MACHINE FAILURE PREDICTION

(Communication Tools Assignment)

BY - NITHESH RAMANNA

Table of Contents

BUSINESS PROBLEM

DESCRIPTION

Currently, a car manufacturing company is not using any predictive models to predict the probability of the machine failure and it is fully relied on failure incidents reported by its employees. To avoid interrupts in production line for several hour due to the machine failure it is important for the company to know in advance whether a machine will fail or not.

GOAL OF THE PROJECT

A new predictive model for the company which predicts the machine failure in advance. The model to be more efficient to avoid the production interrupts and the model should be easy to interpret.

DATA EXPLORATION

DATA INFORMATION

One dataset with 10,000 observations has been provided. Out of 10,000 observations there are 339 incidents of machine failure. There are 9 features, out of 9, 2 are the ID features, one is the categorical feature, and one is the target variable.

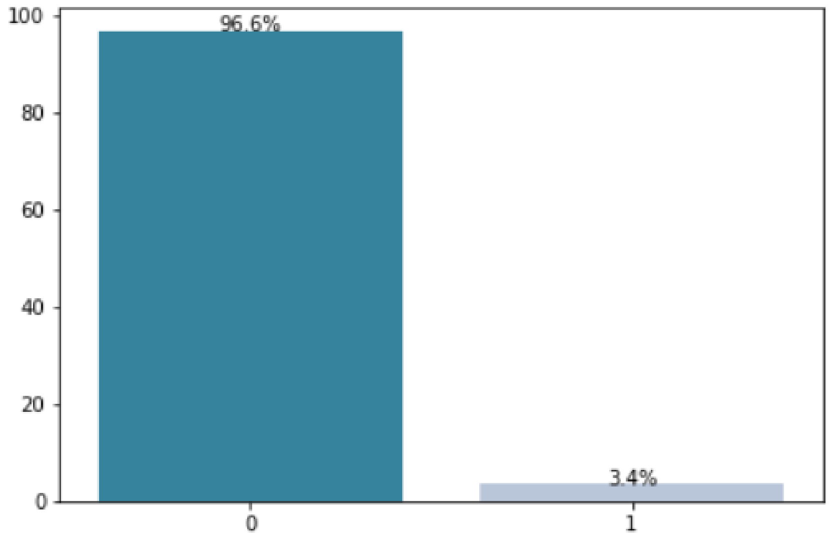

Figure 1 - Percentage of Machines failed and not failed, 0-Not failed, and 1-failed

PAIR PLOT

Figure 2 shows the pair plots. Looking at the figure we can easily say that the features, Process Temperature[K] and Air Temperature[K] are positively related to each other and the features, Torque [NM] and Rotational Speed [rpm] are inversely related to each other.

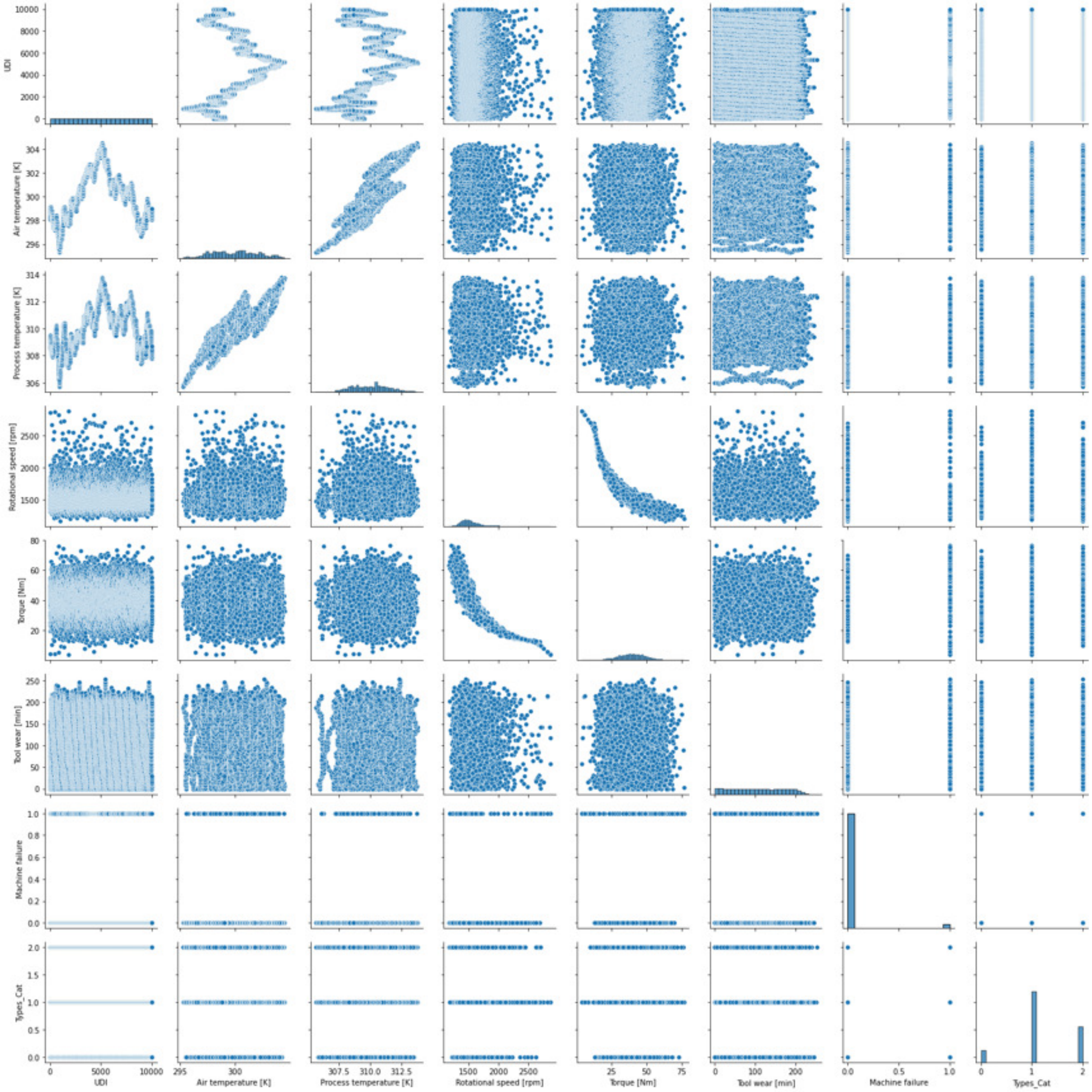

Figure 2 - Pair Plot

CORRELATION BETWEEN THE VARIABLES

The correlation values between the variables will give us the idea of how strongly the variables are correlated, either positively or negatively. Figure 3, like figure 2 provides the information of how the variables are correlated, addition to that the later gives the correlation values for the given pair of variables.

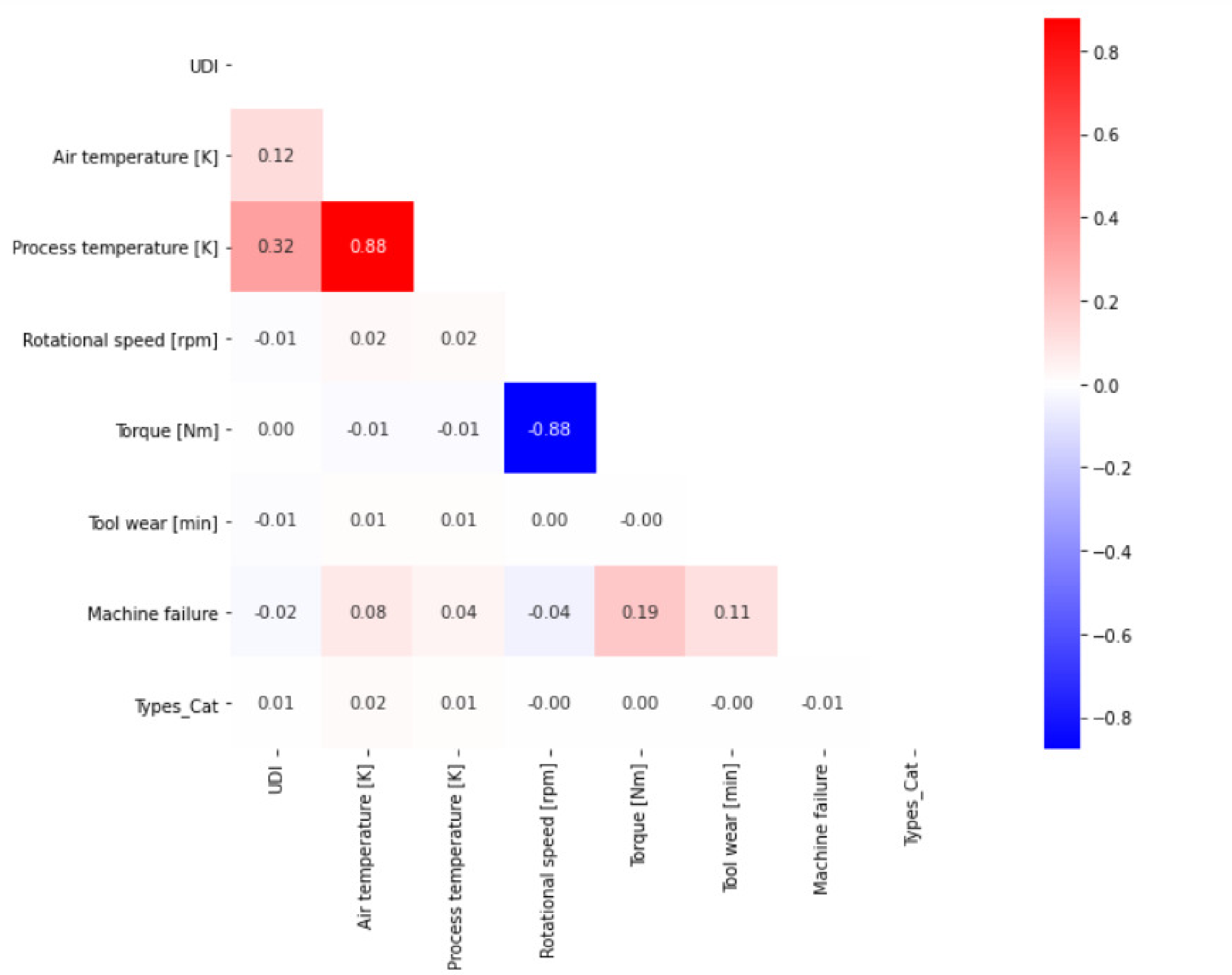

Figure 3 - Correlation between the variables

Looking at the above figure, Process Temperature[K] and Air Temperature[K] have a strong positive correlation, Torque [NM] and Rotational Speed [rpm] have a strong negative correlation. Rest of the variables do not have or have least correlations with other features.

RELATIONSHIP BETWEEN MACHINE FAILURE AND OTHER VARIABLES

Since we have few features, let's see the relationship between these features and the Machine failure. This may give us the fair idea which features are playing the role in Machine failures.

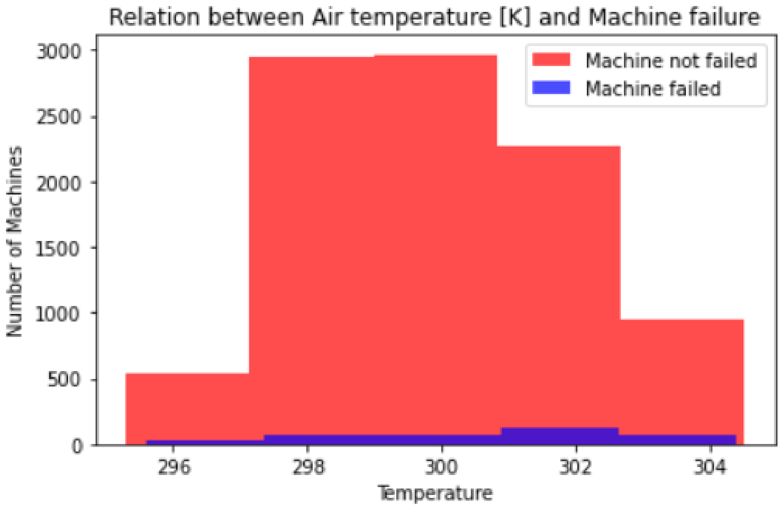

Figure 4 - Air temperature[K] VS Number of Machines

By looking at the above figure 4, apparently, machines failures have been occurred in all the Air temperatures.

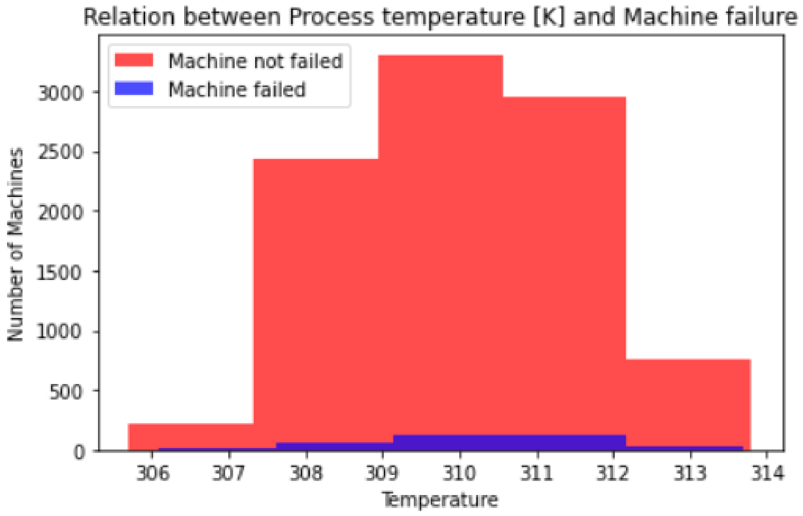

Figure 5 - Process temperature VS Number of Machines

Like relation between Air temperature and Machine failure, relation between Process temperature and Machine failure apparently doesn't gives any direct insights.

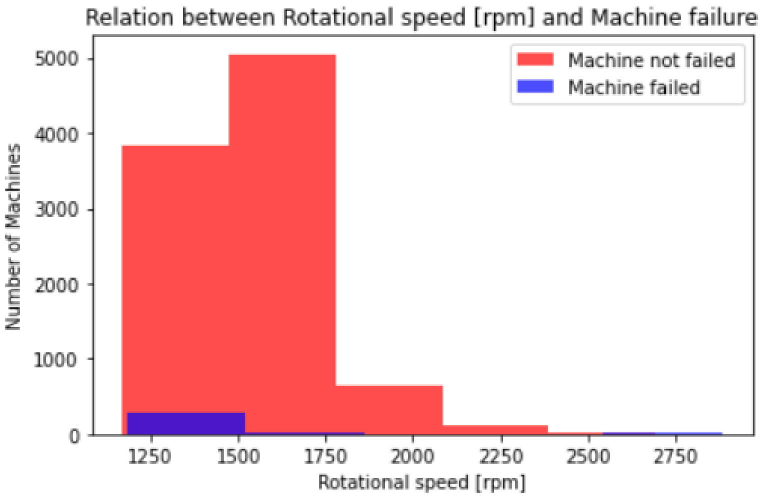

Figure 6 - Rotational speed[rpm] VS Number of Machines

Relation between Rotational speed [rpm] and Machine failure, we can clearly see from figure 6 most of the machines with higher RPM have failed.

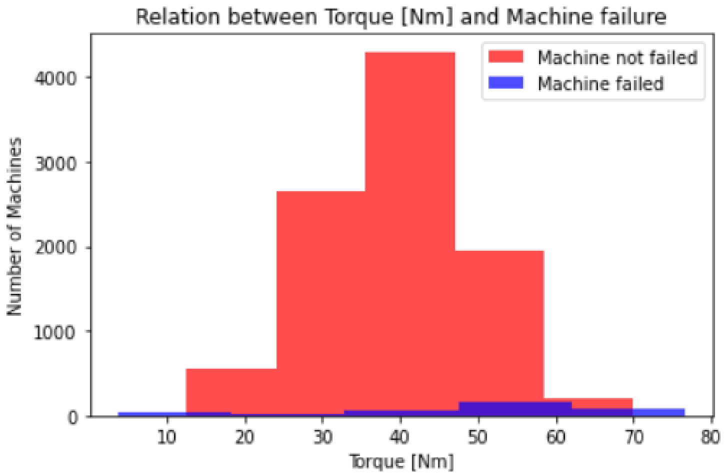

Figure 7 – Torque [Nm] VS Number of Machines

From figure 7 we can clearly see that there are more machine failure incidents at higher machine Torques. We can see that there are high percentage machine failures for Torque greater than 62 Nm.

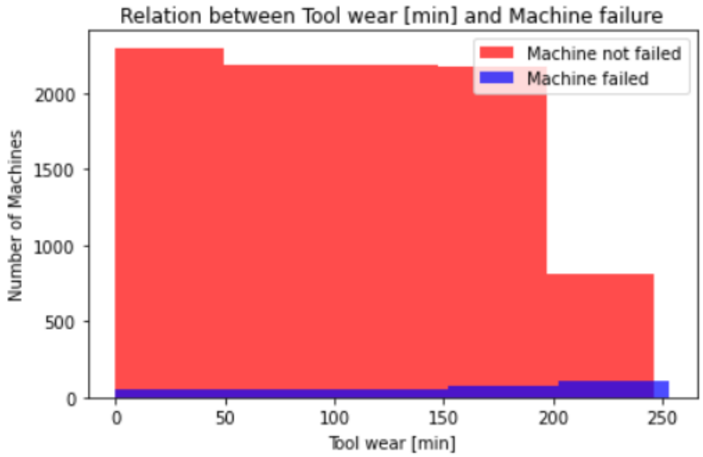

Figure 8 - Tool wear [min] VS Number of Machines

Machines with tool wears more than 200 minutes have higher chances of break down.

MODELS AND INTERPRETATION

DATA PREPROCESSING

There was not much work in processing data as there were no NAs in any of the columns. The only processing conducted is label encoding for the feature, ‘Type’ as it has the categorical values ‘L, M, H’, the encoded feature name is Types_Cat. Later, the data has been subset with the features mentioned below.

- Tool wear [min]
- Torque [Nm]
- Rotational speed [rpm]
- Air temperature [K]
- Process temperature [K]
- Types_Cat

The data has been split into trainset (70%) and testset (30%). By doing that we have 237 machine failure incidents in trainset and 102 machine failure incidents in testset. This check helps to have a comparison with our models’ prediction.

LOGISTIC REGRESSION

Logistic regression is used as our first model as it is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes.

The model has fitted with trainset and validated with test set. The ROC cure for the train and test sets can be seen in the below figure. We can see that it's a good fit with Area Under Curve (AUC) of train equal to 81.56 % and AUC of test equal to 80.24%, which means the model has good discriminatory ability.

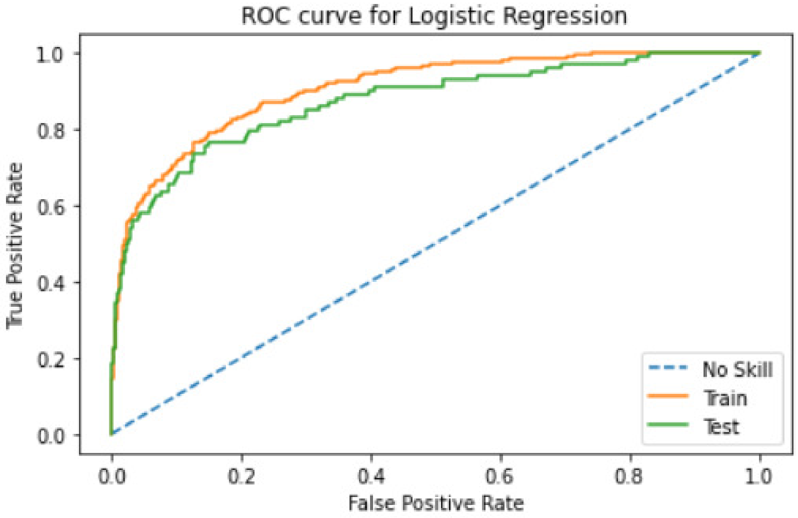

Figure 9 - Logistic regression ROC Curve

The threshold of 0.040175 is chosen to predict the machine failure incidents in test set. The best threshold is calculated through ROC curve by taking the geometric means of thresholds and selecting the threshold with best True Positive Rate. It is important to choose optimum threshold to predict as the default threshold i.e., 0.5 can lead in poor performance of the model. With the threshold chosen through ROC the accuracy of the trainset is 85.21% and test set is 84.67%.

Figure 10 shows the confusion matrix for train and test sets. The model predicted 184 machine failure incidents properly in train set which is 77% of the actual machine failures in trainset. Confusion matrix on test set says that the model predicted 77 machine failures correctly which is 75% of the actual machine failures. The model performed well in predicting the non-failure incidents which is around 85% for both train and test.

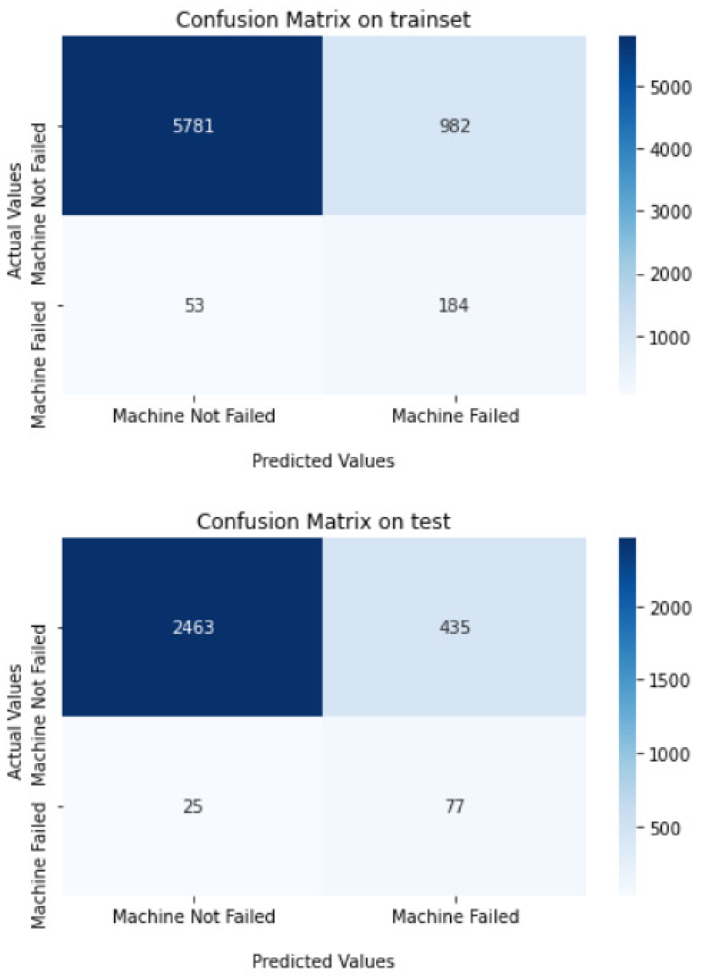

Figure 10 - Confusion Matrix for Logistic regression

MODEL INTERPRETABILITY – LOGISTIC REGRESSION

After the model is trained and predicted the machine failures for test, interpretability techniques are used to make a more informed decision.

PARTIAL DEPENDENCE PLOT

The partial dependence plot is used because it shows the marginal effect of one or two features have on the predicted outcome of a machine learning model and show the relationship between the target variable and a feature is linear or complex. Figure 11 show below is the partial dependence plots of features for logistic regression.

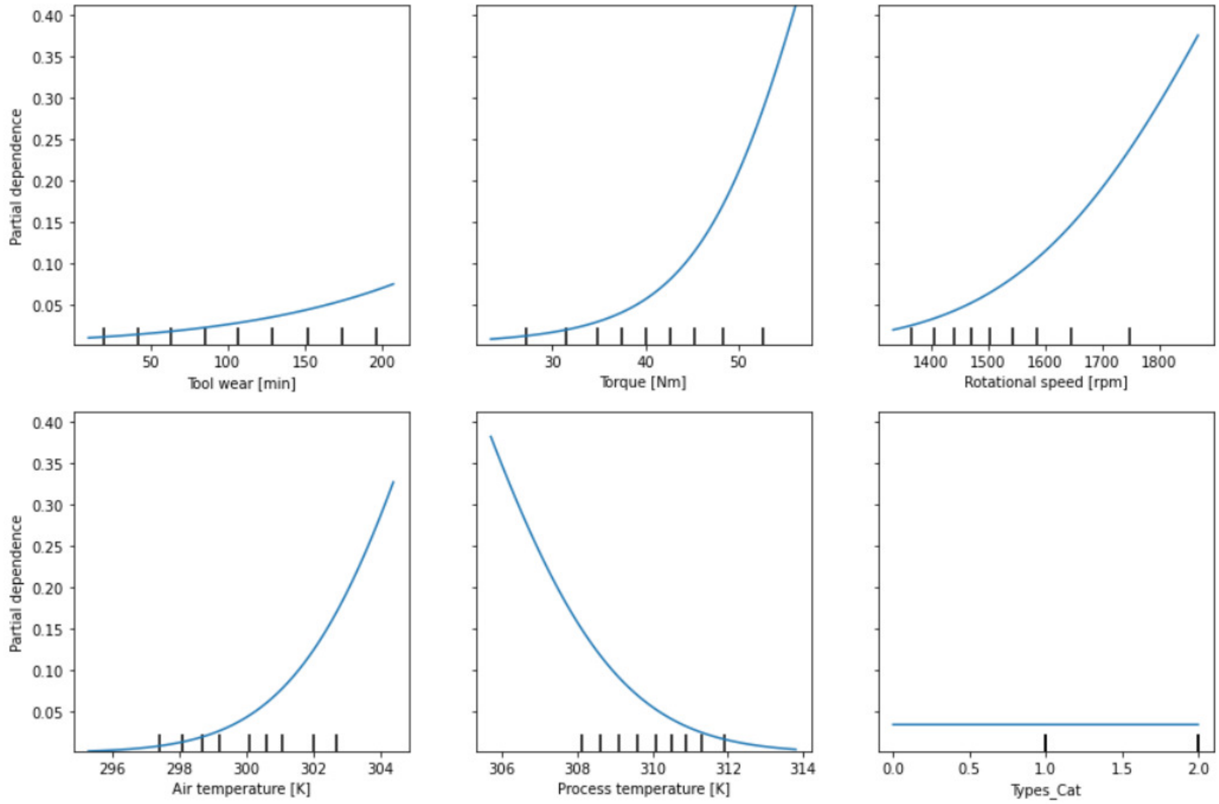

Figure 11 - Partial Dependence Plot - Logistic Regression

It can be seen from the plot that Torque [Nm], Rotational speed and Air temperature have more positive impact on machine failure and Process Temperature has the negative impact. Though Tool wear has some positive impact it is negligible as seen when compared to other features. The Type of machine has no impact on failure incidents.

ACCUMULATED LOCAL EFFECT – LOGISTIC REGRESSION

ALE plots highlight the effects of features have on the predictions of a machine learning model by partially isolating the effects of other features. The resulting ALE explanation is centered around the mean effect of the feature, such that the main feature effect is compared relative to the average prediction of the data.

Figure 12 show below is the ALE for individual features. The plots say Torque [Nm] and Rotation speed [rpm] have higher impact on machine failures.

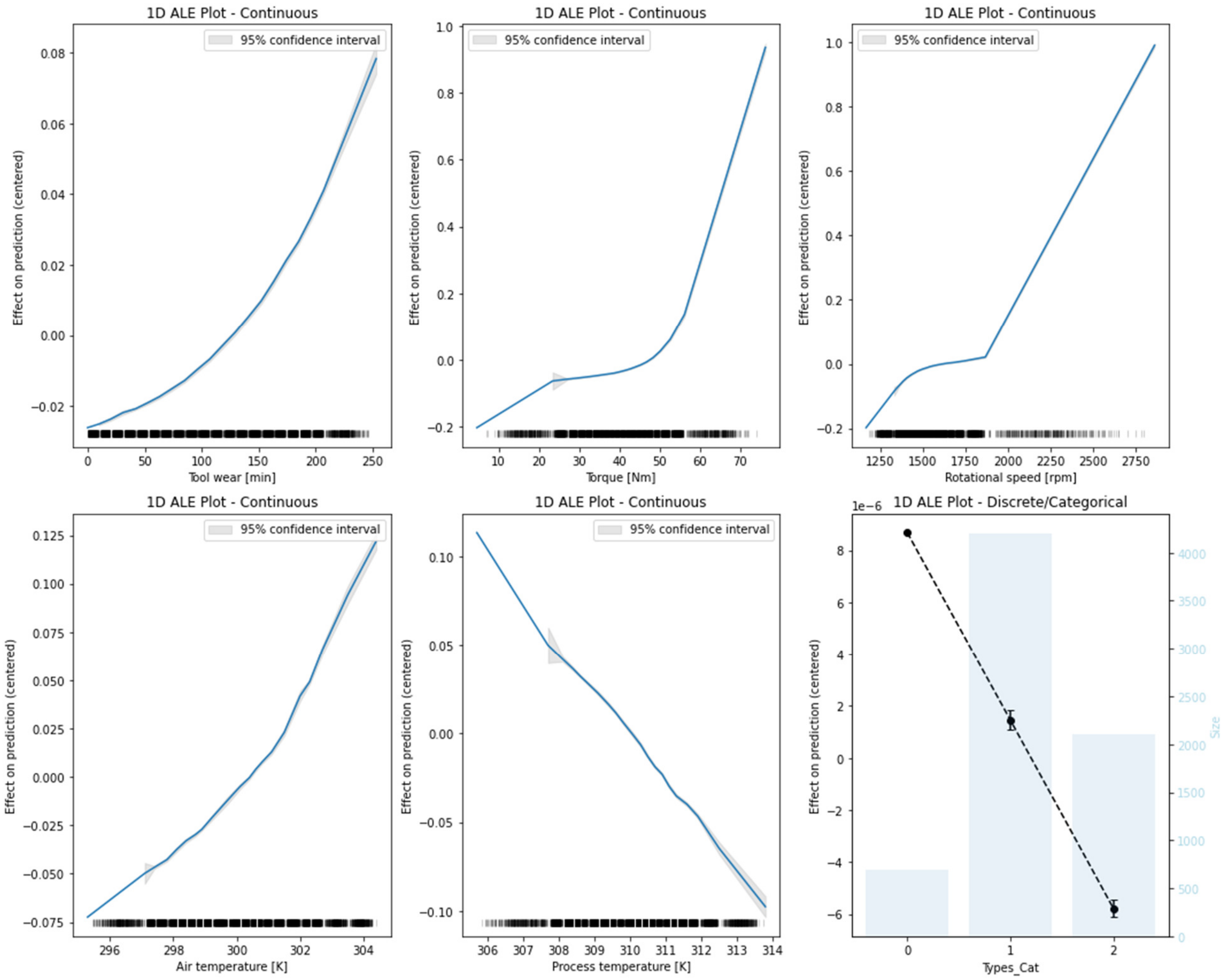

Figure 12 - ALE Plot Logistic Regression

SHAPLEY VALUES – LOGISTIC REGRESSION

The Shapley value is the average expected marginal contribution of one feature after all possible combinations have been considered. Shapley value helps to determine a payoff for all the features when each feature might have contributed more or less than the others.

Figures 13 and 14 shown below are the global interpretation of Shapley values and the impact of features on model output for Logistic Regression. Torque has the highest impact on Machine failure followed by Rotational speed, Air and Pressure temperature. Figure 14 says that Torque and Rotational speed have highest positive impact on model output followed by Air temperature which means as they increase the machine failure probability increases while Process temperature has negative impact on model output which means as it decreases the failure probability increases.

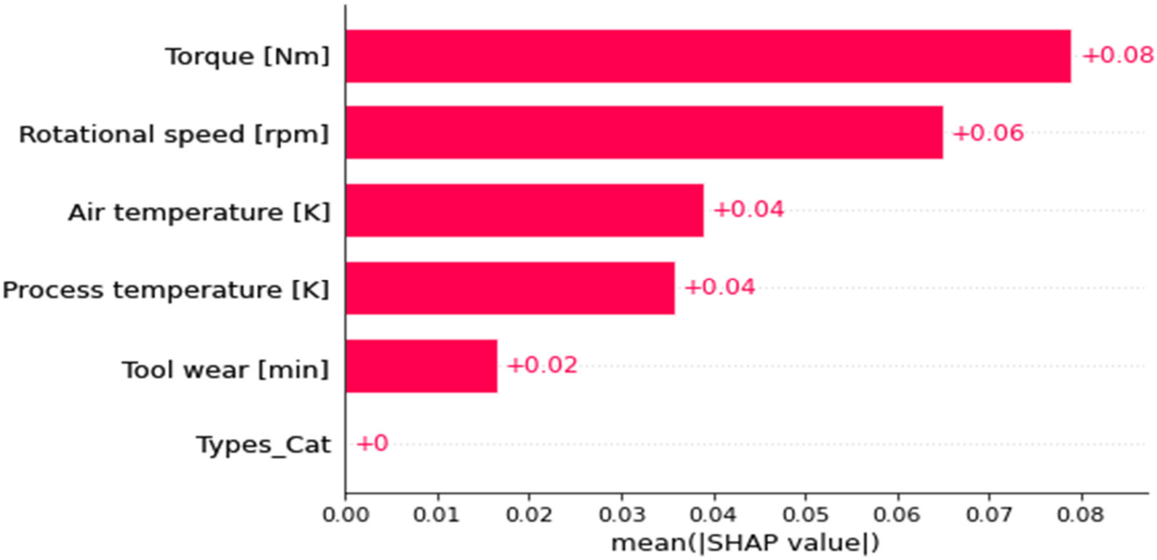

Figure 13 - Global Shapley Values for Logistic Regression

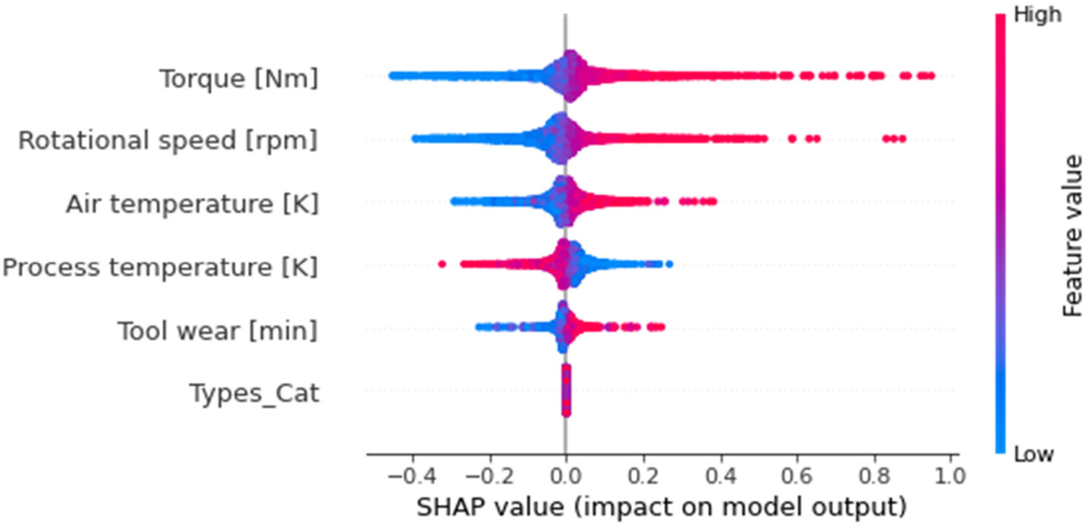

Figure 14 - Shap Value impact on model output (Logistic Regression)

RANDOM FOREST

Random Forest is a machine learning algorithm which is widely used in classification and regression problems. It performs exceptionally well for classification problems.

The model has fitted with trainset with 200 estimators and maximum depth of 6 and validated with test set. The ROC cure for the train and test sets can be seen in the below figure. We can see that it’s a good fit with Area Under Curve (AUC) of train equal to 93.13 % and AUC of test equal to 90.96%, which means the model has excellent discriminatory ability.

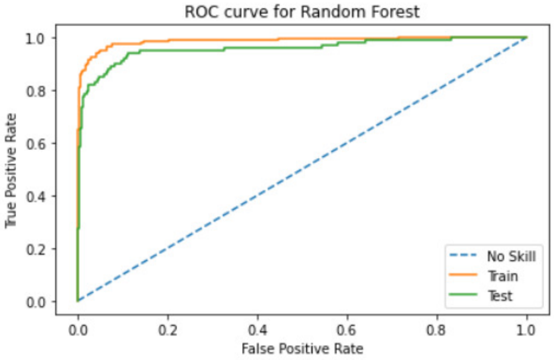

Figure 15 - ROC Curves for Random Forest

The threshold of 0.035077 is chosen to predict the machine failure incidents in test set. The best threshold is calculated through ROC curve by taking the geometric means of thresholds and selecting the threshold with best True Positive Rate. With the threshold chosen through ROC the accuracy of the trainset is 88.69% and test set is 88.93%.

Figure 16 shows the confusion matrix for train and test sets. The model predicted 232 machine failure incidents properly in train set which is 97.8% of the actual machine failures in trainset. Confusion matrix on test set says that the model predicted 95 machine failures correctly which is 93.13% of the actual machine failures. Here the model performed exceptionally well in predicting the machine failure.

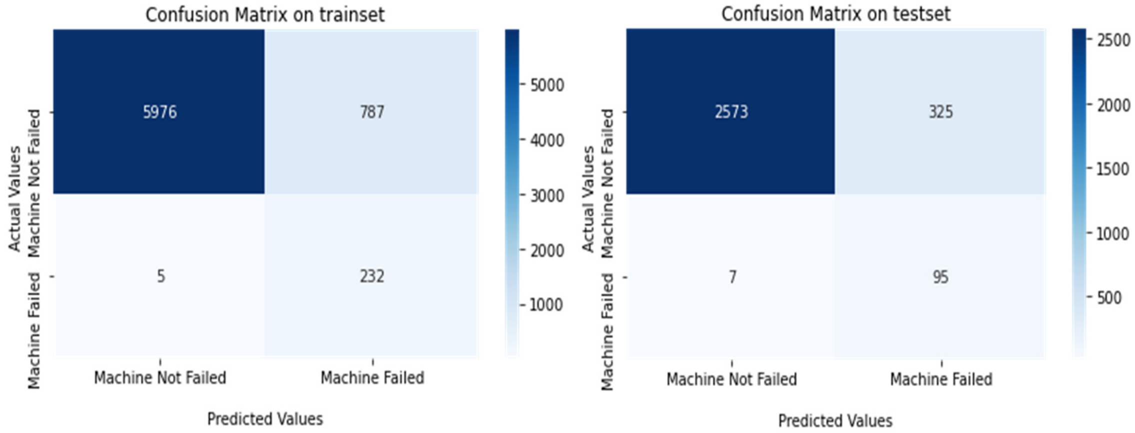

Figure 16 - Confusion Matrix for Random Forest

MODEL INTERPRETABILITY – RANDOM FOREST

After the model is trained and predicted the machine failures for test, interpretability techniques are used to make a more informed decision.

PARTIAL DEPENDENCE PLOT

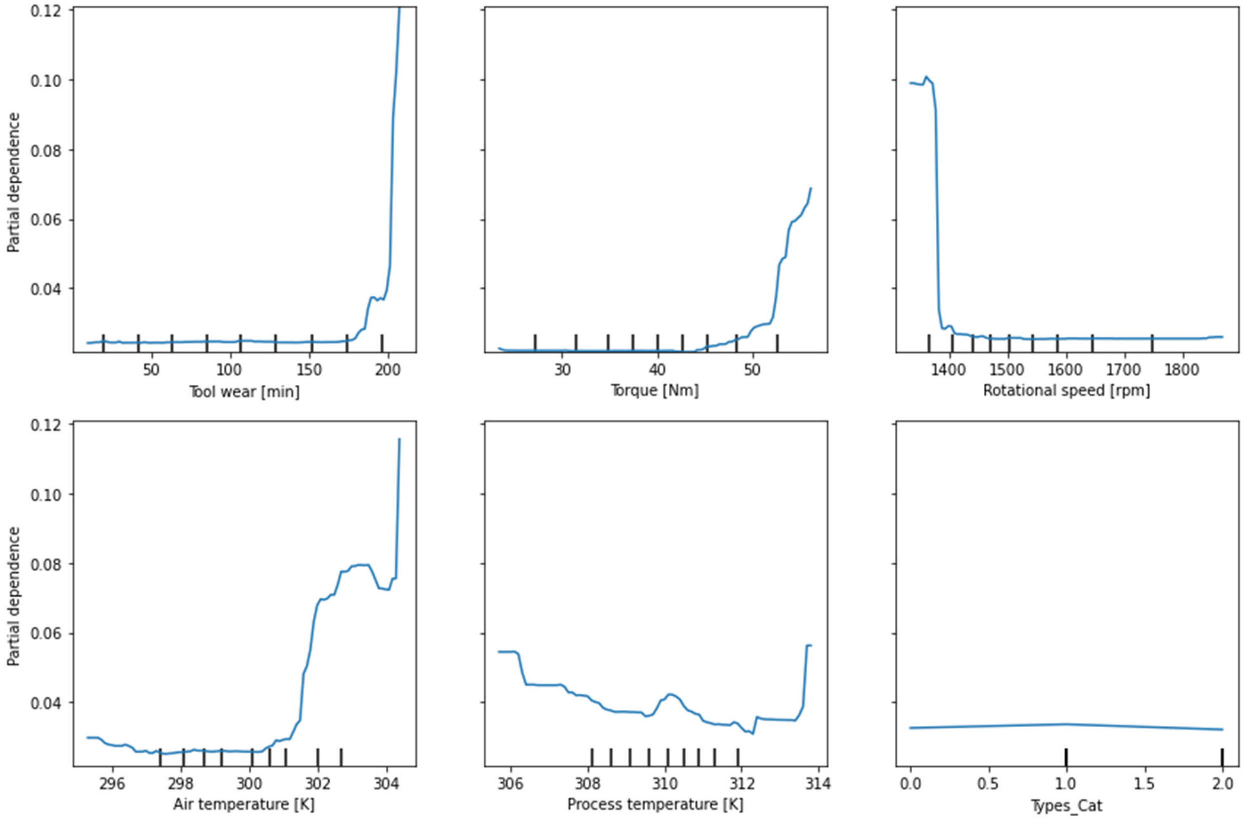

Figure 17 - Random Forest Partial dependency plot

It can be seen from the plot that Torque [Nm], Air temperature and Tool wear have more positive impact on machine failure and Rotational speed has the negative impact. The Type of machine has no impact on failure incidents and Process temperature impact can't be predicted with plot.

ACCUMULATED LOCAL EFFECT – RANDOM FOREST

Figure 18 show below is the ALE for individual features. The plots say Torque [Nm] and Rotation speed [rpm] have higher impact on machine failures. When the torque of machine is greater than 60 Nm there is a positive impact and negative impact when lesser than 22 Nm and rotational speed is above 2000 rpm.

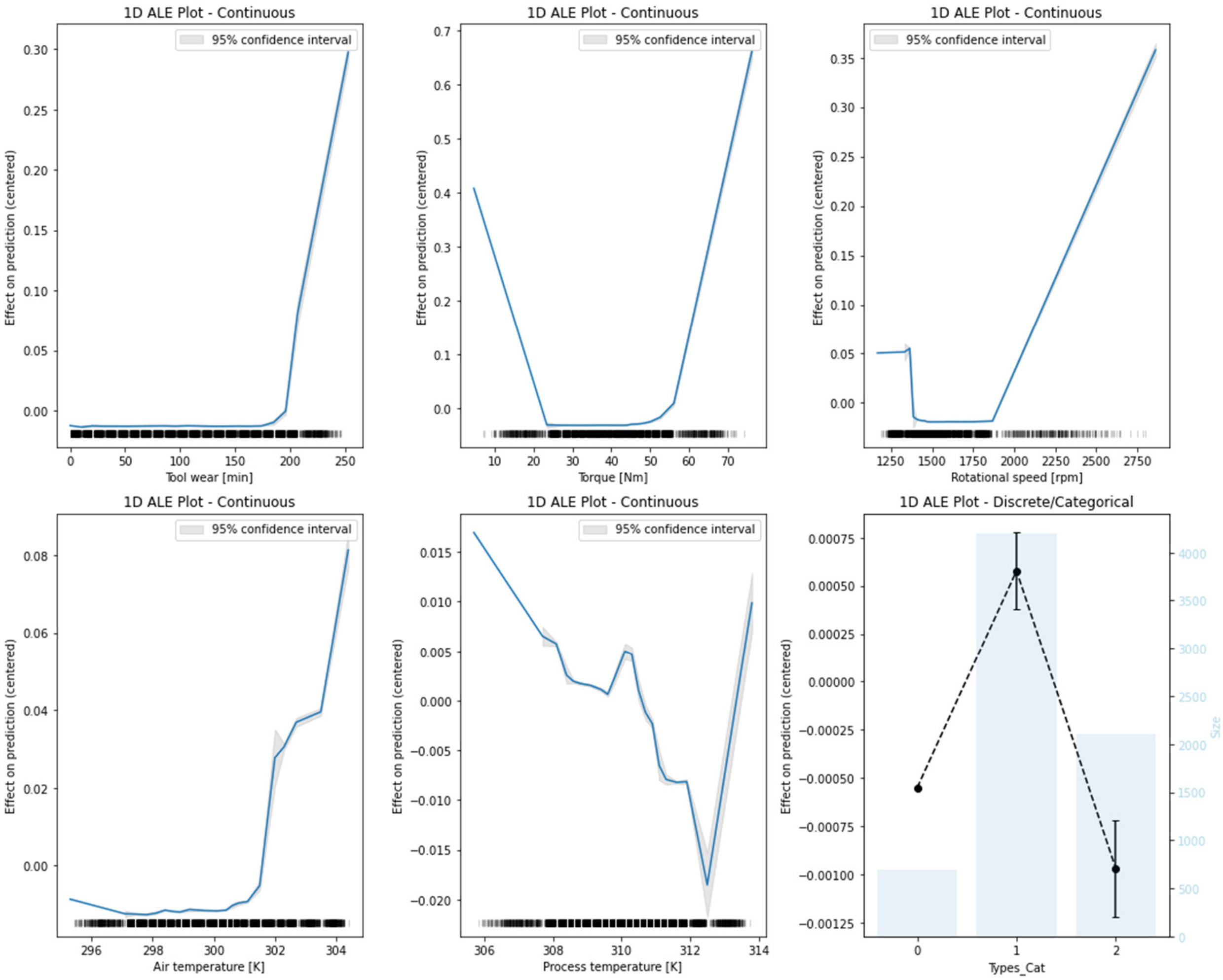

Figure 18 - ALE - Random Forest

SHAPLEY VALUES – RANDOM FOREST

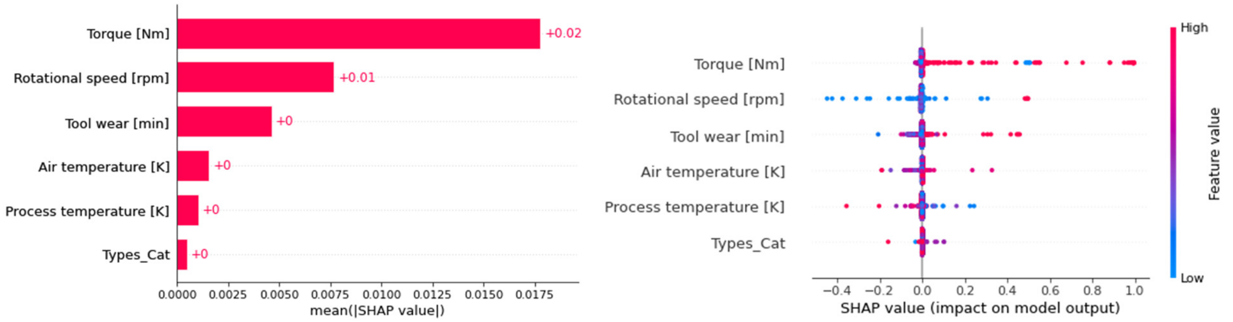

Figure 19 - Shapley Values

Figures 19 is the global interpretation of Shapley values and the impact of features on model output for Random Forest. Torque has the highest impact on Machine failure followed by Rotational speed. It says that Torque and Rotational speed have highest positive impact on model output which means as they increase the machine failure probability increases. While other features have very less impact on machine failure.

CONCLUSION

Two models have been used with same features. It is also clear from our models that Torque, and Rotational speed features have higher impact on the machine failure. Looking at the predictions of both the models we conclude that Random Forest performed better than Logistic Regression. Random Forest has better accuracy and AUC. Model evaluation was on test set with 3000 observations in which there were 102 machine failure incidents. Logistic Regression model predicted 77 incidents properly whereas Random Forest Model predicted 95 incidents properly. Therefore, it is advisable for the company to implement Random Forest model to predict the machine failure incidents.

REFERENCES

- https://scikit-learn.org/
- https://seaborn.pydata.org/