ENPM808Y – Homework 2

Sandeep Thalapanane
UID - 119402535

Introduction: To make predictions of unseen data by training and testing the model using classification and regression algorithms and compare the results of those algorithms and evaluate them. The dataset used in this problem is synthetic data with 14 features which reflects AI4I 2020 Predictive Maintenance Dataset in the industry. The features are UID, product ID, Air temperature [K], Process temperature [K], Rotational speed [rpm], Torque [Nm], and Tool wear [min]. The output feature is 'machine failure' which consists of five independent failure modes - Tool wear failure (TWF), Heat dissipation failure (HDF), Power failure (PWF), Overstrain failure (OSF), and Random failures (RNF). The three algorithms used to solve this problem are Linear regression, Logistic regression, and Naive Bayes. Linear regression predicts a numerical output whereas logistic and Naive Bayes classify the data.

EDA: In the exploratory data analysis, we check if there are any unassigned values (NAN) and other anomalies. The duplicate values are checked using the pandas duplicated.sum() function. Missing values are checked using the pandas isnull.sum() function. The first letter of the product ID, which is an alphabet, is removed and the numerical part of the product ID is replaced in the same column. The 'Type' column has only three values which are low, medium, and high, and are converted to integers 0, 1, and 2 using label encoding. The correlation matrix shows the correlation coefficient between two variables. In EDA, we learn about the data.

Model Implementation and evaluation: The data is split into two categories in which 70 percent of the data is used to train the model and the rest 30 percent of data is used to evaluate the model in all three algorithms.

1) Linear Regression – It is used to estimate the relationship between multiple independent variables and one dependent variable i.e., the 'Type' column. The equation for the multiple linear regression is $y = \beta_0 + \beta_1 x + \beta_2 x + \cdots .. + \beta_n x$, where y is the dependent variable, β_0 is the intercept and $\beta_1, \beta_2, \ldots . \beta_n$ are the coefficients of the independent variables. Evaluation of the linear regression model is done using mean squared error and R^2 score.

2) Logistic Regression – This model produces three separate probability estimates for three classes and selects the highest probability as the final prediction for each case in multinomial regression. The sigmoid function is used to generate predictions. Training accuracy and model accuracy scores are used to evaluate this model along with the classification report & confusion matrix.

3) Naive Bayes – This model assumes all the features to be conditionally independent and classifies the data using the Bayes theorem. This model is evaluated using training accuracy, model accuracy scores, classification report, and confusion matrix.

Conclusion: In the Linear regression model, the R^2score value should be close to 1 to be considered relatively strong. R^2score of the linear model is near zero, which is considered to be a bad predictor and it has a low mean squared error. The training accuracy and model accuracy scores of logistic regression are higher than the Naive Bayes algorithm, with both scores near 60 percent. Among the three, probably logistic regression model might be the best fit for the data and to make predictions.