Yoseph Kebede

ENPM808Y

HW2

02/14/2023

Report on Implementation of regression and classification algorithms

A manufacturing industry uses a certain machine that has been monitored for failure in various scenarios, ie 10,000 instances for the purpose of this dataset. Thus, in order to identify the likely causes of the machine failure, these instances have been categorized in to input features that can be grouped into environmental state, machine performance, and failure type conditions which all attribute one way or another to machine failure. Moreover, the machine can fail due to anyone of the failure types, which are tool wear failure (TWF), heat dissipation failure (HDF), power failure (power failure), overstrain failure (OSF), and random failures (RNF). The challenge here is to identify which failure modes likely caused the machine to fail when more than one failure mode is ON at any given moment, and the correlation of the environment states as well as the machine performance to the failure modes.

Before any visualization work on the data, some pre-processing actions were performed where the data types, presence of missing data points, as well as data duplication were checked. Then, nonnumeric columns not needed for analysis such as *Product ID* and *UID* were removed while others such as *Type* and failure modes were converted to numeric.

First, the correlation matrix was computed and plotted in a heatmap so that the normalized relationship of the features to each other is visualized, as well as by illustrating scatter plots where each of the numerically identified columns are plotted individually with respect to the remaining variables. Thus, this section answers the first question on how the parameters relate to each other. Then, after standardizing the data by scaling it using the mean and standard deviation, it was then split into the testing set and training set in preparation for the modeling with the features transformed (scaled) using the mean and standard deviation of the feature data.

The trained dataset was then tested using linear regression, logistic regression and naïve bayes models. First, the linear regression model showed that the trained and tested data had a mean squared error of 0.44, but an r2 score of 0.001 showing that the variance of the model and total variance of data are off. This however can be improved by tuning the data for better performance.

The logistic regression model on the other hand showed that the training accuracy was 59.5% and the model accuracy score was 61.3% , and a precision score of 61%, where 100% being trained and test data correlating to be identical.

Finally, the naïve bayes model had a lower model accuracy than the trained accuracy, ie. 30.4% to 31.9% respectively, and precision was 68% for the type L, 29% for M, and 5% for H.

In conclusion, although still better performance data tuning can be done, results showed that the logical regression model yielded results with better correlation than the other models. I have, thus, learnt a bit about what it takes to setup a data before model implementation and how to display processed results as well as managing larger datasets to understand how target variables are impacted

by features chosen for analysis. I would still like to learn more on how, for example, the machine failure is determined from the other failure modes. Hence, I hope to gain more experience as the course progresses, so that I can perform startup analysis if given any dataset and asked to find correlation for a target feature within the data.