0.9537
NO	NOX	0.9236
NO2	NOX	0.8704
NO2	O3	-0.6725
PM-10	PM-2.5	0.6658
NO2	NO	0.6192
O3	NOX	-0.6047
CO	NOX	0.5795
CO	NO	0.5556
BC	N_CPC	0.5147

Table 1: Most Relevant Correlations

2.3 Effects of Temperature and Humidity on air pollutants

2.3.1 Temperature

Positive Correlation with O_3 ($r = 0.361$): Higher temperatures are associated with higher levels of ozone. Ozone formation is strongly influenced by sunlight and temperature. Warmer temperatures enhance the photochemical reactions involving nitrogen oxides (NOx) and volatile organic compounds (VOCs) that produce ozone.

Negative Correlation with PM-2.5 ($r = -0.146$) and PM-1.0 ($r = -0.234$): Higher temperatures are associated with lower levels of PM-2.5 and PM-1.0. Higher temperatures can enhance the dispersion of particulate matter and reduce its concentration near the ground. Additionally, higher temperatures can lead to increased volatilization of certain particulate matter components, reducing their concentration in the atmosphere.

Negative Correlation with CO ($r = -0.309$): Higher temperatures are associated with lower levels of carbon monoxide. During warmer periods, atmospheric mixing and dispersion are more efficient, which can lead to lower concentrations of CO near the ground. Additionally, the combustion processes that produce CO may be less intensive during warmer periods.

Very Weak or No Significant Correlation with Other Pollutants: The correlations with other pollutants such as BC (Black Carbon), NO_2, NO, and NO_X are either very weak or not statistically significant, indicating that temperature may not have a substantial direct impact on these pollutants under the conditions studied.

2.3.2 Humidity

Positive Correlation with PM-2.5 ($r = 0.254$) and PM-1.0 ($r = 0.267$): Higher humidity levels are associated with higher concentrations of PM-2.5 and PM-1.0. Humidity can influence the formation of secondary particulate matter through chemical reactions in the atmosphere.

Negative Correlation with CO ($r = -0.158$): Higher humidity levels are associated with lower levels of CO. High humidity can enhance the removal processes for CO, such as wet deposition. Additionally, during high humidity conditions, the atmospheric stability may change, enhancing the dispersion and dilution of CO.

Weak or No Significant Correlation with Other Pollutants: The correlations between humidity and other pollutants such as BC, NO_2, NO, NO_X, and

SO_2 are weak or not statistically significant, indicating that humidity may not have a strong direct influence on these pollutants under the conditions studied.

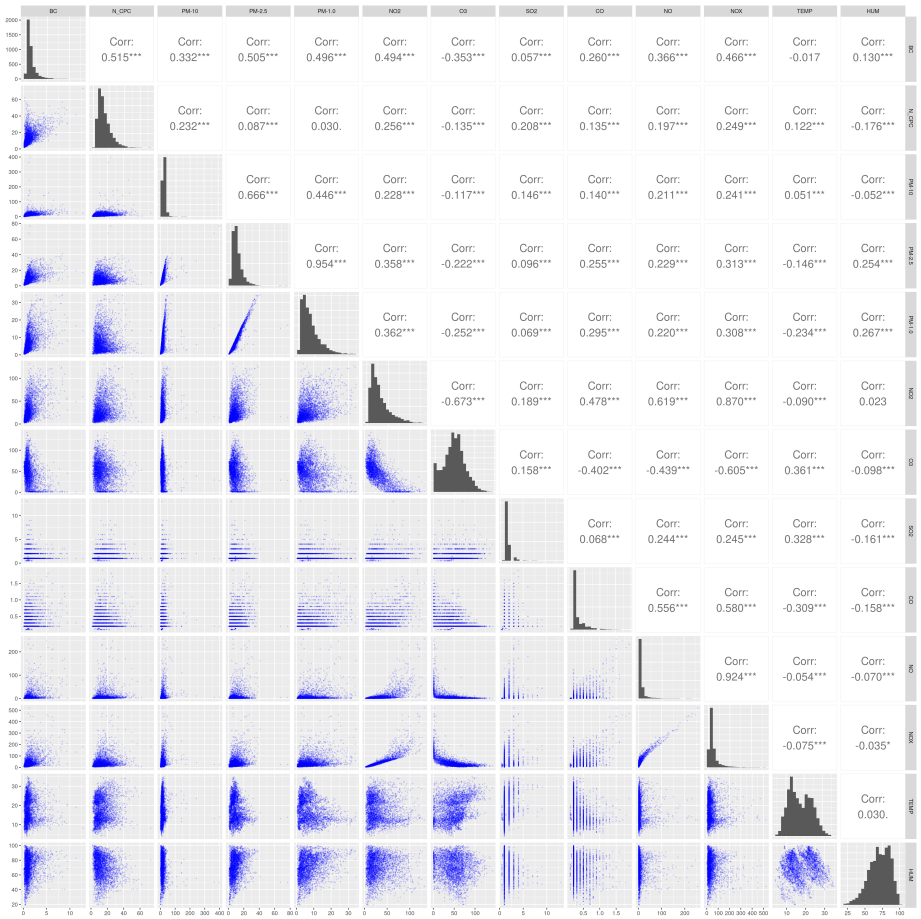
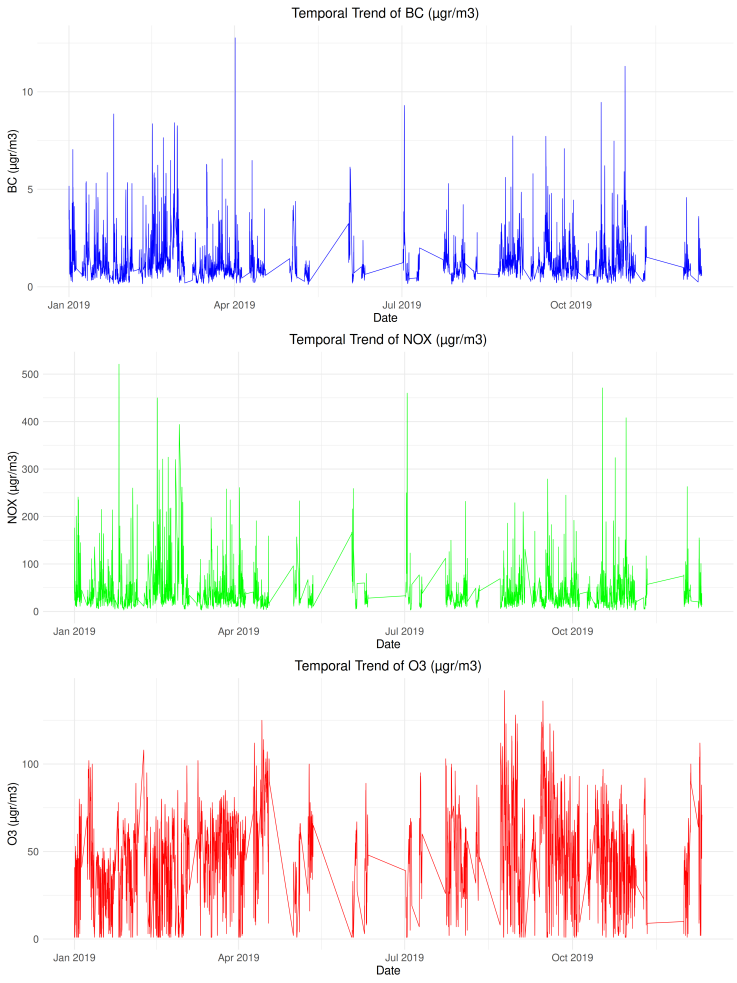


Figure 1: Scatter Plots across all variables

2.4 Temporal Trends

2.4.1 Seasonal Trends



- **Black Carbon (BC):** The concentration of BC (Black Carbon) shows significant variability throughout the year. Peaks are observed at multiple points, indicating short-term spikes.
  During the winter, higher BC concentrations are observed, likely due to an increase in residential heating usage; cold weather also reduces atmospheric dispersion, causing pollutants to accumulate.
  High BC levels are often associated with combustion processes such as vehicle emissions, residential heating, and industrial activities.
- **Nitrogen Oxides (NOX):** Similarly to BC, higher NO_X levels in winter can be attributed to increased use of heating systems, which often rely

on fossil fuels, leading to more combustion and higher NO_Xemissions. Additionally, temperature inversions during winter can trap pollutants close to the ground, increasing NO_Xconcentrations.

- **Ozone (O3):** Differently, Ozone levels are highest in the summer due to increased sunlight and higher temperatures, which enhance the photochemical reactions that produce ozone.
  In winter, the lack of strong sunlight and lower temperatures reduce the photochemical activity, leading to lower ozone levels.

Season	**BC**	**N_CPC**	**PM-10**	**PM-1.0**	**O3**	**CO**	**NOX**
Fall	1.27	13.5	12.4	5.71	45.6	0.289	38.7
Spring	1.12	12.9	15.8	6.41	54.8	0.306	37.5
Summer	1.33	14.2	15.4	7.00	56.3	0.259	38.6
Winter	1.42	13.8	16.6	9.05	40.9	0.371	48.0

Table 2: Average Concentrations by Season

The differences in seasonal averages has been tested with ANOVA and post-hoc Tukey HSD (see Appendix A.1).
The ANOVA results show significant seasonal variations for all pollutants analyzed (BC, N_CPC, PM-10, PM-1.0, O3, CO, NOX), with p-values significantly less than 0.05 for each pollutant.
The post-hoc Tukey HSD tests reveal specific seasonal pairs where these differences are significant. For example, Winter shows higher levels of combustion-related pollutants (BC, PM-10, PM-1.0, CO, NOX) compared to the other seasons, which is likely due to increased heating activities and less atmospheric dispersion. Summer shows higher levels of ozone compared to Fall and Winter, due to increased photochemical activity driven by higher solar radiation.

2.4.2 Daily Trends

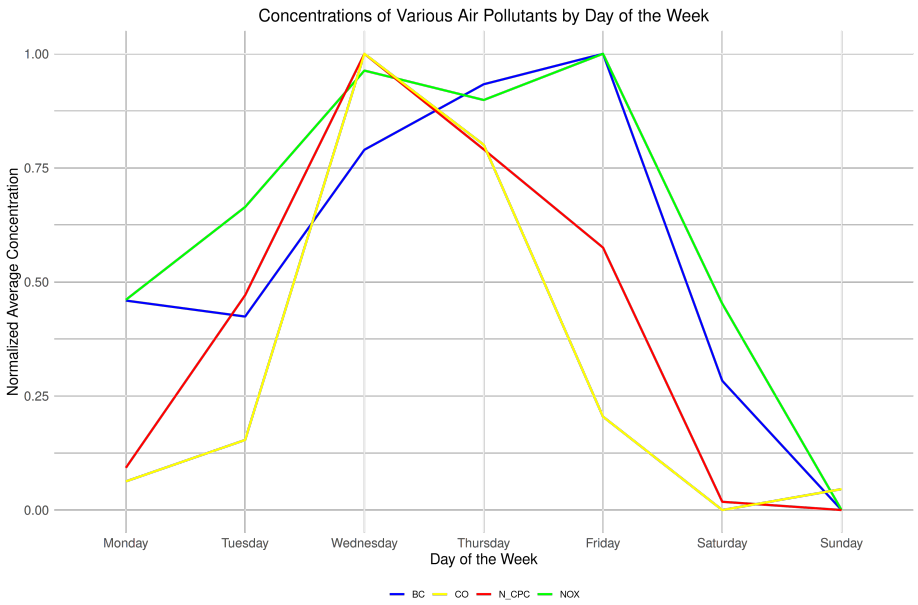


Figure 2: Normalized average concentrations by different days of the week

The differences in daily averages has been tested with ANOVA and post-hoc Tukey HSD (see Appendix A.2).
The ANOVA results show significant differences in pollutant levels by day of the week for all four pollutants analyzed (BC, NOX, N_CPC, CO), with p-values significantly less than 0.05 for each pollutant.
The post-hoc Tukey HSD tests further reveal specific days that differ significantly; there is a significant reduction in pollutant levels during the weekend (especially on Sunday) compared to weekdays. This is likely due to decreased traffic, industrial activities, and overall human activity on weekends.

3 Black Carbon Proxy

Methodology To properly train and test the models, the dataset is split into a training set (80%) and a test set (20%). Both the training and the test sets are then standardized, forcing their mean to be equal to 0 and their standard deviation to be equal to 1.

Note that the data is standardized after the split, not before. If the entire dataset is scaled before splitting, the scaling process uses the information from the entire dataset, including the test set. This means the mean and variance are calculated from all data, which can cause the model to "see" data it shouldn't have access to during training.

For the following analysis, the data is scaled back using the scale applied in the transformation. To visualize the effect of changing the hyperparameters, I will plot the predicted values of BC against the two main components obtained with **PCA** and observe how the predicted values vary when changing the hyperparameters.

Next, R^2 scores for combinations of several parameter values are tested. From this first analysis, two relevant parameters for each model are chosen. For each model, a large combination of values for these two parameters is tested; the best ones will be used in the following applications.

Finally, for each model are compared:

- The highest R^2 from the parameters combination testing.
- The highest R^2 value from FSS.
- The mean R^2 from the shuffling, using the best parameters found in previous analysis.

3.1 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a simple, instance-based learning algorithm used for regression. It makes predictions based on the average of the nearest neighbors' values.

n_neighbors

The number of neighbors to consider for making predictions. Increasing the number of neighbors can smooth the model and reduce variance but may also dilute important local patterns.

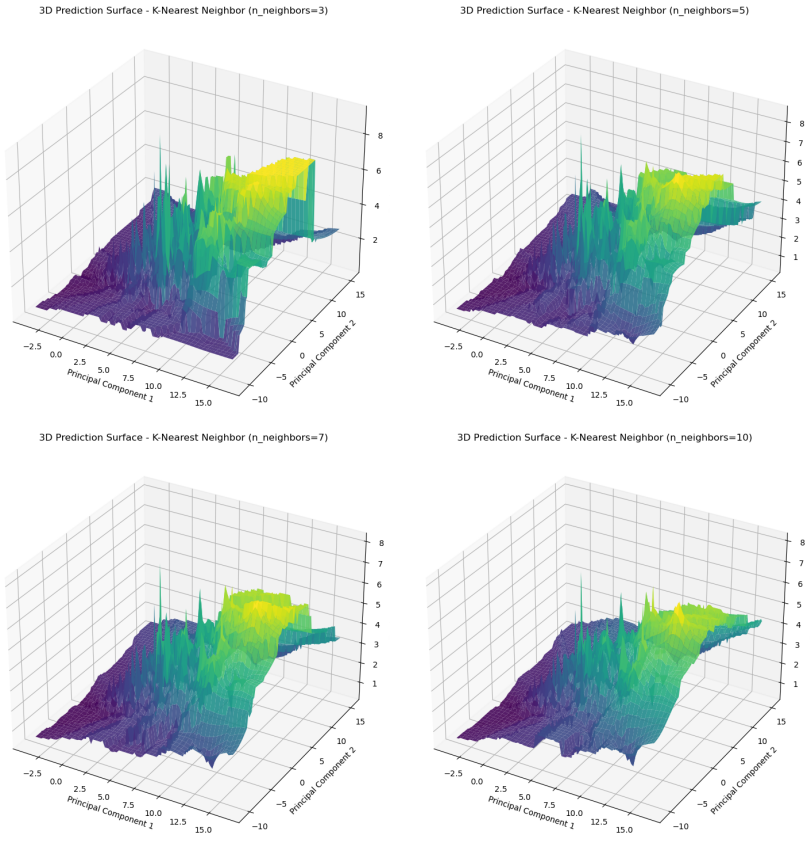


Figure 3: K-Nearest Neighbor with different values of n_neighbors

weights

The weight function used in prediction. Different weighting schemes can affect the influence of neighbors:

- **uniform**: All neighbors are weighted equally.
- **distance**: Closer neighbors have a greater influence, which can improve predictions in variable-density data.

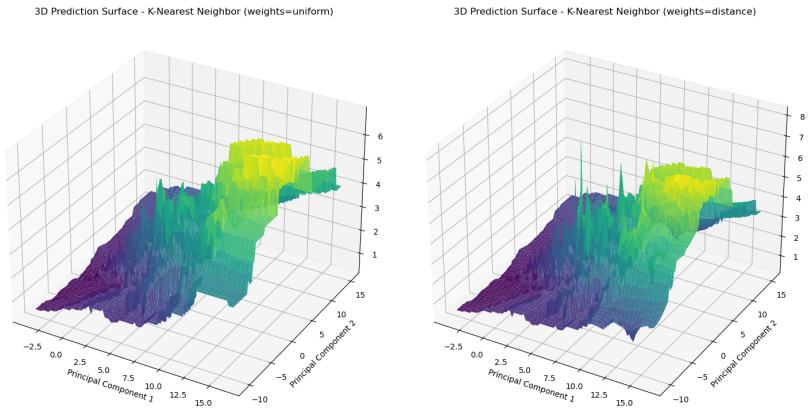


Figure 4: K-Nearest Neighbor with different weighting schemes

algorithm

The algorithm used to compute the nearest neighbors. Different algorithms can handle the data structure and dimensionality more efficiently:

- **auto**: Automatically selects the best algorithm based on data.
- **ball_tree**: Efficient for large datasets.
- **kd_tree**: Suitable for low-dimensional data.
- **brute**: Performs brute-force search.

leaf_size

The size of the leaf in tree-based algorithms (Ball Tree, KD Tree). Larger leaf sizes can speed up the search but may reduce accuracy.

Performance of K-Nearest Neighbor

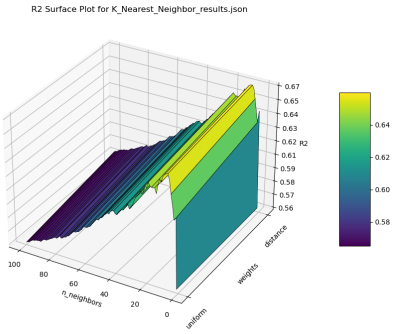


Figure 5: R2 Plot for K-Nearest Neighbor

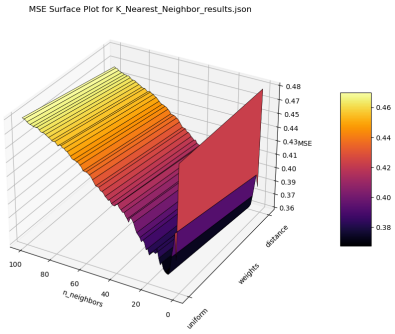


Figure 6: MSE Plot for K-Nearest Neighbor

The parameter n_neighbors is a clear factor in the performance of the model. A very low value leads to bad performances; the performance quickly risens until $n\_neighbors = 7$, and then progressively decreases while increasing the number of neighbors. While the optimal value for the n_neighbor parameter is pretty clear, the weights parameter similarly performed for both 'uniform' and 'distance'.

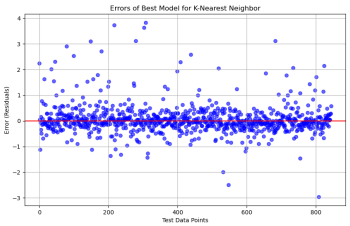


Figure 7: Errors for K-Nearest Neighbor

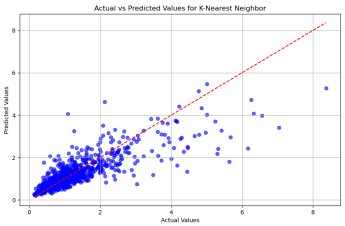


Figure 8: Actual vs Predicted for K-Nearest Neighbor

Forward Subset Selection (FSS) Results for K-Nearest Neighbor

- Selected Features: ['NO2', 'N_CPC', 'PM-1.0', 'TEMP', 'HUM', 'O3']
- Scores for each step: [0.08141883402368755, 0.2829263210663832, 0.5185465257541939, 0.5917637919852528, 0.6419290935503226, 0.6498396631771198]

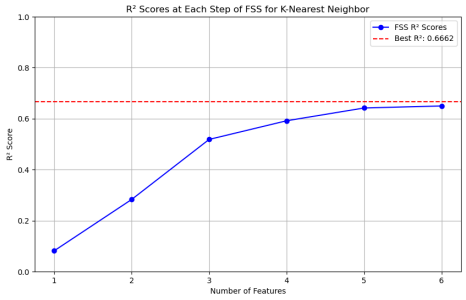


Figure 9: R2 Scores at Each Step of FSS for K-Nearest Neighbor

With five or six features, the model has approximately the same performance as the full model, while gaining in simplicity, interpretability, and reduced computational costs. It includes some of the most correlated features, but curiously, it also includes 'temperature', the variable that has the lowest correlation with Black Carbon.

Shuffling Statistics for K-Nearest Neighbor

After shuffling the data and evaluating the model K-Nearest Neighbor, the following statistics were obtained:

- Original Score: 0.6661613476367061
- Mean Shuffled Score: 0.6661613476367061

- Standard Deviation of Shuffled Scores: 2.220446049250313e-16
- 95% Confidence Interval of Shuffled Scores: [0.6661613476367061, 0.6661613476367061]
- Average Difference: 0.0000000000000000

Standard Deviation and Confidence Interval suggest the Shuffling procedure has no impact on the performance of the model.

Performance Metrics of K-Nearest Neighbor

- Parameters: {n_neighbors: 7, weights: 'distance'}
- Standard R^2: 0.6661613476367061
- FSS R^2: 0.6498396631771198
- Shuffle Average R^2: 0.6661613476367061

K-Nearest Neighbor has good performance, with decent R^2 scores. FSS only slightly reduces the performances, and Shuffling does not impact the performance of the model.

3.2 Support-Vector Regression

Support-Vector Regression (SVR) is a type of Support Vector Machine (SVM) that is used for regression problems. SVR aims to find a function that deviates from the actual observed targets by a value no greater than ϵ for each training point, and at the same time, is as flat as possible. This method is particularly effective for high-dimensional data and can handle non-linear relationships through the use of kernels.

Kernel

The kernel function in SVR transforms the input data into a higher-dimensional space, allowing the model to handle non-linear relationships.
Four different types of kernels are explored:

- **Linear:** No transformation, the model remains in the original feature space.
- **Polynomial:** Transforms the input data into polynomial features.
- **Radial Basis Function (RBF):** Maps the input data into an infinite dimensional space, allowing the model to handle very complex relationships.

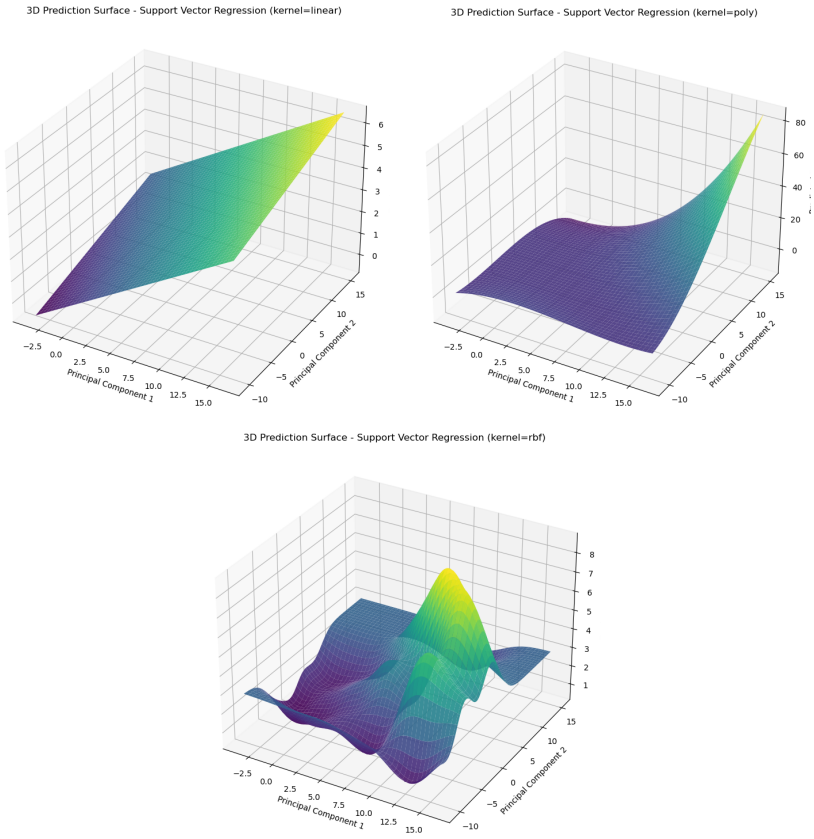


Figure 10: SVR with different kernels

C

The regularization parameter C in SVR controls the trade-off between achieving a low error on the training data and maintaining a flat regression function. A small C value makes the decision surface smooth, while a large C value aims to fit the training data as well as possible, which may lead to overfitting.

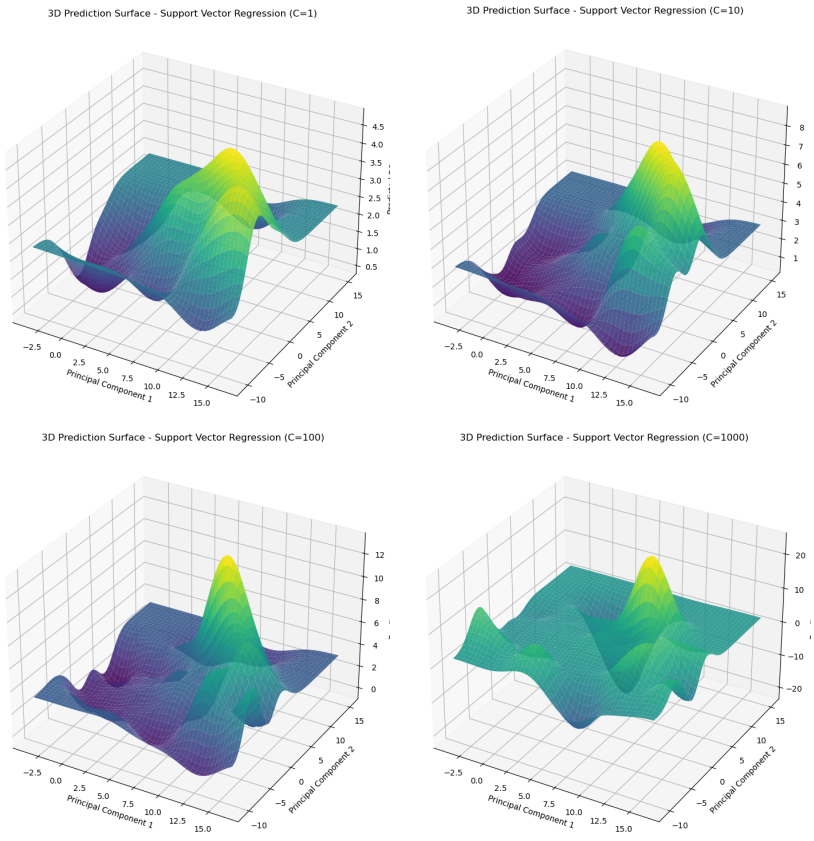


Figure 11: SVR with different values of C

Gamma

The parameter γ defines how far the influence of a single training example reaches. Low values of γ mean a far reach, affecting more points, leading to a smoother decision boundary. High values of γ mean a close reach, making the decision boundary more sensitive to individual points, which can lead to overfitting.

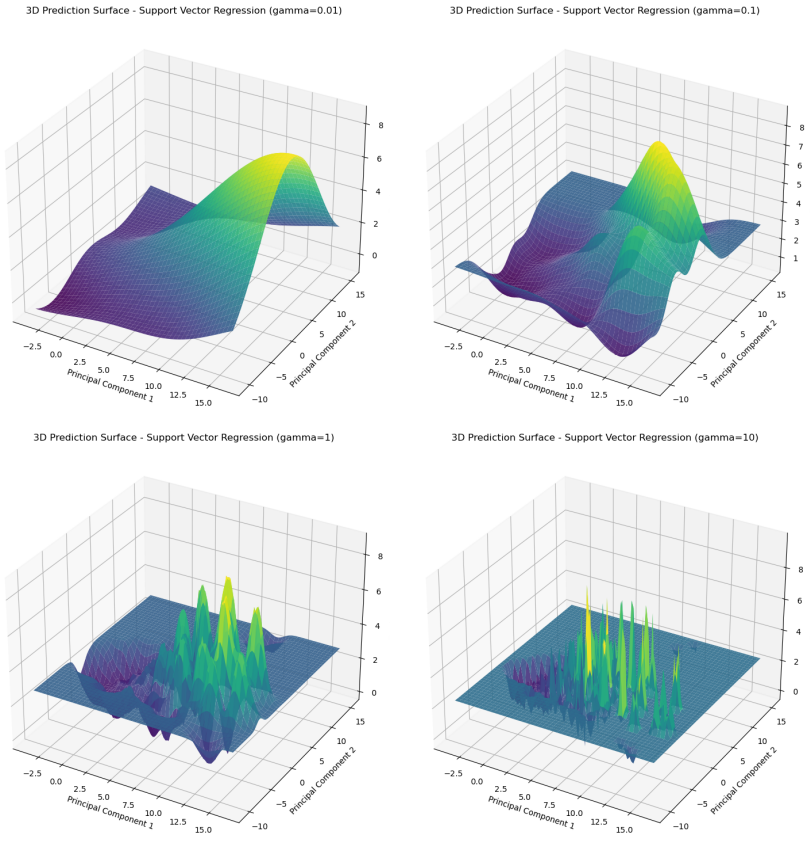


Figure 12: SVR with different values of Gamma

Epsilon

The ϵ parameter in SVR defines the width of the margin around the regression line within which no penalty is given to errors. This margin allows the model to ignore small errors and focus on larger deviations. A small ϵ value makes the model sensitive to the training data, potentially overfitting, while a large ϵ value makes the model more robust to noise but may underfit the data.

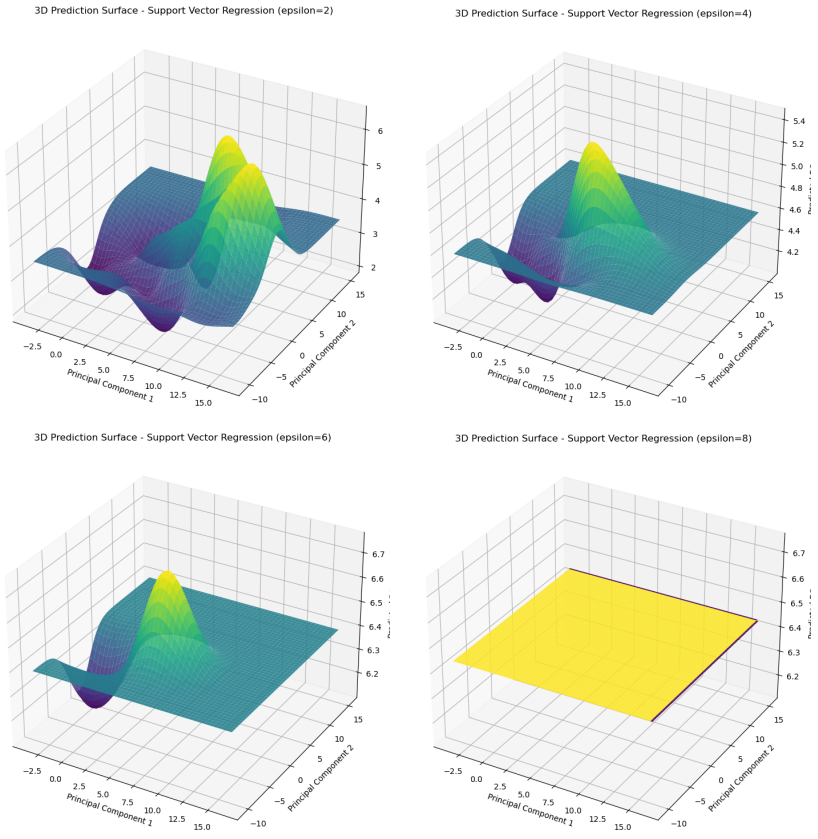


Figure 13: SVR with different values of Epsilon

Performance of Support Vector Regression

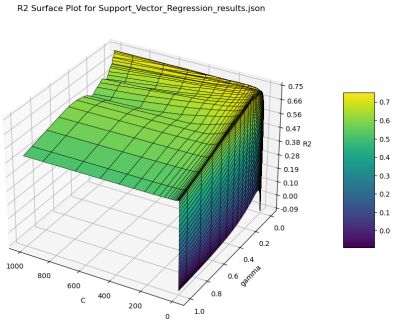


Figure 14: R2 Plot for Support Vector Regression

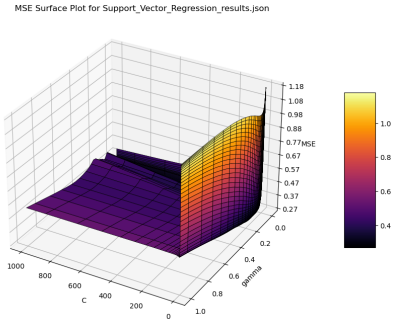


Figure 15: MSE Plot for Support Vector Regression

The parameter C, except for very low values, performs consistently across the whole range. Gamma has a slight impact on the performance of the model, with lower values generally returning higher precision.

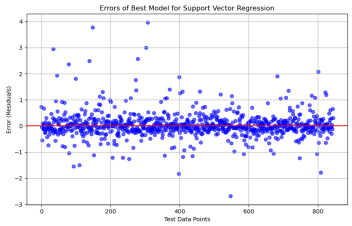


Figure 16: Errors for Support Vector Regression

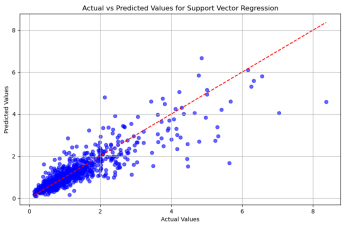


Figure 17: Actual vs Predicted for Support Vector Regression

Forward Subset Selection (FSS) Results for Support Vector Regression

- Selected Features: ['NOX', 'N_CPC', 'PM-1.0', 'TEMP', 'SO2', 'O3', 'HUM', 'PM-2.5', 'NO2']
- Scores for each step: [0.16462297185713023, 0.36592826298181547, 0.5834830926407293, 0.6412600206461635, 0.6875337278163641, 0.6976046486792852, 0.7027724127368008, 0.7040149803826904, 0.7045492915051355]

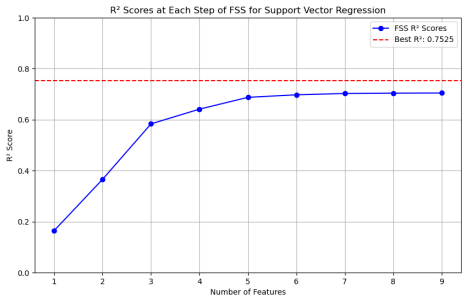


Figure 18: R2 Scores at Each Step of FSS for Support Vector Regression

From five features and above, the model shows a reasonable performance compared to the full model, and gains in simplicity, interpretability, and reduced computational costs. The gain from five to nine features is quite low and not worthy in most of the use cases. The selected features include some of the most correlated features, but curiously, it also include 'temperature' and 'SO2', the variables that have the lowest correlation with Black Carbon.

Shuffling Statistics for Support Vector Regression

After shuffling the data and evaluating the model Support Vector Regression, the following statistics were obtained:

- Original Score: 0.7524588500733992
- Mean Shuffled Score: 0.7524588500733992
- Standard Deviation of Shuffled Scores: 2.220446049250313e-16
- 95% Confidence Interval of Shuffled Scores: [0.7524588500733992, 0.7524588500733992]
- Average Difference: 0.0000000000000000

Standard Deviation and Confidence Interval suggest the Shuffling procedure has no impact on the performance of the model.

Performance Metrics of Support Vector Regression

- Parameters: {C: 40.37017258596558, gamma: 0.03199267137797385}
- Standard R^2: 0.7524588500733992
- FSS R^2: 0.7045492915051355
- Shuffle Average R^2: 0.7524588500733992

Support Vector Regression performs well, with a quite high R^2 score. FSS led to a loss in performance; Shuffling did not impact the performance of the model.

3.3 Elastic Net Regression

Elastic Net Regression is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. It is particularly useful when there are multiple correlated features.

alpha

The regularization parameter that controls the overall strength of the penalty. Increasing alpha adds more regularization, which can improve generalization by reducing overfitting.

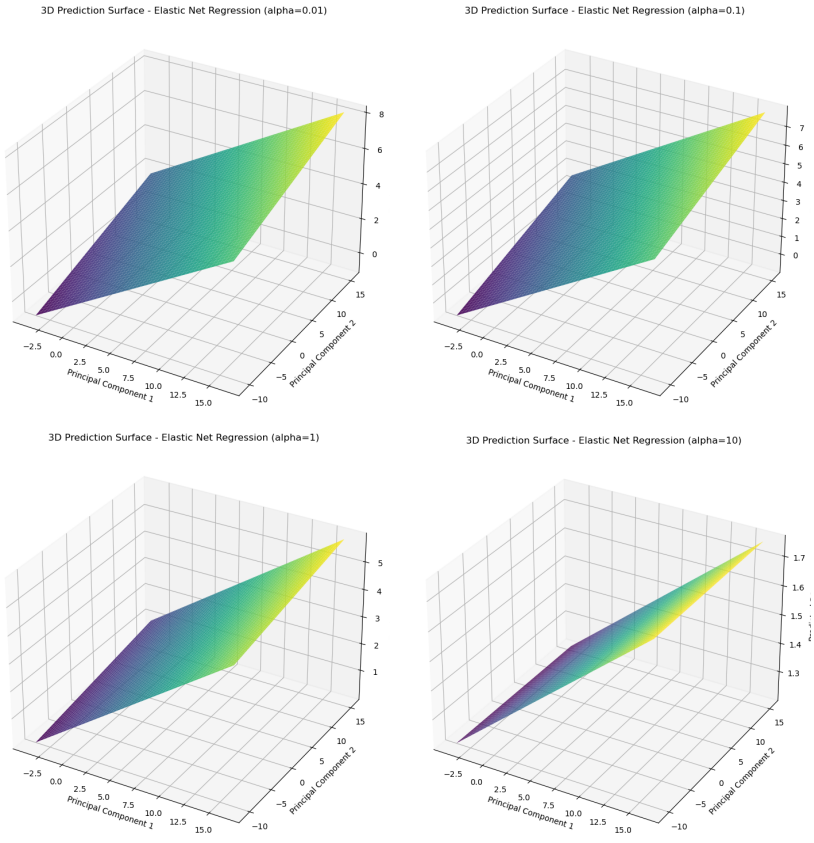


Figure 19: Elastic Net Regression with different values of alpha

l1_ratio

The mixing parameter between L1 (lasso) and L2 (ridge) penalties. Adjusting the l1_ratio balances the contribution of the L1 and L2 penalties, with different ratios capturing different sparsity levels and correlated features.

max_iter

The maximum number of iterations for the optimization algorithm to converge. Increasing max_iter allows more time for convergence, potentially improving model accuracy.

Performance of Elastic Net Regression

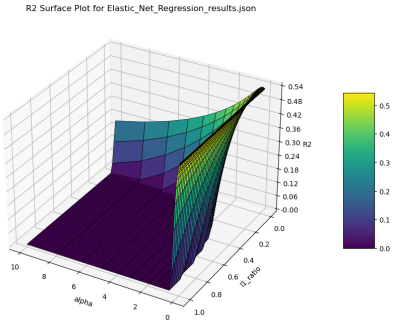


Figure 20: R2 Plot for Elastic Net Regression

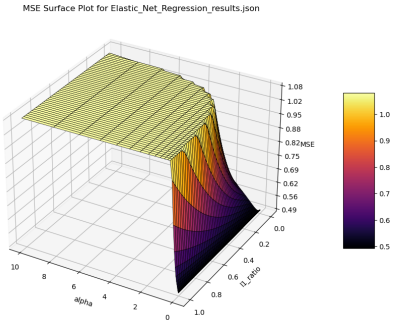


Figure 21: MSE Plot for Elastic Net Regression

The model shows better performances for low alpha values and low l1_ratio values, indicating that the L1 penalty returned an higher accuracy than the L1 penalty.

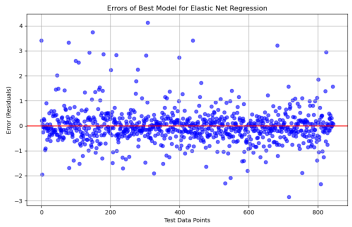


Figure 22: Errors for Elastic Net Regression

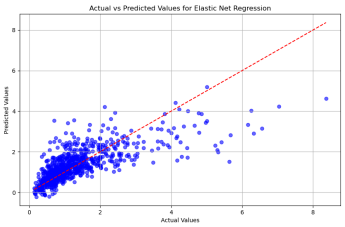


Figure 23: Actual vs Predicted for Elastic Net Regression

Forward Subset Selection (FSS) Results for Elastic Net Regression

- Selected Features: ['PM-2.5', 'N_CPC', 'NO2', 'HUM', 'SO2', 'PM-1.0', 'NOX', 'CO', 'PM-10']
- Scores for each step: [0.23255777372600867, 0.43573270071635195, 0.49085420726490947, 0.5042904018856552, 0.5136514604551177, 0.5215156312026693, 0.5270451108848533, 0.5274115665375864, 0.527412174782219]

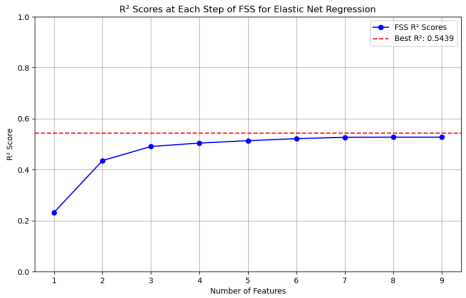


Figure 24: R2 Scores at Each Step of FSS for Elastic Net Regression

From four features and above, the model shows a reasonable performance compared to the full model, and gains in simplicity, interpretability, and reduced computational costs. The gain from five to nine features is quite low and not worthy in most of the use cases. The selected features include some of the most correlated features, but curiously, it also include 'SO2', a variable that have a low correlation with Black Carbon.

Shuffling Statistics for Elastic Net Regression

After shuffling the data and evaluating the model Elastic Net Regression, the following statistics were obtained:

- Original Score: 0.543906377386355
- Mean Shuffled Score: 0.543906377386355
- Standard Deviation of Shuffled Scores: 2.220446049250313e-16
- 95% Confidence Interval of Shuffled Scores: [0.543906377386355, 0.543906377386355]
- Average Difference: 0.0000000000000000

Standard Deviation and Confidence Interval suggest the Shuffling procedure has no impact on the performance of the model.

- Parameters: {alpha: 0.11513953993264481, l1_ratio: 0.061224489795918366}

- Standard R^2: 0.543906377386355
- FSS R^2: 0.527412174782219
- Shuffle Average R^2: 0.543906377386355

Elastic Net Regression has low R^2 scores, indicating a low performance in predicting the BC values. The linear nature of the model might be not as effective as other non-linear models in capturing the complex relationship in the data.

Polynomial Elastic Net A strategy to improve the model might be to transform the original features into a new set of features that include polynomial terms. In my implementation, the feature set is expanded to include all polynomial terms up to 2, using the function *PolynomialFeatures* from sklearn. Then, the Elastic Net model is applied to the expanded set.
Polynomial Elastic Net Regression drastically improves on the basic Elastic Net, going from $R^2 = 0.544$ to $R^2 = 0.680$ using the same parameters of the basic model analysis. This suggests that introducing polynomial features helps the model capture more complex patterns in the data.

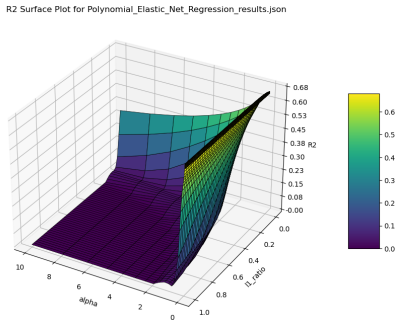


Figure 25: R2 Plot for Elastic Net Regression

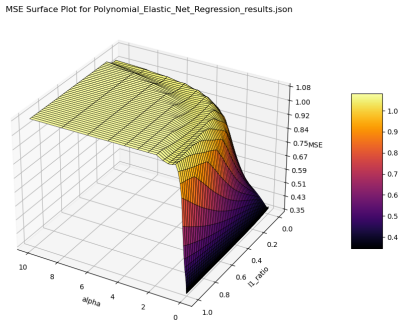


Figure 26: MSE Plot for Elastic Net Regression

The R^2 and MSE distribution is similar to the Linear Elastic Net, but it generally returns higher performances.

3.4 AdaBoost Regression

AdaBoost Regression is an ensemble method that combines the predictions of several base estimators to improve robustness against overfitting. It focuses on difficult cases by adjusting weights.

n_estimators

The number of boosting stages to be run. Increasing the number of stages generally improves performance but also increases the risk of overfitting and computational cost.

learning_rate

The learning rate shrinks the contribution of each base estimator. Lower learning rates require more stages to achieve good performance but can lead to better generalization.

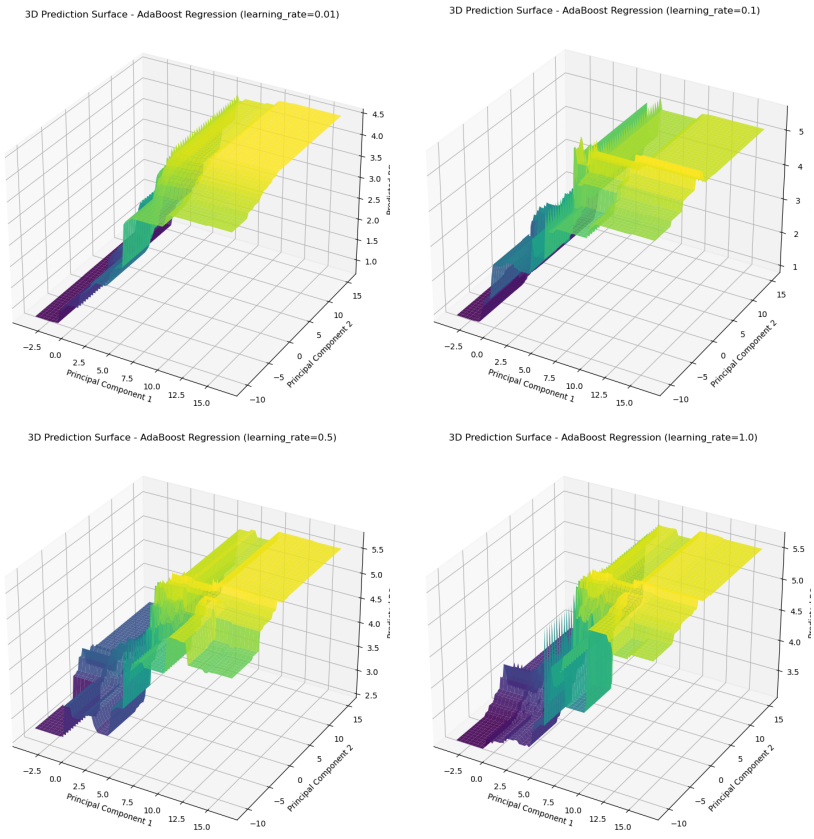


Figure 27: AdaBoost Regression with different values of learning_rate

loss

The loss function to be minimized. Different loss functions can affect the sensitivity to outliers and the robustness of the model:

- **linear**: Linear loss function, suitable for general purposes.
- **square**: Squared loss function, which penalizes larger errors more heavily.
- **exponential**: Exponential loss function, which can be more sensitive to outliers.

Performance of AdaBoost Regression

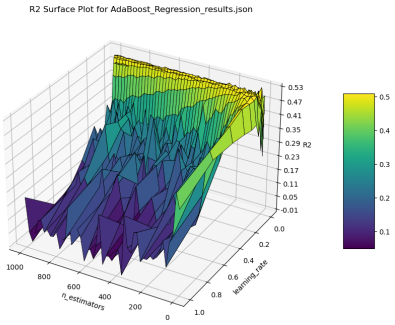


Figure 28: R2 Plot for AdaBoost Regression

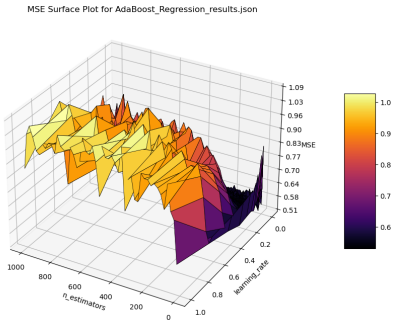


Figure 29: MSE Plot for AdaBoost Regression

The parameter learning_rate is a clear factor in the performance of the model: lower values corresponds to higher accuracy. This is expected by the model; however, low learning rates values lead to an higher computational complexity. While the optimal value for learning_rate is pretty clear, the n_estimators parameter has a variegate and unclear distribution. While usually an higher number of estimators should lead to an improvement in performance, it could also lead to overfitting, and consequently to a lower accuracy on unseen test data.

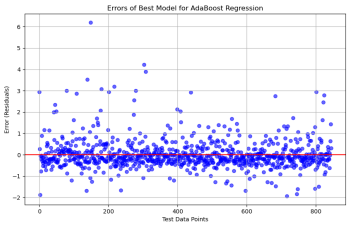


Figure 30: Errors for AdaBoost Regression

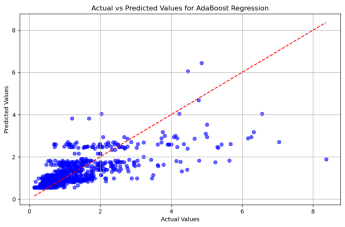


Figure 31: Actual vs Predicted for AdaBoost Regression

Forward Subset Selection (FSS) Results for AdaBoost Regression

- Selected Features: ['PM-2.5', 'N_CPC', 'PM-1.0', 'NOX', 'O3', 'TEMP', 'CO', 'SO2']
- Scores for each step: [0.21971633678902264, 0.45591843530359333, 0.4730695592193843, 0.4828612536125503, 0.4987089026758304, 0.5027297097278038, 0.5056377891365827, 0.5062346290838653]

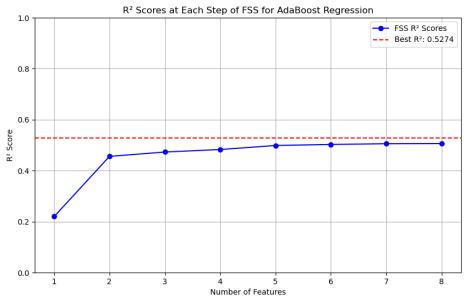


Figure 32: R2 Scores at Each Step of FSS for AdaBoost Regression

From four features and above, the model shows a reasonable performance compared to the full model, and gains in simplicity, interpretability, and reduced computational costs. The gain from five to nine features is quite low and not worthy in most of the use cases. The selected features include some of the most correlated features, but curiously, it also include 'temperature' and 'SO2', the variables that have the lowest correlation with Black Carbon.

Shuffling Statistics for AdaBoost Regression

After shuffling the data and evaluating the model AdaBoost Regression, the following statistics were obtained:

- Original Score: 0.5274003309640303
- Mean Shuffled Score: 0.4656316530906181
- Standard Deviation of Shuffled Scores: 0.028735692903445634
- 95% Confidence Interval of Shuffled Scores: [0.4599011436544216, 0.4713621625268146]
- Average Difference: -0.06176867787341217

Shuffling the data led on average to a decrease of 0.0618 in the R2 score. Standard Deviation and Confidence Interval suggest that the negative impact on the accuracy of the model is statistically relevant.

Performance Metrics of AdaBoost Regression

- Parameters: {n_estimators: 21, learning_rate: 2.5595479226995335e-10}
- Standard R^2: 0.5274003309640303
- FSS R^2: 0.5062346290838653
- Shuffle Average R^2: 0.4656316530906181

AdaBoost Regression shows lower performance, with relatively lower R^2 scores. This suggests that the model may not be as effective in capturing complex patterns in the data. The loss in performance by applying FSS is moderate, and the Shuffling procedure significantly worsens the model.

3.5 Kernel Ridge Regression

Kernel Ridge Regression (KRR) combines Ridge Regression with the kernel trick. It can model non-linear relationships by applying kernel functions.

alpha

The regularization parameter that controls the trade-off between fitting the training data and keeping the model weights small. Higher alpha values increase regularization, reducing the risk of overfitting but possibly underfitting the data.

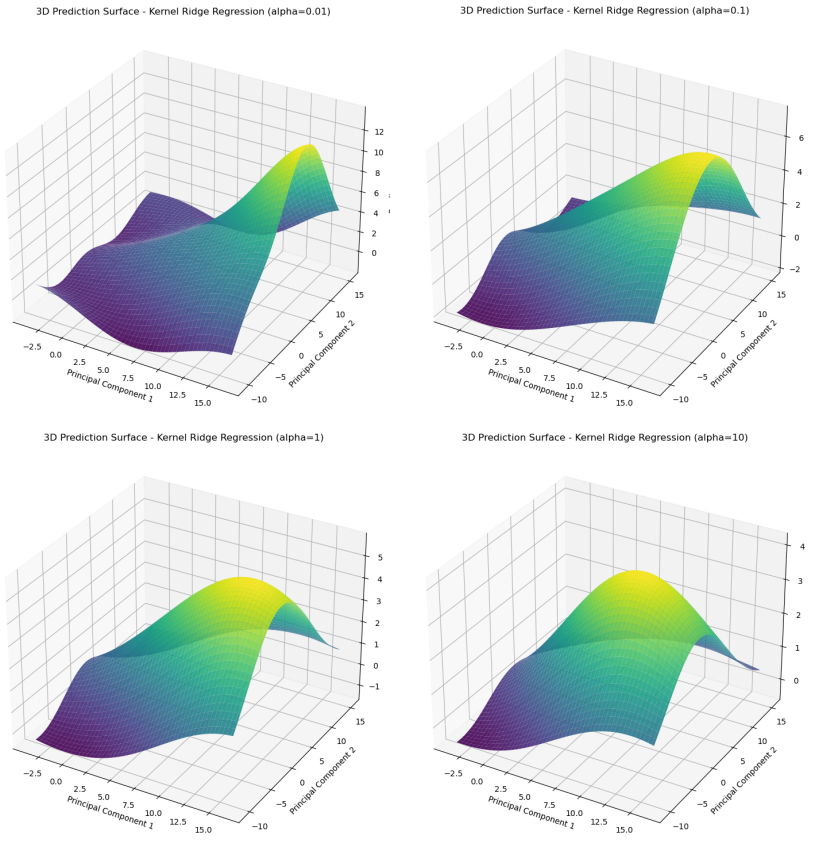


Figure 33: Kernel Ridge Regression with different values of alpha

kernel

The kernel function used to transform the input data. Different kernels can capture different types of relationships:

- **Linear**: No transformation, suitable for linear relationships.
- **Radial Basis Function (RBF)**: Maps the input data into an infinite-dimensional space, allowing the model to handle very complex relationships.
- **Polynomial**: Maps the input data into a higher-dimensional space, allowing the model to capture polynomial relationships.
- **Sigmoid**: Maps the input data using the sigmoid function, useful for capturing certain types of non-linear relationships.

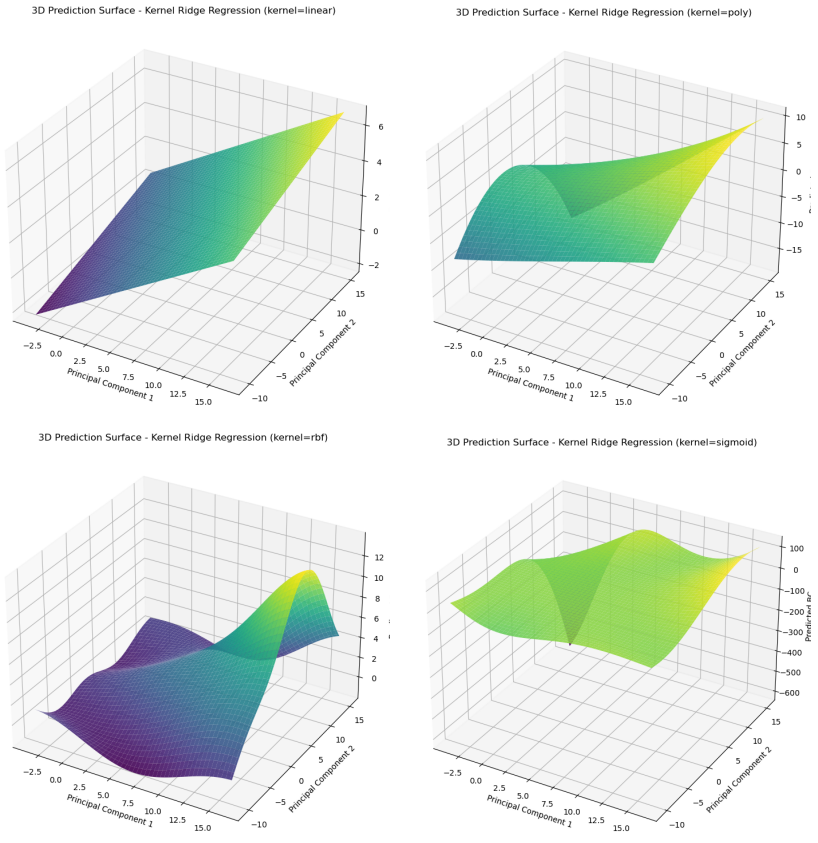


Figure 34: Kernel Ridge Regression with different kernels

gamma

The parameter for the RBF kernel that defines how far the influence of a single training example reaches. A higher gamma value makes the influence more localized, which can capture fine details but increase the risk of overfitting.

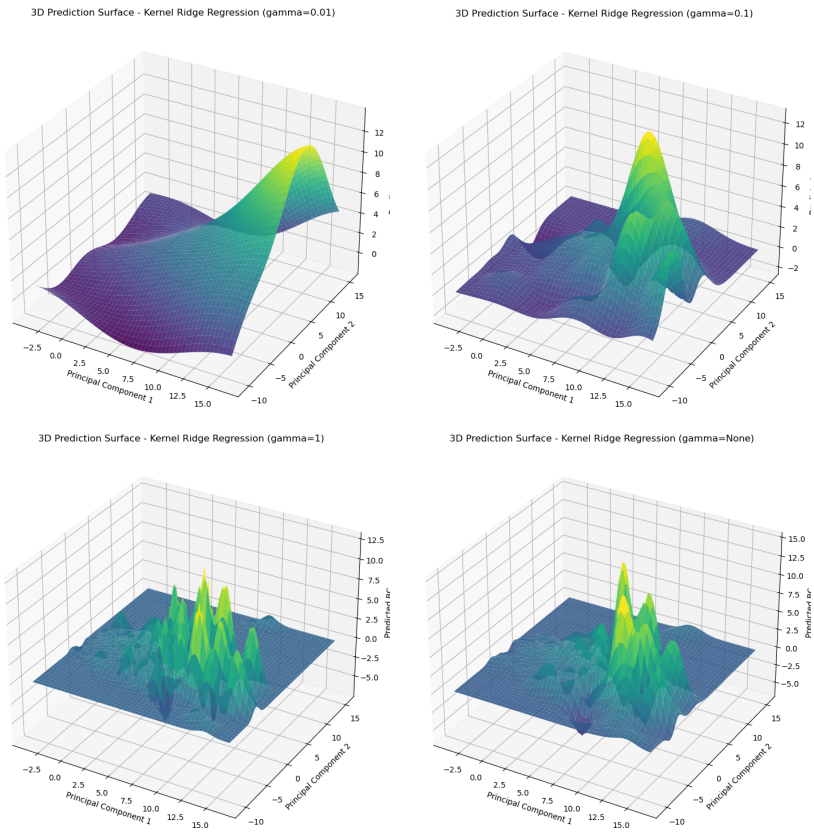


Figure 35: Kernel Ridge Regression with different values of gamma

Performance of Kernel Ridge Regression

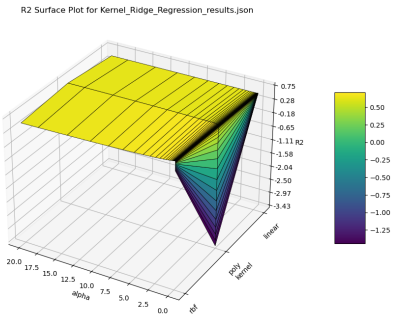


Figure 36: R2 Plot for Kernel Ridge Regression

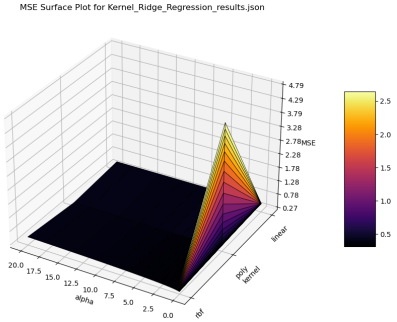


Figure 37: MSE Plot for Kernel Ridge Regression

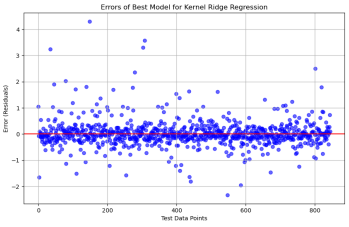


Figure 38: Errors for Kernel Ridge Regression

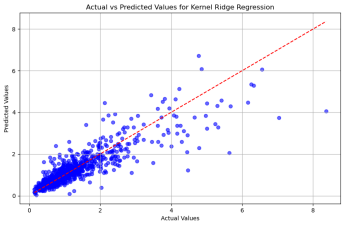


Figure 39: Actual vs Predicted for Kernel Ridge Regression

Except for very low alpha values, where the the polynomial kernel returns inaccurate predictions, the model seems to perform consistently across whole alpha range, no matter the kernel type.

Forward Subset Selection (FSS) Results for Kernel Ridge Regression

- Selected Features: ['PM-2.5', 'N_CPC', 'O3', 'TEMP', 'PM-10']
- Scores for each step: [0.24949491093110415, 0.40213661299236136, 0.5457446052897723, 0.6255878794250085, 0.6305131234477133]

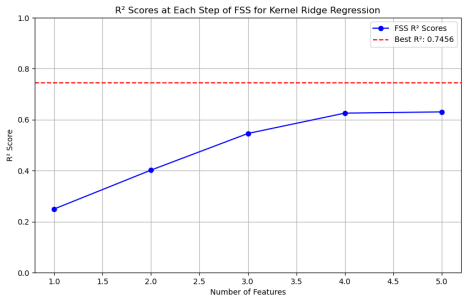


Figure 40: R2 Scores at Each Step of FSS for Kernel Ridge Regression

Even with five different features, FSS fails to accurately represent the full model. It includes some of the most correlated features, but curiously, it also includes 'temperature', the variable that has the lowest correlation with Black Carbon.

Shuffling Statistics for Kernel Ridge Regression

After shuffling the data and evaluating the model Kernel Ridge Regression, the following statistics were obtained:

- Original Score: 0.7456385447437905
- Mean Shuffled Score: 0.7456385447437905
- Standard Deviation of Shuffled Scores: 1.1102230246251565e-16
- 95% Confidence Interval of Shuffled Scores: [0.7456385447437905, 0.7456385447437905]
- Average Difference: 0.0000000000000000

Standard Deviation and Confidence Interval suggest the Shuffling procedure has no impact on the performance of the model.

Performance Metrics of Kernel Ridge Regression

- Parameters: {alpha: 0.04756504112736628, kernel: 'rbf'}
- Standard R^2: 0.7456429217123808
- FSS R^2: 0.6305131234477133
- Shuffle Average R^2: 0.7456429217123808

Kernel Ridge Regression shows strong performance, with quite high R^2 value. This indicates that the model is highly effective at capturing the relationships in the data. FSS led to a significant decrease of R^2 value; Shuffling did not impact the performance of the model.

3.6 Decision Tree Regression

Decision Tree Regression is a non-parametric model that splits the data into subsets based on feature values to make predictions. It is easy to interpret but can easily overfit.

max_depth

The maximum depth of the tree. Deeper trees can model more complex patterns and reduce bias, but they can also lead to overfitting if they become too complex.

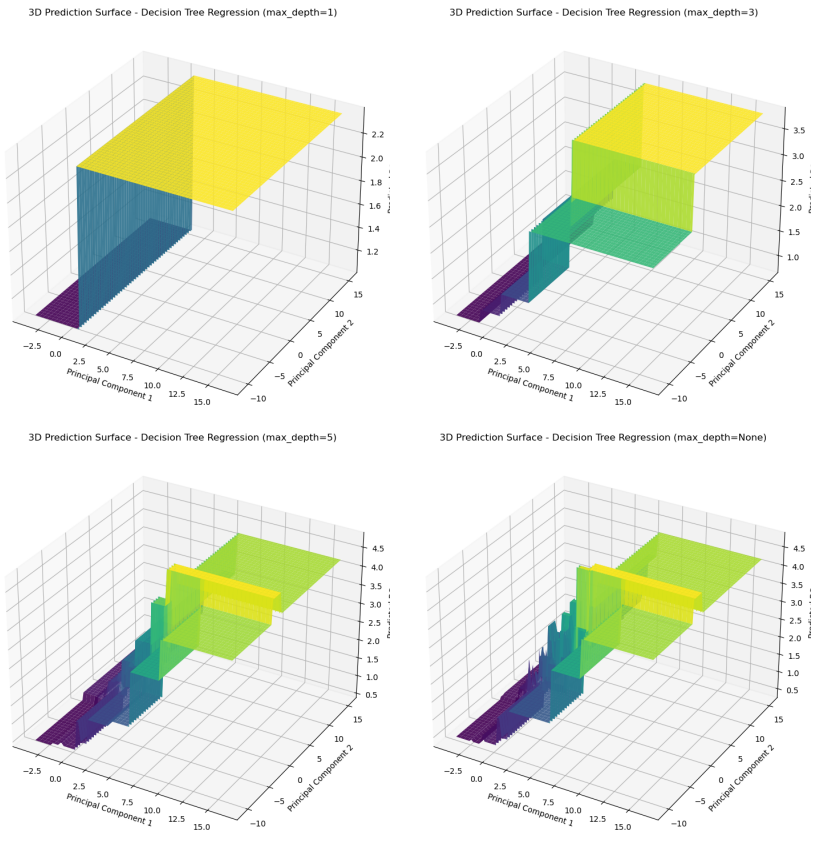


Figure 41: Decision Tree Regression with different values of max_depth

min_samples_split

The minimum number of samples required to split an internal node. Increasing this value helps prevent the model from learning overly specific patterns, reducing overfitting and improving generalization.

min_samples_leaf

The minimum number of samples required to be at a leaf node. Higher values ensure that leaf nodes have sufficient data, smoothing the model and reducing the likelihood of overfitting to noise.

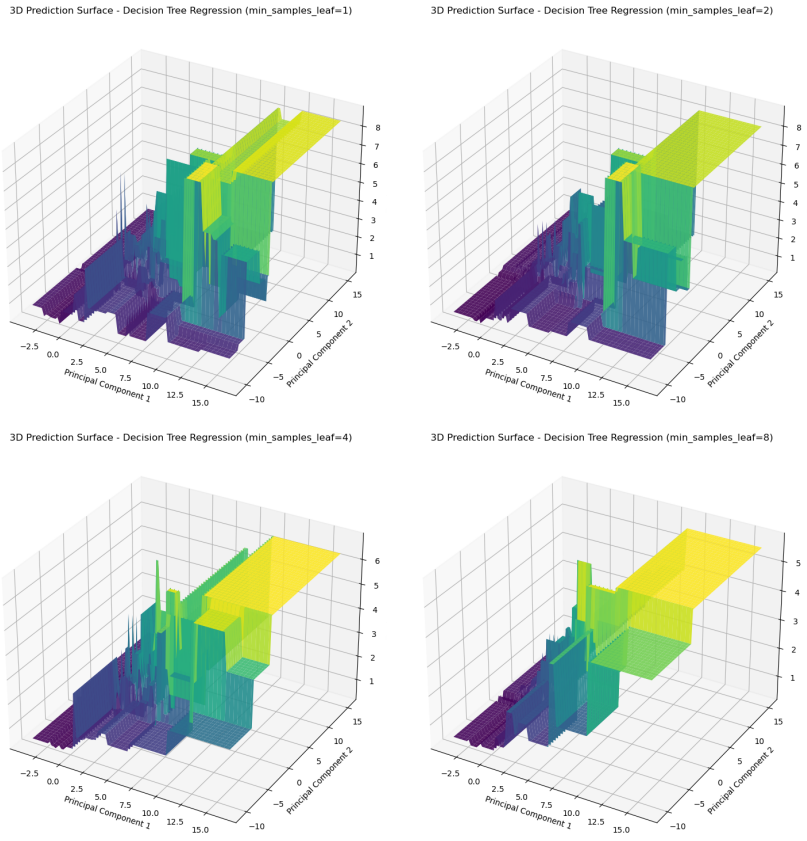


Figure 42: Decision Tree Regression with different values of min_samples_leaf

Performance of Decision Tree Regression

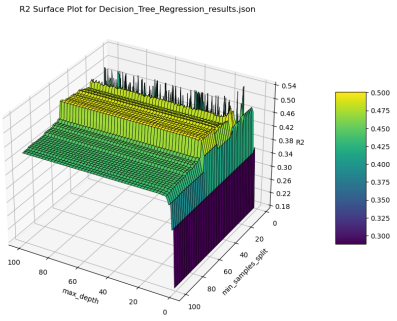


Figure 43: R2 Plot for Decision Tree Regression

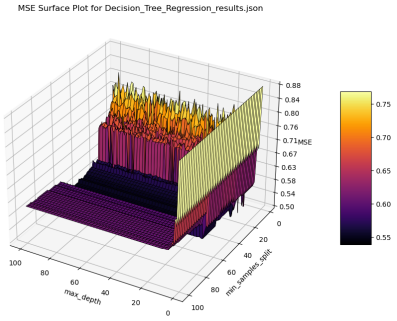


Figure 44: MSE Plot for Decision Tree Regression

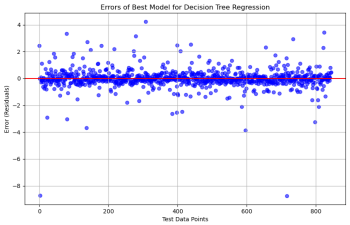


Figure 45: Errors for Decision Tree Regression

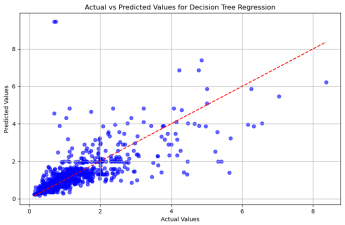


Figure 46: Actual vs Predicted for Decision Tree Regression

For $max\_depth \geq 10$ the accuracy is stable, indicating that a high value of this parameters is not needed and it only results in higher computational costs. Regarding the min_samples_split parameter, the best R^2 values are found between 10 and 50.

Forward Subset Selection (FSS) Results for Decision Tree Regression

- Selected Features: ['O3', 'N_CPC', 'PM-1.0', 'TEMP', 'SO2']
- Scores for each step: [0.18663637663343038, 0.3262331607145642, 0.45078288244317033, 0.5079431245856122, 0.5435792868822141]

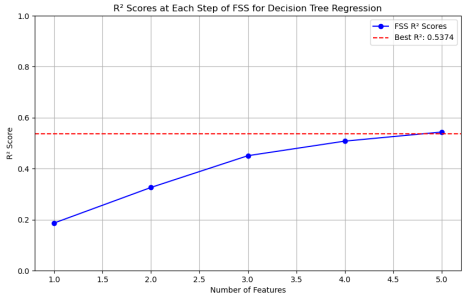


Figure 47: R2 Scores at Each Step of FSS for Decision Tree Regression

The model actually gains in accuracy with the FSS method. With only five features, the model shows a better performance compared to the full model, and gains in simplicity, interpretability, and reduced computational costs. The selected features include some of the most correlated features, but curiously, it also include 'temperature', the variable that has the lowest correlation with Black Carbon.

Shuffling Statistics for Decision Tree Regression

After shuffling the data and evaluating the model Decision Tree Regression, the following statistics were obtained:

- Original Score: 0.5374358169688502
- Mean Shuffled Score: 0.41473951980673535
- Standard Deviation of Shuffled Scores: 0.012119125881591867
- 95% Confidence Interval of Shuffled Scores: [0.407049964169335, 0.4224290754441357]
- Average Difference: -0.12269629716211489

Shuffling the data led on average to a high loss in accuracy, going from $R^2 = 0.5374$ to $R^2 = 0.4147$. Standard deviation and Confidence Interval confirm that the Shuffling procedure has a huge negative impact on the performance of the model.

Performance Metrics of Decision Tree Regression

- Parameters: {max_depth: 9, min_samples_split: 13}
- Standard R^2: 0.5374358169688502
- FSS R^2: 0.5435792868822141
- Shuffle Average R^2: 0.41473951980673535

Decision Tree Regression has a moderate performance, with an R^2 of 0.537. This indicates that the model captures some of the data's structure but may not be as effective as more complex models. FSS actually slightly improved the model; the Shuffling procedure led to an high loss in performance.

3.7 Random Forest

Random Forest is an ensemble learning method used for regression that operates by constructing multiple decision trees during training and outputting the mean prediction of the individual trees. It reduces overfitting by averaging multiple trees.

n_estimators

The number of trees in the forest. Increasing the number of trees generally improves performance by reducing variance and overfitting, but it also increases computation time and memory usage.

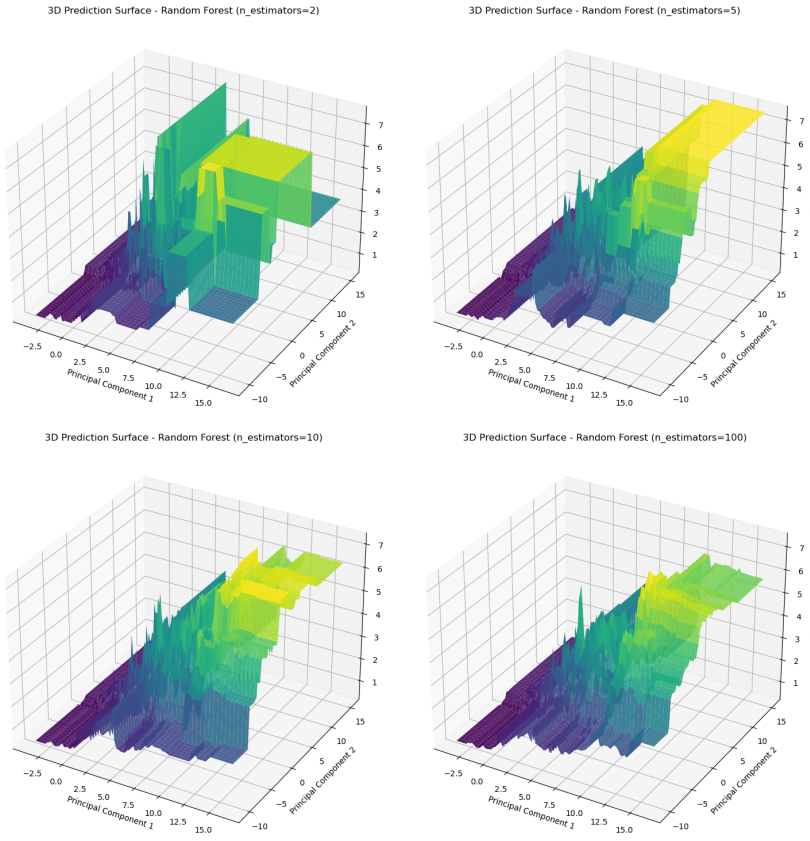


Figure 48: Random Forest with different values of n_estimators

max_depth

The maximum depth of each tree. Deeper trees can model more complex patterns and capture fine details in the data, but they can also lead to overfitting if the trees become too complex.

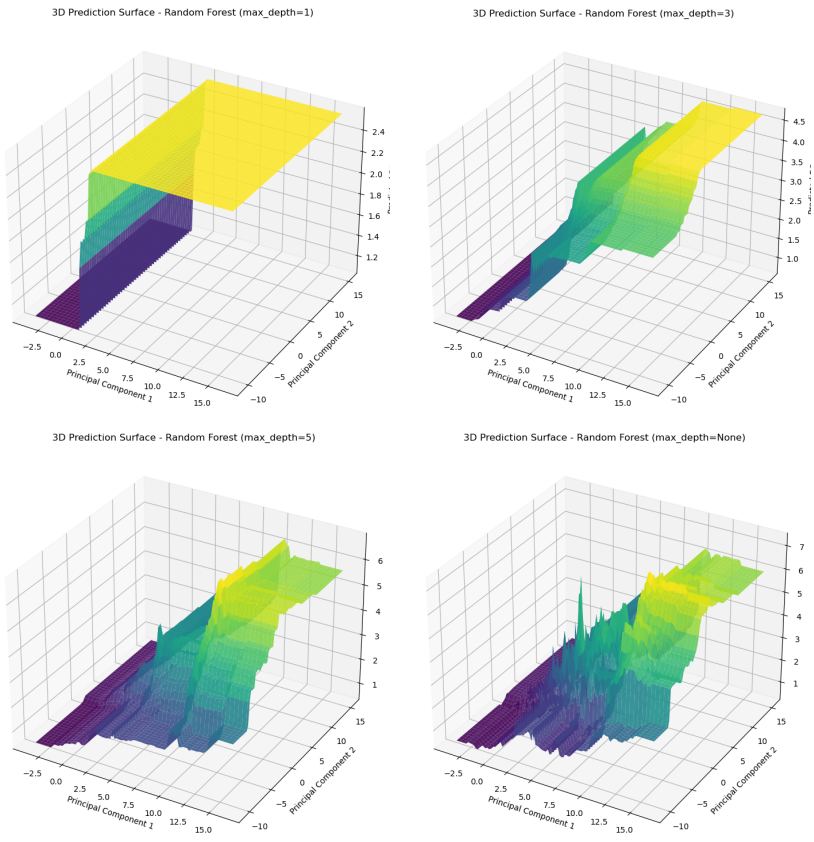


Figure 49: Random Forest with different values of max_depth

min_samples_split

The minimum number of samples required to split an internal node. Higher values prevent the model from learning overly specific patterns, thus reducing overfitting and making the model more robust.

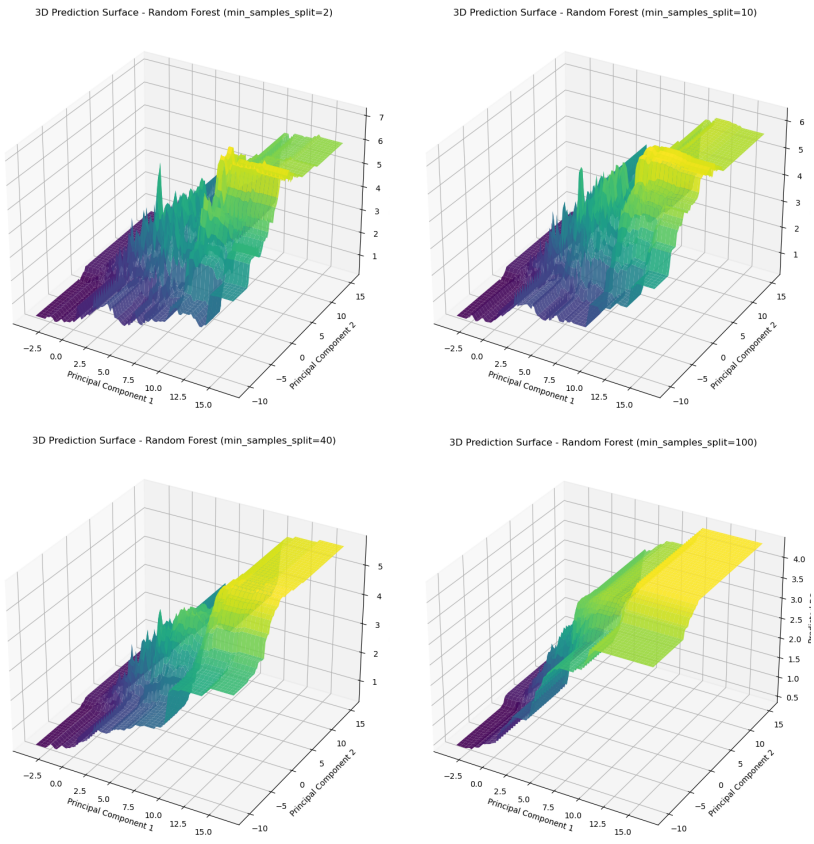


Figure 50: Random Forest with different values of min_samples_split

min_samples_leaf

The minimum number of samples required to be at a leaf node. Higher values can smooth the model by ensuring that leaf nodes have sufficient data, which reduces the likelihood of overfitting to noise.

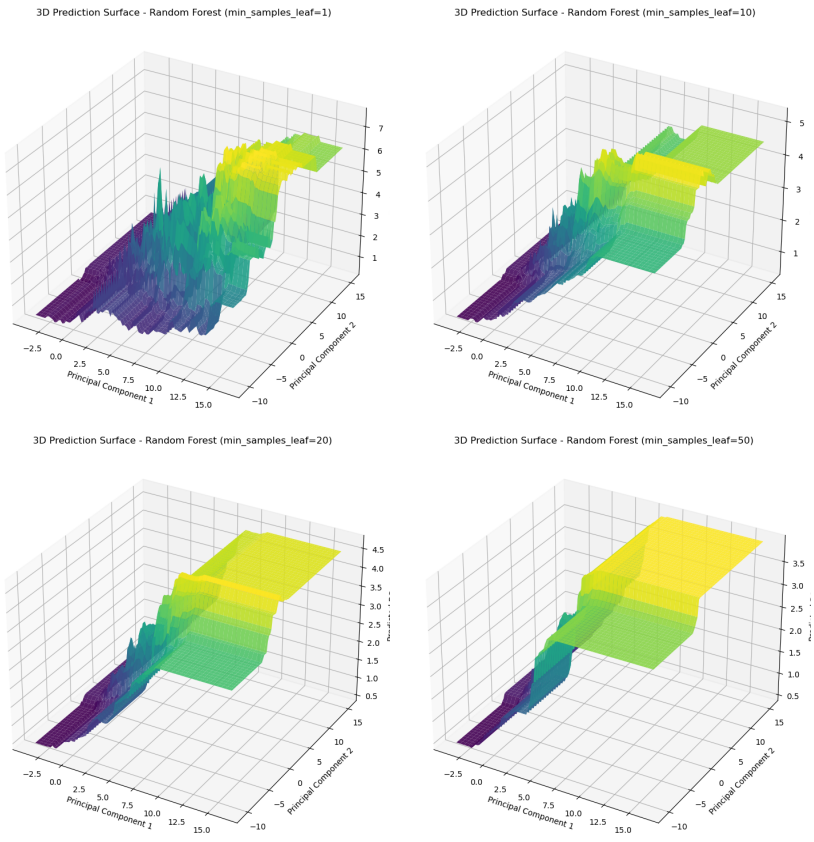


Figure 51: Random Forest with different values of min_samples_leaf

Performance of Random Forest

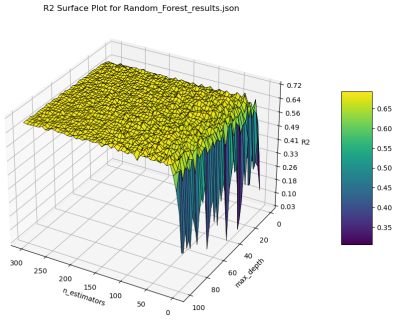


Figure 52: R2 Plot for Random Forest

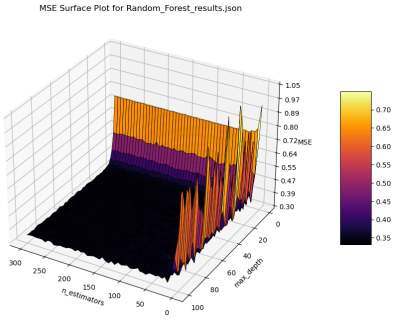


Figure 53: MSE Plot for Random Forest

For both the parameters, the performance is consistent across the whole ranges, except for very low values. After a certain threshold (around 10 for both n_estimators and max_depth), an increased value of the parameters only results in higher computational costs, with no additional advantage.

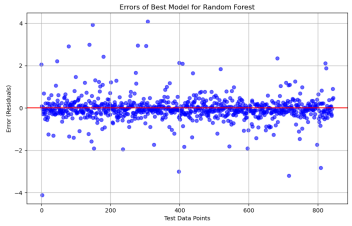


Figure 54: Errors for Random Forest

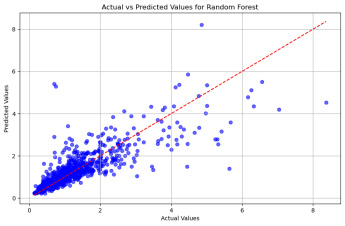


Figure 55: Actual vs Predicted for Random Forest

Forward Subset Selection (FSS) Results for Random Forest

- Selected Features: ['O3', 'N_CPC', 'PM-1.0', 'TEMP', 'NO', 'PM-10', 'PM-2.5']
- Scores for each step: [0.18423853992854658, 0.3130499729204039, 0.55665348688085, 0.6432216807060016, 0.671069648882167, 0.6768081638176464, 0.681359846937063]

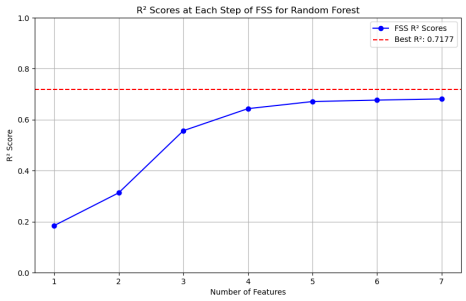


Figure 56: R2 Scores at Each Step of FSS for Random Forest

From four features and above, the model shows a reasonable performance and gains in simplicity, interpretability, and reduced computational costs. The gain from five to seven features is quite low and not worthy in most of the use cases. The selected features include some of the most correlated features, but curiously, it also include 'temperature', the variable that has the lowest correlation with Black Carbon.

Shuffling Statistics for Random Forest

After shuffling the data and evaluating the model Random Forest, the following statistics were obtained:

- Original Score: 0.7177306741506901
- Mean Shuffled Score: 0.6698122770203646
- Standard Deviation of Shuffled Scores: 0.014147911421238384
- 95% Confidence Interval of Shuffled Scores: [0.6669908820334994, 0.6726336720072298]
- Average Difference: -0.04791839713032553

Shuffling the data led on average to a decrease of 0.0479 in the R2 score. Standard Deviation and Confidence Interval suggest the decrease is statistically significant and somehow expected.

Performance Metrics of Random Forest

- Parameters: {n_estimators: 25, max_depth: 97}
- Standard R^2: 0.7177306741506901
- FSS R^2: 0.681359846937063
- Shuffle Average R^2: 0.6698122770203646

Random Forest shows consistent performance with high R^2 scores. FSS led to a slight decrease in performance; Shuffling significantly lowered the accuracy.

3.8 Gaussian Process

Gaussian Process (GP) is a non-parametric, Bayesian approach to regression that defines a distribution over functions and makes predictions based on observed data. It is highly flexible and can model uncertainty directly.

alpha

The noise level in the data. A higher alpha value assumes more noise in the observations, which can make the model more robust to overfitting but may reduce accuracy on clean data.

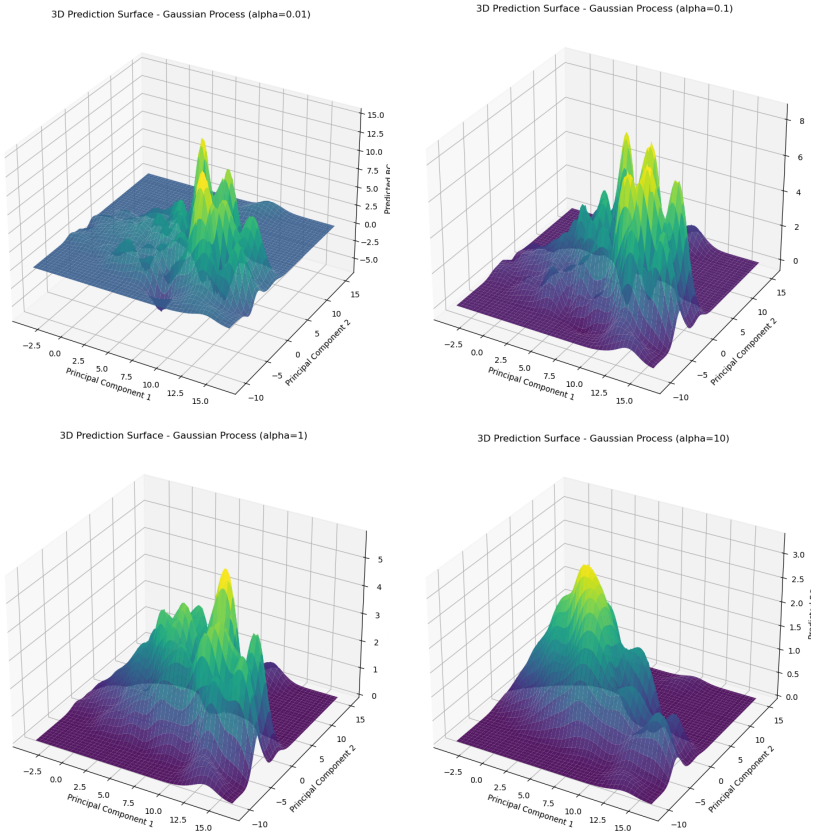


Figure 57: Gaussian Process with different values of alpha

n_restarts_optimizer

The number of restarts of the optimizer for better convergence to a global optimum. Increasing the number of restarts can help avoid local minima and improve model performance but also increases computational cost.

Performance of Gaussian Process

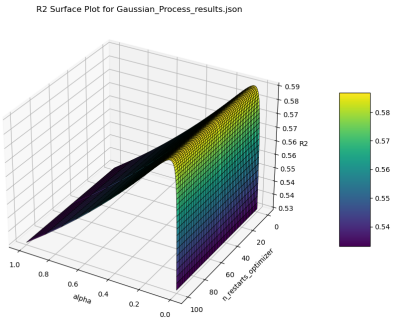


Figure 58: R2 Plot for Gaussian Process

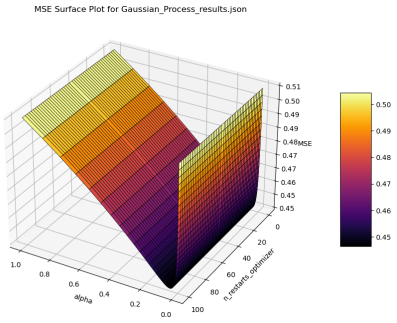


Figure 59: MSE Plot for Gaussian Process

The parameter alpha is a clear factor in the performance of the model. A very low value leads to bad performances; the performance quickly risens until $alpha = 0.05$, and then progressively decreases as we increase the value of alpha. While the optimal value for the alpha parameter is pretty clear, the n_restart_optimizer parameter performed equally across the whole range of its values.

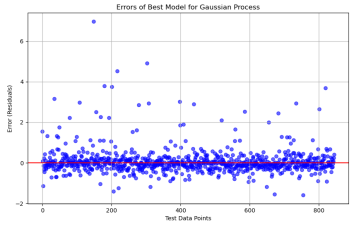


Figure 60: Errors for Gaussian Process

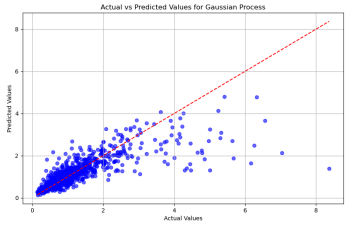


Figure 61: Actual vs Predicted for Gaussian Process

Forward Subset Selection (FSS) Results for Gaussian Process

- Selected Features: ['PM-2.5', 'N_CPC', 'O3', 'TEMP']
- Scores for each step: [0.24963803542091773, 0.40436455015906125, 0.5090856058086002, 0.5555511689911053]

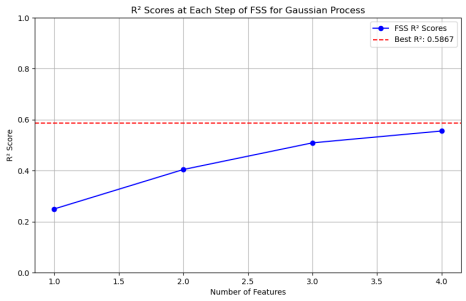


Figure 62: R2 Scores at Each Step of FSS for Gaussian Process

With just four features, the model has almost the same performance as the full model, while gaining in simplicity, interpretability, and reduced computational costs. It includes some of the most correlated features, but curiously, it also includes 'temperature', the variable that has the lowest correlation with Black Carbon.

Shuffling Statistics for Gaussian Process

After shuffling the data and evaluating the model Gaussian Process, the following statistics were obtained:

- Original Score: 0.5867455779355718
- Mean Shuffled Score: 0.5867455779355718
- Standard Deviation of Shuffled Scores: 1.1102230246251565e-16
- 95% Confidence Interval of Shuffled Scores: [0.5867455779355718, 0.5867455779355718]
- Average Difference: 0.0000000000000000

Standard Deviation and Confidence Interval suggest the Shuffling procedure has no impact on the performance of the model.

Performance Metrics of Gaussian Process

- Parameters: {alpha: 0.0517947467923121, n_restarts_optimizer: 0}
- Standard R^2: 0.5867455779355718
- FSS R^2: 0.5555511689911053
- Shuffle Average R^2: 0.5867455779355718

Gaussian Process has moderate performance, with a moderate R^2 value. This suggests that the model may struggle to capture the underlying patterns in the data as effectively as some other models. FSS led to a moderate loss in performance; Shuffling did not impact the performance of the model.

3.9 Gradient Boosting Regression

Gradient Boosting Regression is an ensemble technique that builds models sequentially, each trying to correct the errors of the previous one. It is highly flexible and can be used with various loss functions.

n_estimators

The number of boosting stages to be run. Increasing the number of stages generally improves model performance by reducing bias but also increases the risk of overfitting and computational cost.

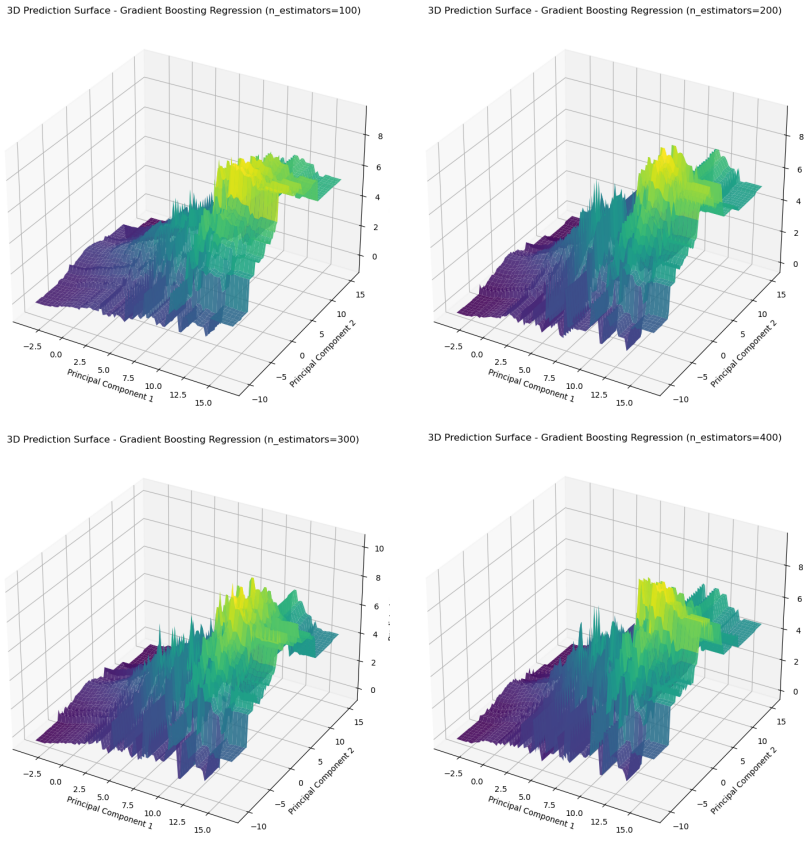


Figure 63: Gradient Boosting Regression with different values of n_estimators

learning_rate

The learning rate shrinks the contribution of each tree. Lower learning rates often lead to better generalization by preventing overfitting, but they require more boosting stages.

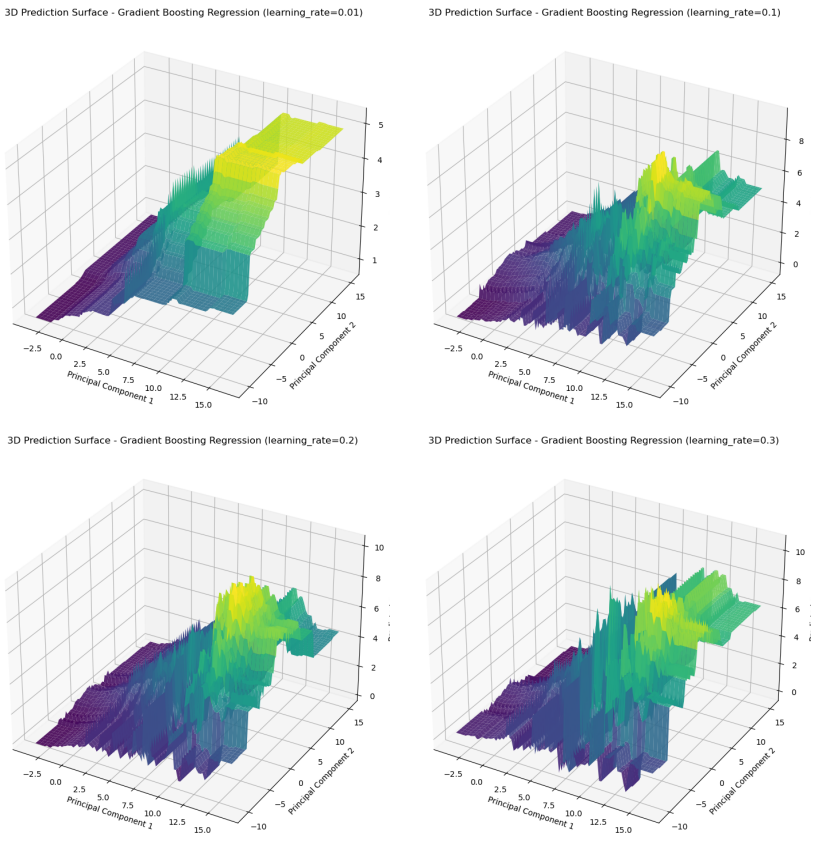


Figure 64: Gradient Boosting Regression with different values of learning_rate

max_depth

The maximum depth of each tree. Deeper trees can capture more complex interactions but can also lead to overfitting if they become too complex.

subsample

The fraction of samples used to fit each base learner. Subsampling introduces randomness, which can help prevent overfitting and improve generalization.

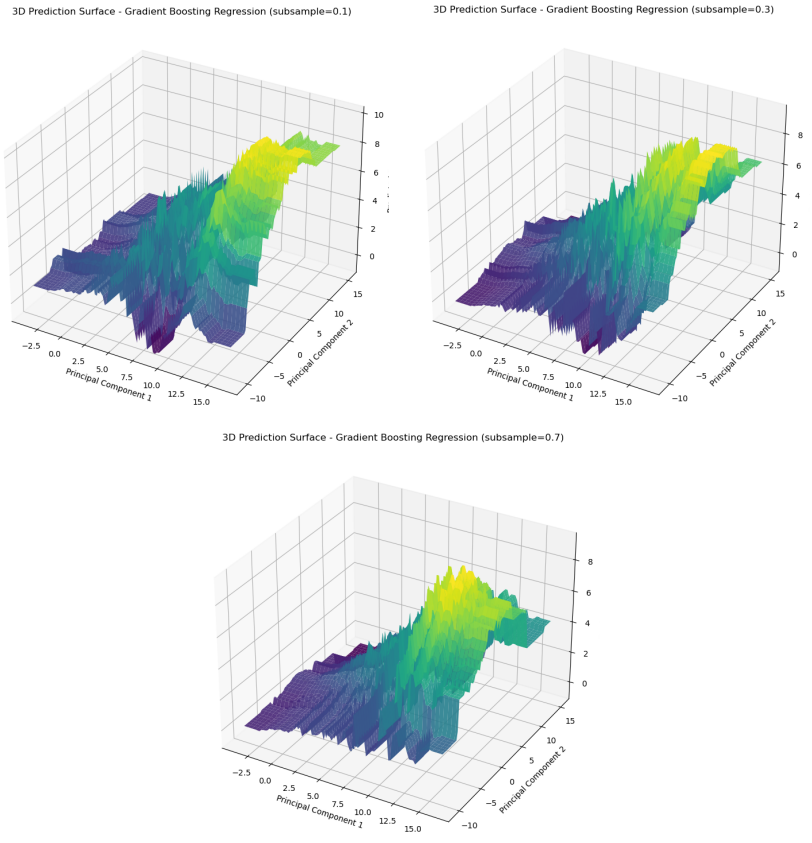


Figure 65: Gradient Boosting Regression with different values of subsample

Performance of Gradient Boosting Regression

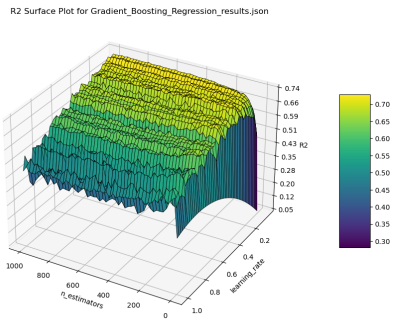


Figure 66: R2 Plot for Gradient Boosting Regression

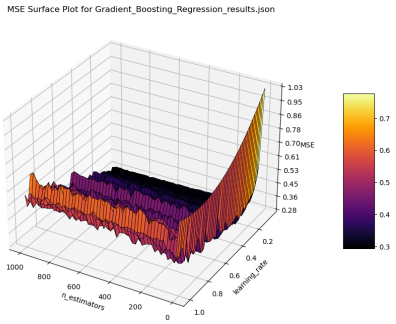


Figure 67: MSE Plot for Gradient Boosting Regression

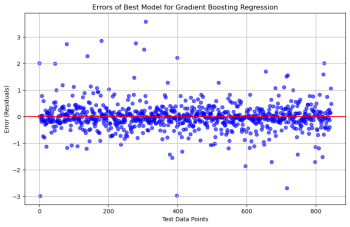


Figure 68: Errors for Gradient Boosting Regression

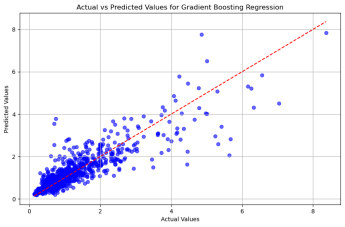


Figure 69: Actual vs Predicted for Gradient Boosting Regression

The parameter learning_rate is a clear factor in the performance of the model: lower values corresponds to higher accuracy. This is expected by the model; however, low learning rates values lead to an higher computational complexity. While the optimal value for learning_rate is pretty clear, the n_estimators parameter has a variegate and unclear distribution. While usually an higher number of estimators should lead to an improvement in performance, it could also lead to overfitting, and consequently to a lower accuracy on unseen test data.

Forward Subset Selection (FSS) Results for Gradient Boosting Regression

- Selected Features: ['O3', 'N_CPC', 'PM-2.5', 'TEMP', 'NO', 'CO', 'PM-1.0', 'SO2']
- Scores for each step: [0.18288847392335858, 0.300972829708265, 0.5453949933577633, 0.6154306298044471, 0.6373895409327599, 0.654065355807539, 0.6602011018778615, 0.666977643886846]

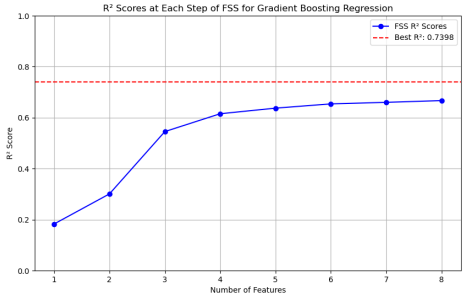


Figure 70: R2 Scores at Each Step of FSS for Gradient Boosting Regression

Even with eight features, the model still performs quite worse than the full model. Also, The gain from six to eight features is quite low and not worthy in most of the use cases. The selected features include some of the most correlated features, but curiously, it also include 'temperature' and 'SO3', the variables that have the lowest correlation with Black Carbon.

Shuffling Statistics for Gradient Boosting Regression

After shuffling the data and evaluating the model Gradient Boosting Regression, the following statistics were obtained:

- Original Score: 0.7397899587681255
- Mean Shuffled Score: 0.7329710242545938
- Standard Deviation of Shuffled Scores: 0.002913890534903964
- 95% Confidence Interval of Shuffled Scores: [0.7323899323956845, 0.7335521161135031]
- Average Difference: -0.00681893451353166

Shuffling the data led on average to slight loss in accuracy. Standard Deviation and Confidence Interval suggest that, although very low, the decrease in R^2 is statistically significant and expected.

Performance Metrics of Gradient Boosting Regression

- Parameters: {n_estimators: 898, learning_rate: 0.10816326530612246}
- Standard R^2: 0.7397899587681255
- FSS R^2: 0.666977643886846
- Shuffle Average R^2: 0.7329710242545938

Gradient Boosting Regression performs well with high scores in R^2. FSS led to a loss in performance; Shuffling led to a very low but expected decrease of the R^2 value.

3.10 Feed-Forward Neural Network

Feed-Forward Neural Network (FFNN) is a type of artificial neural network where connections between nodes do not form a cycle. It is used for regression by learning complex mappings from input features to continuous target values.

hidden_layer_sizes

The number of neurons in each hidden layer. Larger or more numerous layers allow the model to capture more complex patterns but can increase the risk of overfitting and require more computational resources.

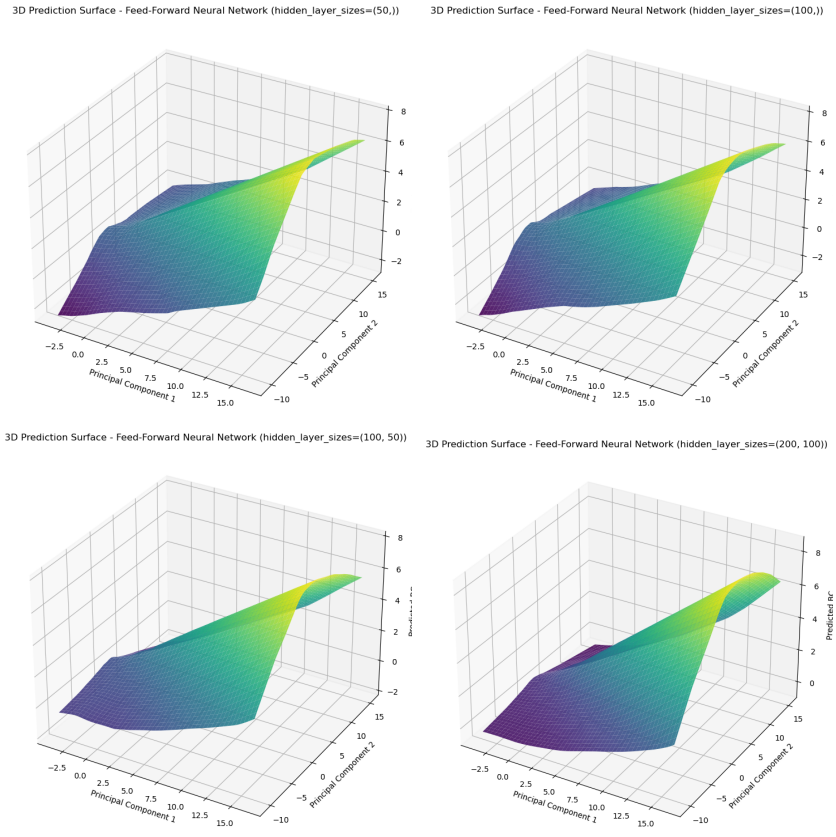


Figure 71: Feed-Forward Neural Network with different values of hidden_layer_sizes

activation

The activation function for the hidden layer. It introduces non-linearity into the model. Choosing the right activation function can significantly affect model performance:

- **ReLU**: Effective and widely used, it helps mitigate the vanishing gradient problem.
- **Tanh**: Useful for zero-centered data, mapping inputs to values between -1 and 1.

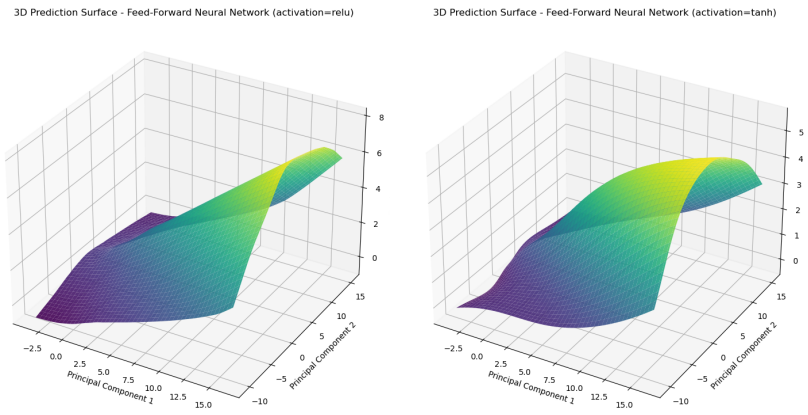


Figure 72: Feed-Forward Neural Network with different activation functions

solver

The algorithm for weight optimization. Different solvers can converge faster or handle large datasets differently:

- **adam**: Stochastic gradient-based optimizer, suitable for large datasets.
- **lbfgs**: Quasi-Newton method, can be more accurate for smaller datasets.

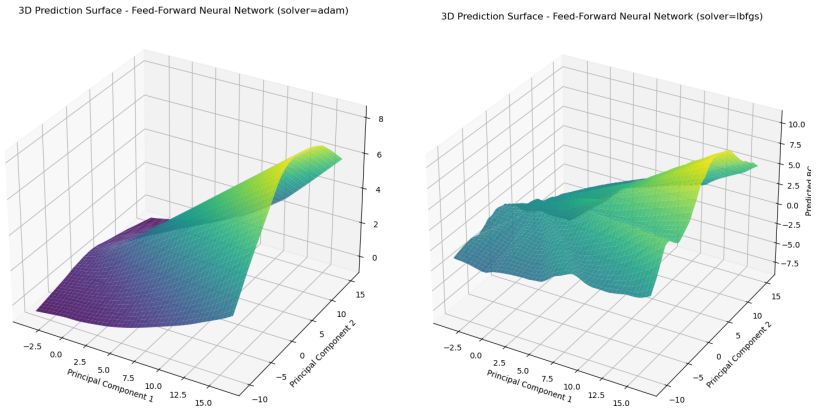


Figure 73: Feed-Forward Neural Network with different solvers

alpha

The regularization term that prevents overfitting by penalizing large weights. Increasing alpha can help regularize the model, improving generalization at the cost of potentially underfitting the training data.

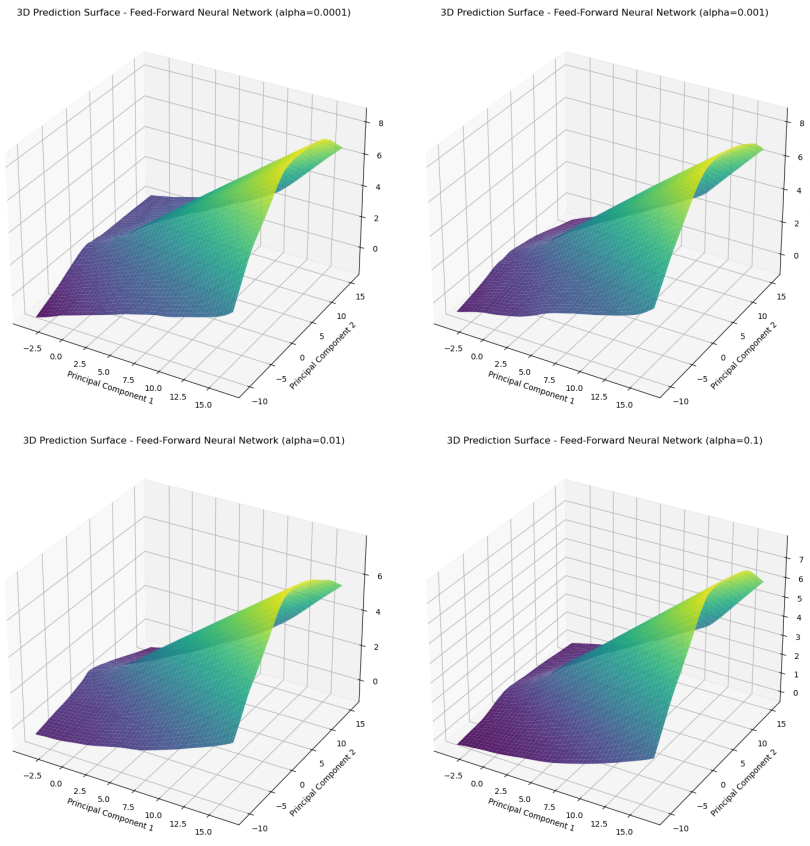


Figure 74: Feed-Forward Neural Network with different values of alpha

learning_rate

The learning rate schedule for weight updates. The choice of learning rate can significantly affect the convergence speed and final model performance:

- **constant**: A fixed learning rate throughout training, which can be simpler but may require tuning.
- **adaptive**: Adjusts the learning rate based on performance, potentially improving convergence and avoiding local minima.
- **invscaling**: Decreases the learning rate at each step, which can help in finding a more precise final set of weights by reducing the step size over time.

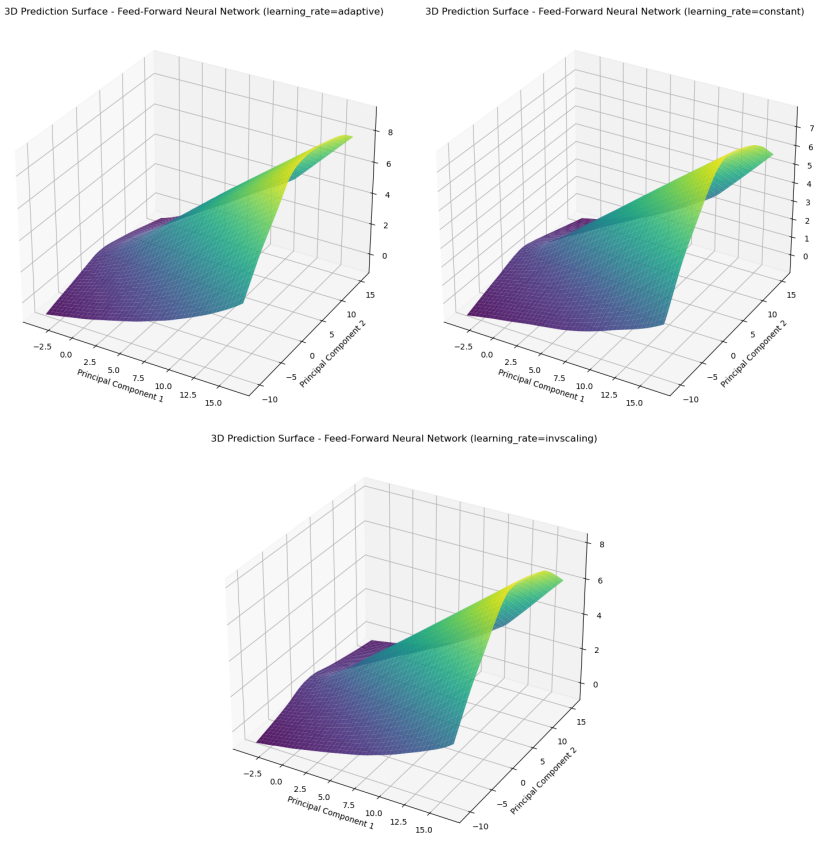


Figure 75: Feed-Forward Neural Network with different learning rates

Performance of Feed-Forward Neural Network

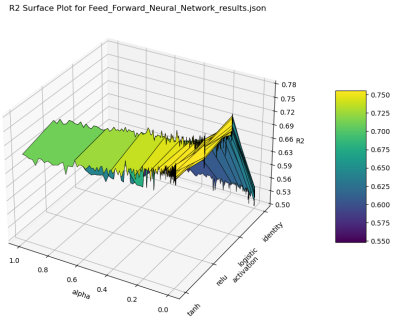


Figure 76: R2 Plot for Feed-Forward Neural Network

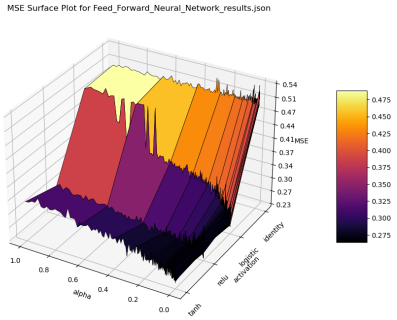


Figure 77: MSE Plot for Feed-Forward Neural Network

Lower values of alpha seem to perform slightly better than higher values; however, the activation parameter has the main role in impacting the performance of the FFNN model. The highest performances are observed for the activation functions 'tanh' and 'relu,' while 'logistic' only leads to moderate performances and 'identity' results in the worst performance among the four parameter values.

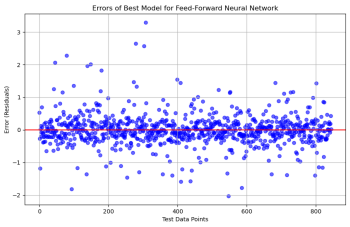


Figure 78: Errors for Feed-Forward Neural Network

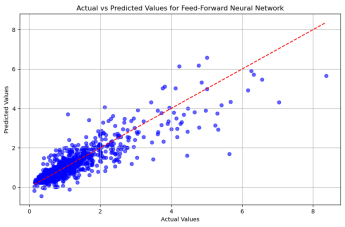


Figure 79: Actual vs Predicted for Feed-Forward Neural Network

Forward Subset Selection (FSS) Results for Feed-Forward Neural Network

- Selected Features: ['PM-2.5', 'N_CPC', 'O3', 'TEMP', 'PM-1.0', 'HUM']
- Scores for each step: [0.22532104493760813, 0.512891022793525, 0.6143776111016266, 0.6685132646714997, 0.6907884572097003, 0.7224365141410439]

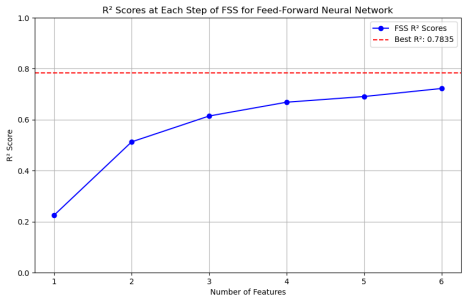


Figure 80: R2 Scores at Each Step of FSS for Feed-Forward Neural Network

Even with six different features, FSS fails to represent accurately the full model. It includes some of the most correlated features, but curiously, it also includes 'temperature', the variable that has the lowest correlation with Black Carbon.

Shuffling Statistics for Feed-Forward Neural Network

After shuffling the data and evaluating the model Feed-Forward Neural Network, the following statistics were obtained:

- Original Score: 0.7834519655180784
- Mean Shuffled Score: 0.7542279077993866
- Standard Deviation of Shuffled Scores: 0.022942456856208323
- 95% Confidence Interval of Shuffled Scores: [0.7517582212167717, 0.7566975943820016]
- Average Difference: -0.029224057718691743

Shuffling the data led on average to a decrease of 0.0292 in the R2 score. Standard Deviation and Confidence Interval suggest that, although moderate, the decrease is significant. An hypothesis is that the model was previously capturing temporal patterns, such as lower BC values during the weekend; the shuffling procedure might have destroyed these patterns, leading to a lower accuracy of the predictive model.

Performance Metrics of Feed-Forward Neural Network

- Parameters: {alpha: 4.928249570040513e-08, activation: 'relu'}
- Standard R^2: 0.7834519655180784
- FSS R^2: 0.7224365141410439
- Shuffle Average R^2: 0.7542279077993866

The Feed-Forward Neural Network provides high performance, with high R^2 values. This suggests that the model captures complex patterns in the data effectively, contributing to its superior predictive accuracy. FSS led to a significant loss in performance; Shuffling led to a moderate decrease in the R^2 value.

3.11 Comparison between models

Model	**Parameters**	**Standard R^2**	**FSS R^2**	**Shuffle R^2**
K-Nearest Neighbor	n_neighbors: 7, weights: distance	0.6662	0.6498	0.6662
Support Vector	C: 40.37, gamma: 0.032	0.7525	0.7045	0.7525
Elastic Net	alpha: 0.115, l1_ratio: 0.061	0.5439	0.5274	0.5439
AdaBoost	n_estimators: 21, learning_rate: 2.56e-10	0.5274	0.5062	0.4656
Kernel Ridge	alpha: 0.048, kernel: rbf	0.7456	0.6305	0.7456
Decision Tree	max_depth: 9, min_samples_split: 13	0.5374	0.5436	0.4147
Random Forest	n_estimators: 25, max_depth: 97	0.7177	0.6814	0.6698
Gaussian Process	alpha: 0.052, n_restarts_optimizer: 0	0.5867	0.5556	0.5867
Gradient Boosting	n_estimators: 898, learning_rate: 0.108	0.7398	0.6670	0.7329
FFNN	alpha: 4.93e-08, activation: relu	0.7835	0.7224	0.7542

Table 3: Performance Metrics of Different Models

In the top performers, several models performed similarly, with quite high accuracy. The Feed-Forward Neural Network (FFNN) achieved the highest R^2 score of 0.7835, followed by Gradient Boosting, Random Forest, Support Vector Regression (SVR) and Kernel Ridge Regression, all showing strong performance with R^2 scores above 0.7.
K-Nearest Neighbor had a moderate performance, with $R^2 = 0.6662$.
Several models perform poorly: Gaussian Process, Elastic Net Regression, AdaBoost, and Decision Tree Regression had R^2 scores lower than 0.6, indicating less accuracy in predicting BC concentrations.

4 Conclusions

Black Carbon (BC) concentrations are positively correlated with N_CPC, PM-2.5, NO_X and other air pollutants, signaling emissions from similar sources and combustion processes.
A significant negative correlation with O_3 is also found, indicating the negative impact that BC sunlight absorption has on ozone formation.
Temporal patterns emerged, with air pollutant concentrations being higher during winter due to residential heating and industrial activities, and lower during weekends due to reduced traffic and less intense human activities.

Strong correlations between BC sensor readings and data from more affordable sensors enable the use of machine learning techniques to estimate BC concentrations without relying on costly sensors. The performances of ten different models have been analyzed, testing their accuracy in predicting BC values.

Support Vector, Kernel Ridge, Random Forest, and Gradient Boosting all achieved $R^2 > 0.7$, with the Feed-Forward Neural Network (FFNN) obtaining the highest $R^2 = 0.784$.
While several models performed well, FFNN has the highest precision and also the ability to capture time series relationships. The non-robustness of a model might not look appealing; however, in the context of Black Carbon measurements, where some temporal pattern is observed, the sensitivity to data order can improve the accuracy of the predictions.

Elastic Net, the only linear model in the analysis, initially performed poorly. However, after expanding its feature set to include **polynomial** terms, the model's accuracy improved drastically, suggesting non-linear relationships in the data.

Interestingly, almost all the models include 'temperature' in their FSS subset; while 'temperature' has the lowest correlation with BC, it has significant correlations with other air pollutants and might serve as a good proxy for representing some of them.

Shuffling has no impact at all on most models; however, for some models, the order of the data can influence the learning process. Shuffling the data destroys the time patterns, significantly worsening the performances of FFNN, Random Forest, AdaBoost, and in particular, Decision Trees.

In conclusion, machine learning techniques demonstrated promising potential to estimate BC concentrations using data from more affordable sensors. Where extreme accuracy is not needed, this approach can significantly reduce the costs and operational challenges associated with traditional monitoring systems, supporting scientific understanding of Black Carbon's environmental impact.

A Data Analysis

A.1 Seasonal Trends

Table 4: ANOVA Results for Pollutants by Season

Pollutant	Source	Df	Sum Sq	Mean Sq	F value	Pr
BC	season	3	47	15.691	12.56	3.58e-08 ***
	Residuals	4219	5271	1.249		
N_CPC	season	3	775	258.22	3.814	0.00962 **
	Residuals	4219	285650	67.71		
PM-10	season	3	11463	3821	19.05	2.89e-12 ***
	Residuals	4219	846374	201		
PM-1.0	season	3	7770	2590.0	91.83	<2e-16 ***
	Residuals	4219	118990	28.2		
O3	season	3	179979	59993	93.85	<2e-16 ***
	Residuals	4219	2697094	639		
CO	season	3	8.26	2.7546	64.17	<2e-16 ***
	Residuals	4219	181.10	0.0429		
NOX	season	3	89293	29764	14.58	1.91e-09 ***
	Residuals	4219	8613242	2042		

Table 5: Post-hoc Test Results for BC (Black Carbon)

Comparison	Difference	Lower Bound	Upper Bound	p-value
Spring-Fall	-0.1462	-0.2822	-0.0102	0.0293
Summer-Fall	0.0646	-0.0632	0.1923	0.5633
Winter-Fall	0.1526	0.0361	0.2691	0.0043
Summer-Spring	0.2108	0.0721	0.3495	0.0006
Winter-Spring	0.2988	0.1704	0.4272	0.0000 ***
Winter-Summer	0.0880	-0.0317	0.2077	0.2326

Table 6: Post-hoc Test Results for N_CPC

Comparison	Difference	Lower Bound	Upper Bound	p-value
Spring-Fall	-0.5397	-1.5407	0.4613	0.5083
Summer-Fall	0.7366	-0.2038	1.6770	0.1832
Winter-Fall	0.3511	-0.5067	1.2088	0.7188
Summer-Spring	1.2763	0.2553	2.2973	0.0072
Winter-Spring	0.8908	-0.0547	1.8362	0.0732
Winter-Summer	-0.3855	-1.2666	0.4956	0.6744

Table 7: Post-hoc Test Results for PM-10

Comparison	Difference	Lower Bound	Upper Bound	p-value
Spring-Fall	3.3638	1.6408	5.0867	0.0000 ***
Summer-Fall	3.0158	1.3971	4.6346	0.0000 ***
Winter-Fall	4.2225	2.7460	5.6990	0.0000 ***
Summer-Spring	-0.3480	-2.1055	1.4096	0.9570
Winter-Spring	0.8587	-0.7687	2.4862	0.5271
Winter-Summer	1.2067	-0.3099	2.7233	0.1718

Table 8: Post-hoc Test Results for PM-1.0

Comparison	Difference	Lower Bound	Upper Bound	p-value
Spring-Fall	0.7007	0.0547	1.3468	0.0273
Summer-Fall	1.2923	0.6853	1.8992	0.0000 ***
Winter-Fall	3.3473	2.7937	3.9009	0.0000 ***
Summer-Spring	0.5915	-0.0675	1.2505	0.0966
Winter-Spring	2.6466	2.0364	3.2568	0.0000 ***
Winter-Summer	2.0550	1.4864	2.6237	0.0000 ***

Table 9: Post-hoc Test Results for O3 (Ozone)

Comparison	Difference	Lower Bound	Upper Bound	p-value
Spring-Fall	9.2342	6.1585	12.3100	0.0000 ***
Summer-Fall	10.7308	7.8411	13.6204	0.0000 ***
Winter-Fall	-4.6526	-7.2882	-2.0169	0.0000 ***
Summer-Spring	1.4965	-1.6409	4.6339	0.6104
Winter-Spring	-13.8868	-16.7920	-10.9817	0.0000 ***
Winter-Summer	-15.3833	-18.0907	-12.6760	0.0000 ***

Table 10: Post-hoc Test Results for CO (Carbon Monoxide)

Comparison	Difference	Lower Bound	Upper Bound	p-value
Spring-Fall	0.0175	-0.0077	0.0427	0.2809
Summer-Fall	-0.0296	-0.0533	-0.0059	0.0073
Winter-Fall	0.0823	0.0607	0.1039	0.0000 ***
Summer-Spring	-0.0471	-0.0728	-0.0214	0.0000 ***
Winter-Spring	0.0648	0.0410	0.0886	0.0000 ***
Winter-Summer	0.1119	0.0897	0.1341	0.0000 ***

Table 11: Post-hoc Test Results for NOX

Comparison	Difference	Lower Bound	Upper Bound	p-value
Spring-Fall	-1.2557	-6.7522	4.2408	0.9360
Summer-Fall	-0.1477	-5.3117	5.0162	0.9999
Winter-Fall	9.2845	4.5745	13.9945	0.0000 ***
Summer-Spring	1.1080	-4.4987	6.7147	0.9572
Winter-Spring	10.5402	5.3486	15.7318	0.0000 ***
Winter-Summer	9.4322	4.5940	14.2704	0.0000 ***

A.2 Daily Trends

Table 12: ANOVA Results for Pollutants by Day of the Week

Pollutant	Source	Df	Sum Sq	Mean Sq	F value	Pr
BC	day_of_week	6	34	5.719	4.563	0.000127 ***
	Residuals	4216	5283	1.253		
NOX	day_of_week	6	82918	13820	6.759	3.83e-07 ***
	Residuals	4216	8619617	2045		
N_CPC	day_of_week	6	3876	646.1	9.64	1.46e-10 ***
	Residuals	4216	282549	67.0		
CO	day_of_week	6	1.56	0.26055	5.849	4.37e-06 ***
	Residuals	4216	187.80	0.04454		

Table 13: Post-hoc Test Results for BC

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Monday-Friday	-0.144	-0.335	0.047	0.281
Saturday-Friday	-0.191	-0.386	0.005	0.061
Sunday-Friday	-0.266	-0.458	-0.075	0.001
Thursday-Friday	-0.018	-0.211	0.176	1.000
Tuesday-Friday	-0.153	-0.346	0.039	0.222
Wednesday-Friday	-0.056	-0.250	0.138	0.980
Saturday-Monday	-0.047	-0.236	0.143	0.991
Sunday-Monday	-0.122	-0.307	0.063	0.449
Thursday-Monday	0.126	-0.061	0.313	0.422
Tuesday-Monday	-0.009	-0.196	0.177	1.000
Wednesday-Monday	0.088	-0.100	0.276	0.813
Sunday-Saturday	-0.075	-0.265	0.115	0.905
Thursday-Saturday	0.173	-0.019	0.365	0.109
Tuesday-Saturday	0.038	-0.154	0.229	0.997
Wednesday-Saturday	0.135	-0.058	0.328	0.377
Thursday-Sunday	0.248	0.061	0.436	0.002
Tuesday-Sunday	0.113	-0.074	0.300	0.563
Wednesday-Sunday	0.210	0.021	0.399	0.018
Tuesday-Thursday	-0.136	-0.325	0.054	0.345
Wednesday-Thursday	-0.038	-0.229	0.153	0.997
Wednesday-Tuesday	0.097	-0.093	0.288	0.740

Table 14: Post-hoc Test Results for NOX

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Monday-Friday	-7.129	-14.836	0.578	0.091
Saturday-Friday	-7.240	-15.138	0.657	0.097
Sunday-Friday	-13.219	-20.948	-5.489	0.000
Thursday-Friday	-1.339	-9.149	6.471	0.999
Tuesday-Friday	-4.439	-12.224	3.347	0.628
Wednesday-Friday	-0.487	-8.339	7.365	1.000
Saturday-Monday	-0.111	-7.765	7.542	1.000
Sunday-Monday	-6.090	-13.569	1.390	0.198
Thursday-Monday	5.790	-1.773	13.353	0.265
Tuesday-Monday	2.690	-4.847	10.228	0.941
Wednesday-Monday	6.642	-0.964	14.248	0.134
Sunday-Saturday	-5.978	-13.655	1.698	0.245
Thursday-Saturday	5.901	-1.856	13.659	0.272
Tuesday-Saturday	2.802	-4.931	10.535	0.937
Wednesday-Saturday	6.753	-1.046	14.553	0.141
Thursday-Sunday	11.880	4.294	19.466	0.000
Tuesday-Sunday	8.780	1.219	16.341	0.011
Wednesday-Sunday	12.732	5.103	20.361	0.000
Tuesday-Thursday	-3.100	-10.743	4.544	0.896
Wednesday-Thursday	0.852	-6.859	8.563	0.999
Wednesday-Tuesday	3.952	-3.734	11.637	0.735

Table 15: Post-hoc Test Results for N_CPC (Ultrafine Particle Number Concentration)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Monday-Friday	-1.258	-2.653	0.137	0.1089
Saturday-Friday	-1.452	-2.882	-0.022	0.0438
Sunday-Friday	-1.499	-2.899	-0.100	0.0266
Thursday-Friday	0.561	-0.853	1.975	0.9055
Tuesday-Friday	-0.272	-1.682	1.137	0.9976
Wednesday-Friday	1.107	-0.315	2.529	0.2455
Saturday-Monday	-0.194	-1.580	1.192	0.9996
Sunday-Monday	-0.241	-1.595	1.113	0.9985
Thursday-Monday	1.819	0.450	3.188	0.0018
Tuesday-Monday	0.986	-0.379	2.351	0.3343
Wednesday-Monday	2.365	0.988	3.742	0.0000088
Sunday-Saturday	-0.047	-1.437	1.343	1.0000
Thursday-Saturday	2.013	0.608	3.417	0.0004818
Tuesday-Saturday	1.180	-0.220	2.580	0.1645
Wednesday-Saturday	2.559	1.147	3.971	0.0000020
Thursday-Sunday	2.060	0.686	3.433	0.0002009
Tuesday-Sunday	1.227	-0.142	2.596	0.1133
Wednesday-Sunday	2.606	1.225	3.987	0.0000006
Tuesday-Thursday	-0.833	-2.217	0.551	0.5649
Wednesday-Thursday	0.546	-0.850	1.942	0.9110
Wednesday-Tuesday	1.379	-0.012	2.771	0.0539

Table 16: Post-hoc Test Results for CO (Carbon Monoxide)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Monday-Friday	-0.007	-0.043	0.029	0.9968
Saturday-Friday	-0.011	-0.047	0.026	0.9801
Sunday-Friday	-0.008	-0.044	0.028	0.9942
Thursday-Friday	0.031	-0.006	0.067	0.1635
Tuesday-Friday	-0.003	-0.039	0.034	0.9999
Wednesday-Friday	0.041	0.004	0.078	0.0167
Saturday-Monday	-0.003	-0.039	0.032	0.99997
Sunday-Monday	-0.001	-0.036	0.034	1.0000
Thursday-Monday	0.038	0.003	0.073	0.0248
Tuesday-Monday	0.005	-0.031	0.040	0.9997
Wednesday-Monday	0.048	0.013	0.084	0.0012
Sunday-Saturday	0.002	-0.033	0.038	0.99999
Thursday-Saturday	0.041	0.005	0.078	0.0136
Tuesday-Saturday	0.008	-0.028	0.044	0.9952
Wednesday-Saturday	0.052	0.015	0.088	0.0006
Thursday-Sunday	0.039	0.004	0.074	0.0203
Tuesday-Sunday	0.006	-0.030	0.041	0.9992
Wednesday-Sunday	0.049	0.014	0.085	0.0009
Tuesday-Thursday	-0.033	-0.069	0.002	0.0841
Wednesday-Thursday	0.010	-0.026	0.046	0.9804
Wednesday-Tuesday	0.044	0.008	0.080	0.0061