D682.3
Interpreting the Given Data

Data Exploration and Analysis:

The given dataset consists of 35 columns of data and 1001 rows (1 of which is headers). Within this table there are a small number of key variables that are most pertinent to solving the given problem. Three key sections for interpreting this data are the Pollution category (Columns H, I and J), “Severity Score” (column AE), and “Month” (column AF).

Both the Pollution category and “Severity Score” are closely correlated to the Health Risk Score (HRS) column, making them key values for predicting a proper estimate for the HRS for a row. Notably, these two columns share an almost parallel regression line with the HRS when the data is ordered; making them most valuable to the estimate calculation.

However, with the “Month” category, it is made observable that all of the given data is retrieved from the same month: September. This may introduce a bias into the ML model if not addressed, and is further important to note that this shows the data is undiverse, which may hinder the effectiveness of this model if used to forecast values for other months (because all of the current data is collected from only one month).

Besides these three sections, there are another set of trends that are crucial for interpreting the given data:

1. The data, when formatted, shows most columns sharing a similar slope
2. Atmospheric conditions are the most correlated to HRS

The first of these observations is observable through the slopes collected by the cLine function within the current solution. When analyzing these values it is clear to see that most of the lines within the dataset, albeit not all, are closely aligned. When these values are then reduced by their current y-intercept and given the bias of the HRS column regression line, the data is seen to be similar.

Secondly, the atmospheric conditions are seen to be the most correlated to the HRS line when the above observation is focused on the specific columns. This is likely due to the nature of the atmosphere being a direct connection to the HRS, while the other variables and columns, such as those regarding data and time, are more distant (though still intertwined by merit of them also being correlated to the atmospheric data themselves separately).

Consequently to the two above points and the overall analysis that has been conducted, the following hypothesis has emerged:

Hypothesis: *The information provided in this dataset is interconnected.*

This is supported by the above realized trend that most columns share a similar slope to the HRS column, and is furthered by how even the other columns in the data that are not as similar still share a resemblance, to some degree, to the atmospheric columns that are provided. This is because the data and time columns, while they are very much separate values than the other columns, share a certain likeness to the atmospheric data. One such instance is with how, on weekends, the air pollution is noted to regularly change. This reveals that, even though the named month or day may not directly otherwise connect to the HRS, the patterns notable in the other columns cast an influencing effect onto them: thus intertwining the information provided in the dataset.

Interpretation of Model Outputs:

As aforementioned in the D682.2 report, the evaluation metrics of Mean Absolute Percentage Error (MAPE), Average Error, Coefficient of Determination (R^2), and Root Mean Squared Error (RMSE) all provide distinct insights into the different levels of effectiveness that the model can predict an accurate estimate for the HRS. Using these key metrics, the optimized ML model is proven to make impactful changes to the solutions accuracy.

The Average Error score is perhaps the simplest and most interpretable metric used in analyzing the models outputs. This value informs the user of the standard amount of error per estimate: revealing what deviance from the actual HRS can be expected. Not only is this paramount for the user, but it also reveals the accuracy increase garnered by the optimization of the solutions ML model. As noted by the software outputs for Average Error before and after the optimization (as exemplified in D682.3), there is a clear reduction of the error value, and hence it is easily observable that the accuracy of the model has been increased.

Likewise, the MAPE revealed a very similar value that changes proportionally to the Average Error, but it provides a deeper analysis of the data by pressing the deviance as a percentage of the total expected value. Its findings and inclinations are equitable to the Average Error outcomes.

The RMSE is a metric that also is dependent on the error of a certain estimate, but makes the larger errors more pronounced. With this, the variance of estimates is seen to be smaller than the original, unoptimized system as this value is lower.

Where the current model still has space for improvement is made noticeable through the Coefficient of Determination value (R^2). Before the optimization, the value is seen to be about ~.4, and after it only increases by one degree to ~.5. This shows that the R^2 variance that is being measured is not heavily relegated by the current implementation of the solution.

Together, the notable changes in these metrics after the model has been optimized shows that the learning techniques utilized do increase the accuracy of the solution outcomes. Thus, the future processes of the program can be presumed reliable (within the established degree).

This reality of the solutions impact on future estimations is pertinent in the use of this allocation. As the understood error assists the user in weighing the accuracy of any certain value, this model will provide users with a consistent approximation of HRS and Air Quality scores. Further, this model can be used to forecast health situations and, if the data is within a deviation of a worrying risk factor, the users can take note of this and declare an Air Quality/HRS warning in caution of one being possible. And, by taking note of what certain categories are being weighted more heavily than others, the end user can observe these key types of data and regard or further research these data to better address the health and air quality factors in their jurisdiction.

The importance of features in the optimized program can be made observable by their influence on the end product. Namely, the HRS before boosting is considerably lower, and within the boosting algorithm it notes the contribution of different optimizations: the most prominent being Weighted Averaging and Pruning. With these two components nested in the greater boosting function, catalyze the most change in the estimated value, and thus deliver the greatest contributions to the optimized model.