D682.2
Regression-Based AI Optimization

Optimizations

In order to optimize the regression machine learning model developed for Air Quality and Health Risk Score prediction, the following techniques were deployed: quantization, pruning, L1 and L2 regularization, weighted averaging,and boosting.

Outcomes:

Before Model Optimizations:

- *Average Error:* **6.7068**
- *Mean Absolute Percentage Error:* **68.2 %**
- *R-Squared:* **0.4654**
- *RMS Error:* **71.42**

Optimized Model:

- *Average Error:* **0.6599**
- *Mean Absolute Percentage Error:* **6.86 %**
- *R-Squared:* **0.5003**
- *RMS Error:* **0.7**

Outcome Analysis:

To further understand and analyze the outputs of the current model, additional evaluation metrics were added: the Coefficient of Determination (R-Squared) and RMSE (Root Mean Squared Error). These two lenses assist in observing the relation of estimated values to a central line and can also aid in showing the magnitude of the deviations. For R-Squared, the variance of data is what is primarily revealed, while with RMSE the mean error between the estimate and actual data is shown. Thus, by including these two metrics we are further able to identify and categorize estimate accuracy.

Before the optimization, the estimated values for each HRS are seen to have a large error, including a MAPE (Mean Absolute Percentage Error) of 68.2%. After the use of weighted averaging, pruning, L1 and L2 regularization, quantization and boosting, the now trained model is observed to have reached lower error metrics of Average Error, RMSE, and MAPE (now with a score of 6.86%). Thus, the optimization techniques utilized are seen to provide a more accurate prediction proven by the lesser error margins.

And while the impact from the current optimizations is measurable, the performance impact is not negligible. The complexity of the project, both time and space, rises exponentially with each iteration. This complexity, in the current program and with the provided data, is manageable; however, if the model were to be better optimized within itself to product more accurate values, or if the source data to train the model were to increase from 1000 rows to say even 10000, the model could soon become slow to process, and cost valuable time and computational resources.

Optimization Techniques:

Optimizations:

Performance Techniques:

To have values be more manageable and calculations less resource intensive, the strategy of *Pruning* and *Quantization* were used in the optimized version of the ML model. Pruning consists of taking out certain data that is not crucial to overall calculation, and this technique was implemented to trim out columns with small weights for the eventual weighted averaging process. Quantization involves making values shorter: this technique was used for most all float values within the program being rounded to smaller numbers and thus saving computation resources.

Regularization Techniques:

For regularizing the data being used to make the estimations within this model, *L1* and *L2* regularization techniques are used. The L1 (Lasso) algorithm adds a weight penalty to manipulate the data to be more generalized. Similarly, the L2 (Ridge) method also uses weight penalties, but adds a square of the coefficients: leading to more distributed values and greater generalization. Together these techniques allow for regression line processing that is less prone to overfitting of data.

Ensemble Techniques:

In order to have different facets of the data affect the final estimate more optimally, the use of two ensemble techniques, *Weighted Averaging* and *Boosting*, were deployed for this solution. With Weighted Averaging, columns of data that had closer regression lines to the Health Risk Score column are given more weight, allowing for their impact to be more in the calculation of the final estimate. Then, with the use of Boosting, hyperparameters are given different ways to route a more accurate path to a final estimate. These two techniques allow for both individual columns and also hyperparameters to influence the estimation algorithm than they otherwise are able to, thus providing a better answer to the situations given problem.