Communication Skills

Individual Assignment

Vinay Rajagopalan

Contents

Problem Statement

To optimize the production process for a car manufacturer which is currently a manual process.

Factors to consider
Investigate machine failure based on dataset collected.

Outcome Expected
Predictive Model that will be able to predict machine failure.
Results to provide insights on how to prevent future failures.

Data Description

The dataset consists of 10000 machine measurements and 9 features, including 339 machine failures.

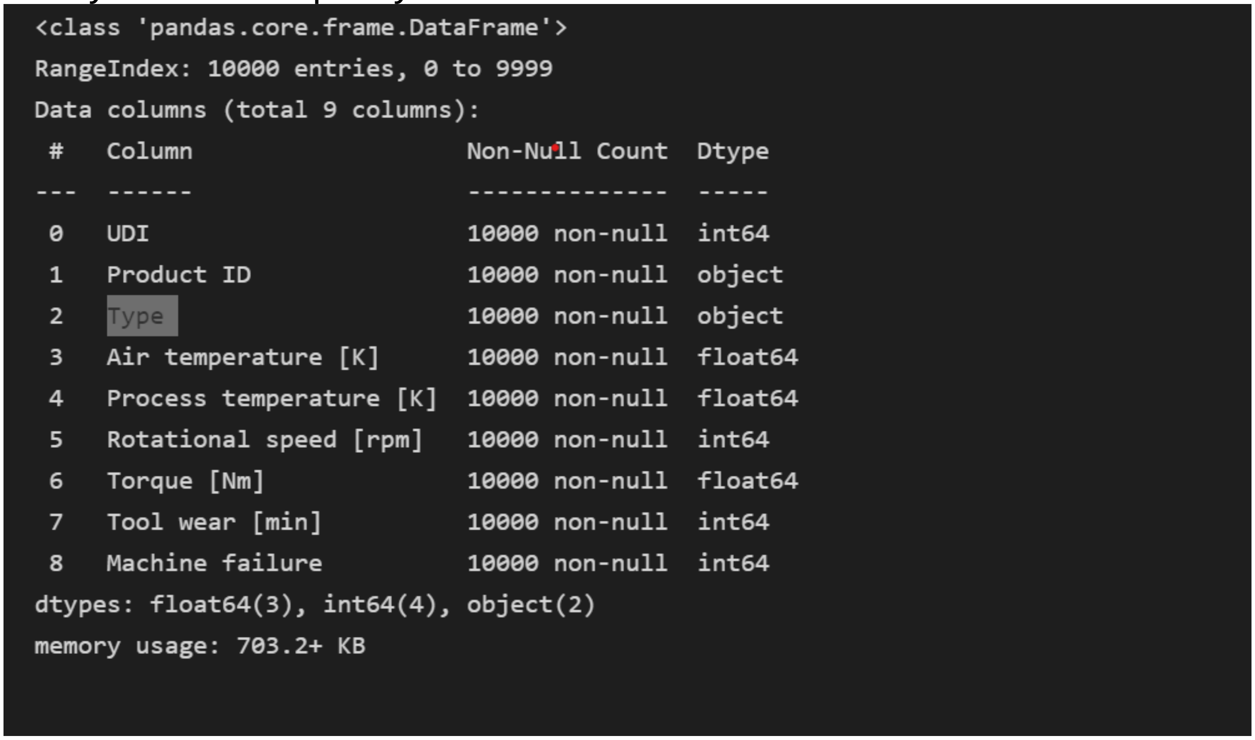
A brief overview on Data before we do some investigation using Exploratory Data Analysis.

UDI	The primary key for the event
Product ID	Product Ids for which measurements were collected
Type	Type of the Product indicating L, H, M.
air temperature [K]	Air temperature in Kelvin
process temperature [K]	Process temperature in Kelvin
rotational speed [rpm]	Rotational speed in rotations per minute
torque [Nm]	Torque in newton-metres
tool wear [min]	Tool wear in minutes
Machine failure	Label that indicates, whether the machine has failed in this particular datapoint

Exploratory Data Analysis

Data Structure

Let’s check the structure for the data to investigate if there any NA values or any other discrepancy that we can notice



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   UDI                      10000 non-null  int64  
 1   Product ID               10000 non-null  object 
 2   Type                     10000 non-null  object 
 3   Air temperature [K]      10000 non-null  float64
 4   Process temperature [K]  10000 non-null  float64
 5   Rotational speed [rpm]   10000 non-null  int64  
 6   Torque [Nm]              10000 non-null  float64
 7   Tool wear [min]          10000 non-null  int64  
 8   Machine failure          10000 non-null  int64  
dtypes: float64(3), int64(4), object(2)
memory usage: 703.2+ KB
```

We see that there are no NA values, and we also see that all the data types are correct and no further corrections are required.

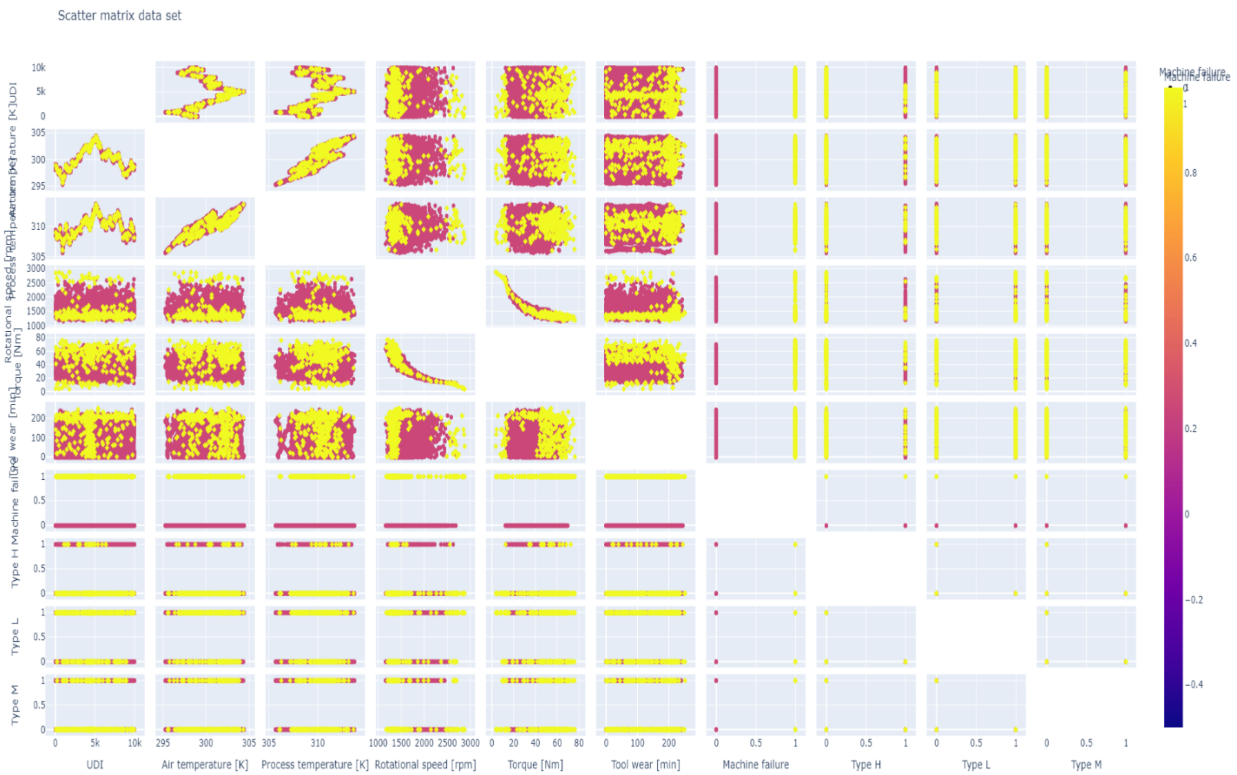
Pre-Processing

Looking at the data we see that Type column is categorical variable, so we dummy encode them and further append them as columns to the data. The Product Id are also unique and thus we have two primary key identifiers so we would be dropping the Product Id column and only keeping the UDI column. Thus, our final data table would consist of 3 additional column Type_L, Type_H and Type_M indicating whether the given product belongs to which type.

Insights

In this section we will try to gain insights on the data to check if can see some patterns that might be causing the machine failures. We are specifically looking for how the data distribution is in general for various variables.

We will first check the scatter plots for all the variables first.



Looking at the scatter plot the first thing to notice is that we do see that at higher rotational speed there is an increased chance of failures. All values above 2695 rpms have seem to have resulted in failures.

In the case of process temperature and air temperature we do not see much of a pattern.

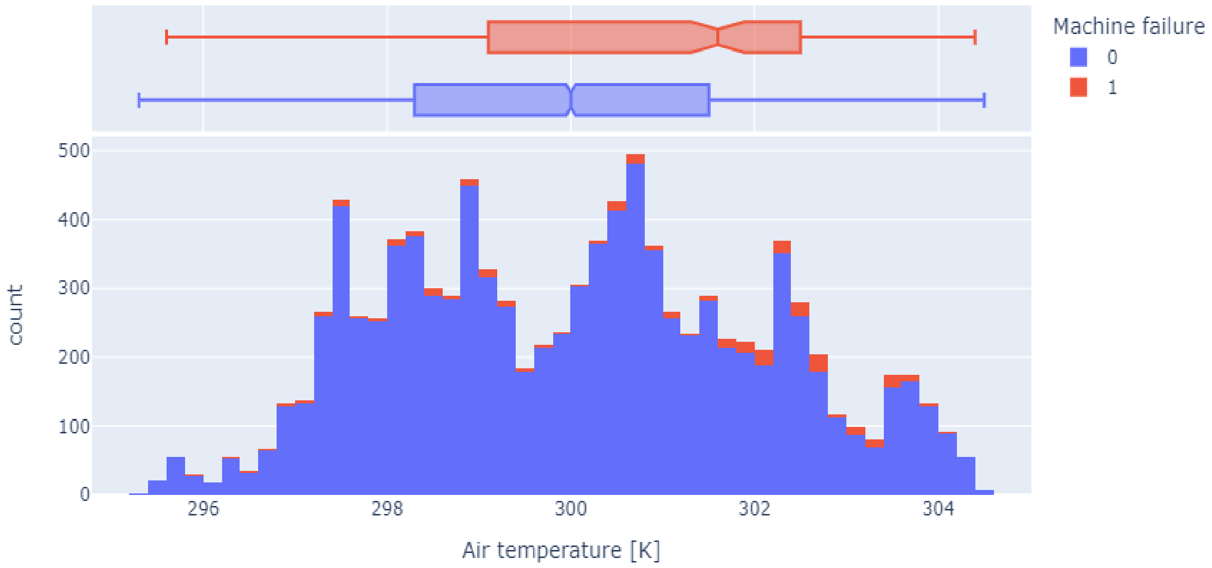
Torque also seems to be a contributing factor for the machine failures we do see that all reported cases with torque higher than 70 Nm has resulted in failures. Another thing to notice is that all cases where torque was lower than 12.6 Nm also seems to signify failures. Thus, implying that higher and lower values of torque seem to be causing problems.

Tool wear and all dummy encoded type features do not seem to show any pattern that stands out.

We already gained two important insights which shows that higher and lower values of torque and higher rotational speeds seems to be causing machine failures.

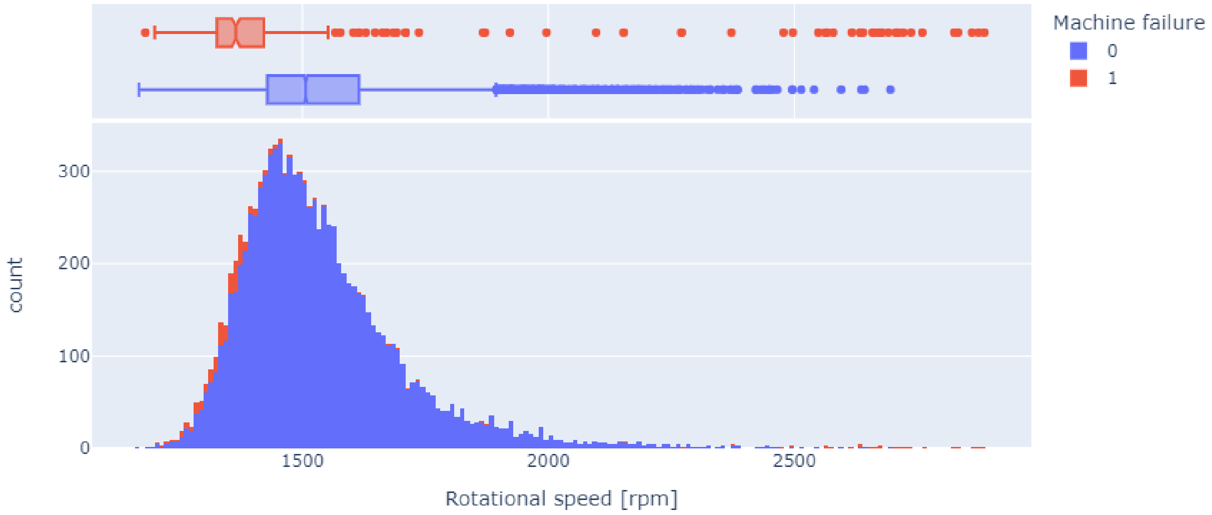
Next, we will try to dive into the distribution of these variables and check if there is any pattern we could detect.

Air Temperature



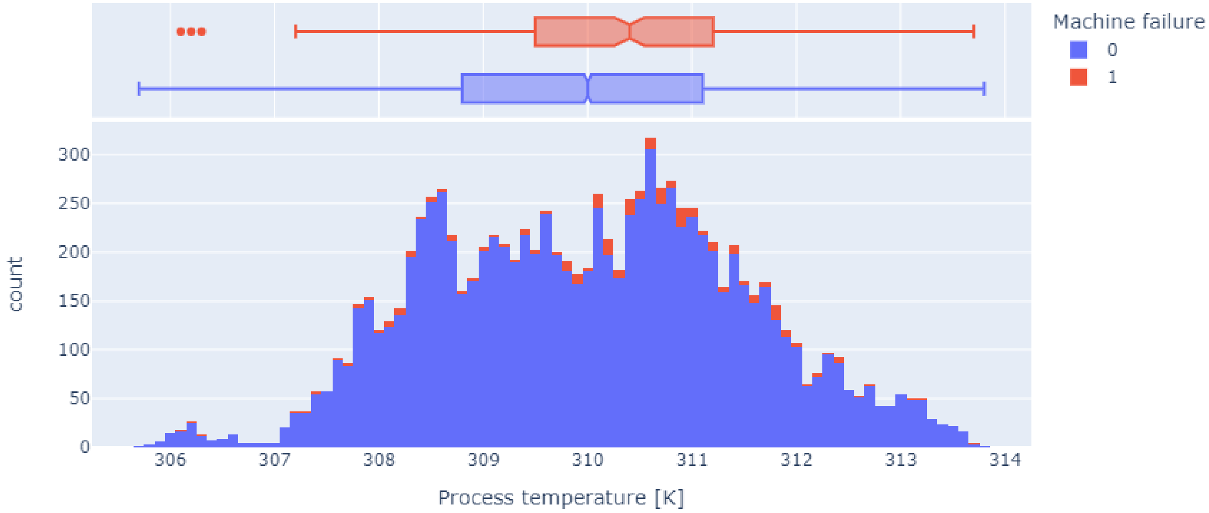
Looking at the distribution plot of Air Temperature we can see that there does not seem to be any outliers for the data. The data does not have any skewness. The median value of Air temperature for machine failure is 301.6 K. We also see that in case of no failures the median value is 300 K thus signifying that higher air temperature seems to be causing machine failure.

Rotational Speed



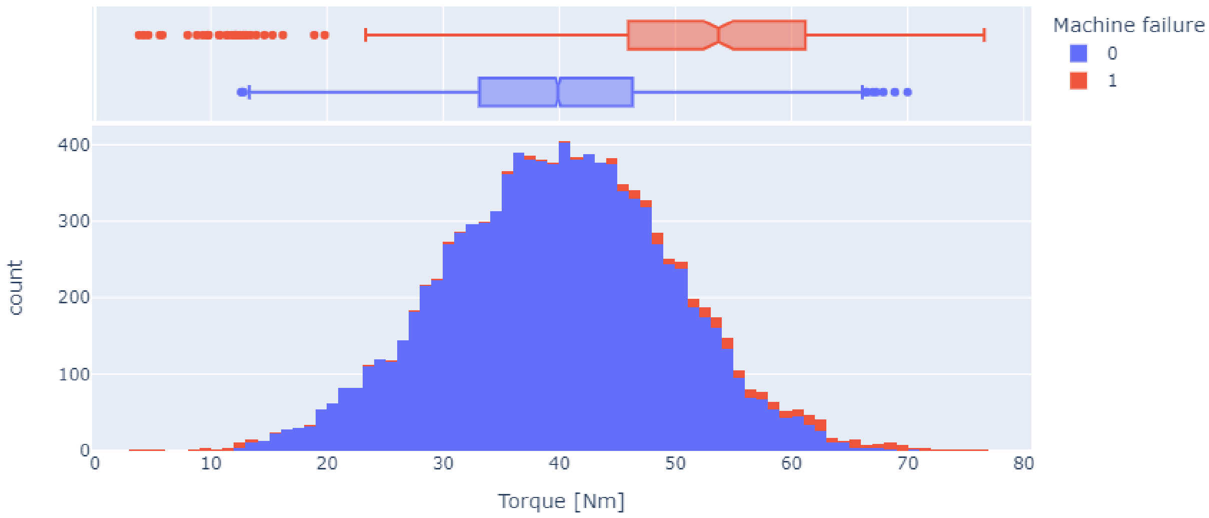
The distribution for the rotational speed has a skewness towards the right and we also see a lot of outliers in this case. The median value of rotational speed for machine failure is1365 rpm. We also see that in case of no failures the median value is 1507 rpm. Due to the outliers in the data, we cannot say if lower values signify lower speed cause failures. But we already saw that in the case of scatter plots higher speeds were causing failures.

Process Temperature



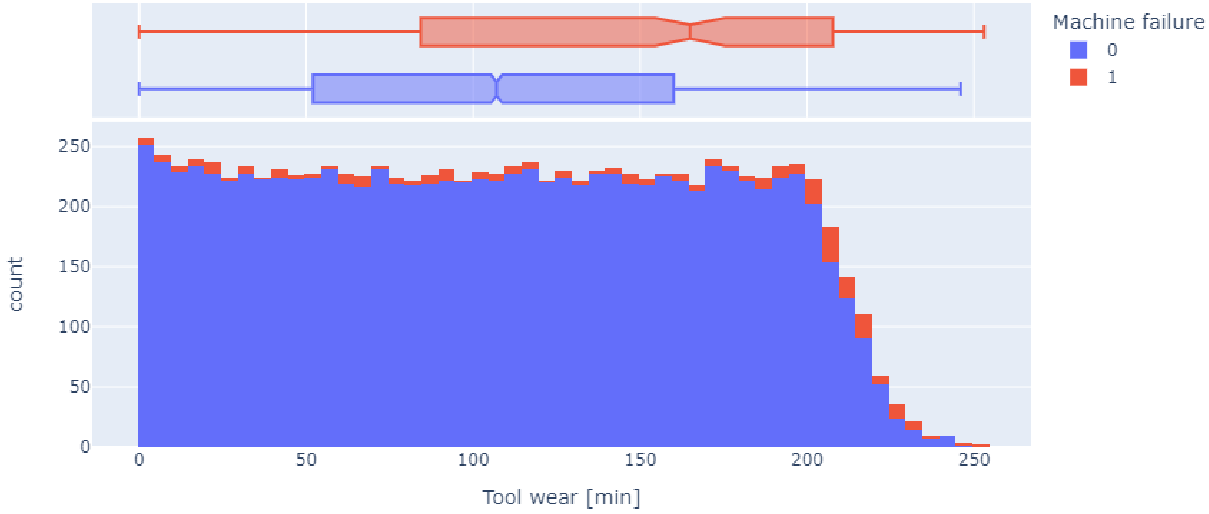
The Process Temperature has a skewness towards the left and there seems to be some outliers in case of process temperatures being 1. In case of Process Temperature, the median value for non-failure and failure is 310 K and 310.4 K. The difference much is not much.

Torque



In case of torque, we see that data is has some skewness towards the right. There are some outliers as well. In case of Torque the median for failures is 53.7 rpm and for non-failures is 39 rpm thus indicating that higher the torque greater is the chance of failure.

Tool wear



In case of Tool wear, we see that data there is no outliers and the mostly is following a uniform distribution with skewness on the right. In case of Tool wear the median for failures is 165 min and for non-failures is 107 min thus indicating that higher the value of toolwear greater is the chance of failure.

Primary Conclusion

Approximately 3% of the data results in machine failure. Based on the insights we see that higher air temperature, Tool wear, Torque increases probability of machine failure while lower rotational speed also increases the probability of machine failure. Process temperature and Type does not seem to contribute much.

Predictive Maintenance Model

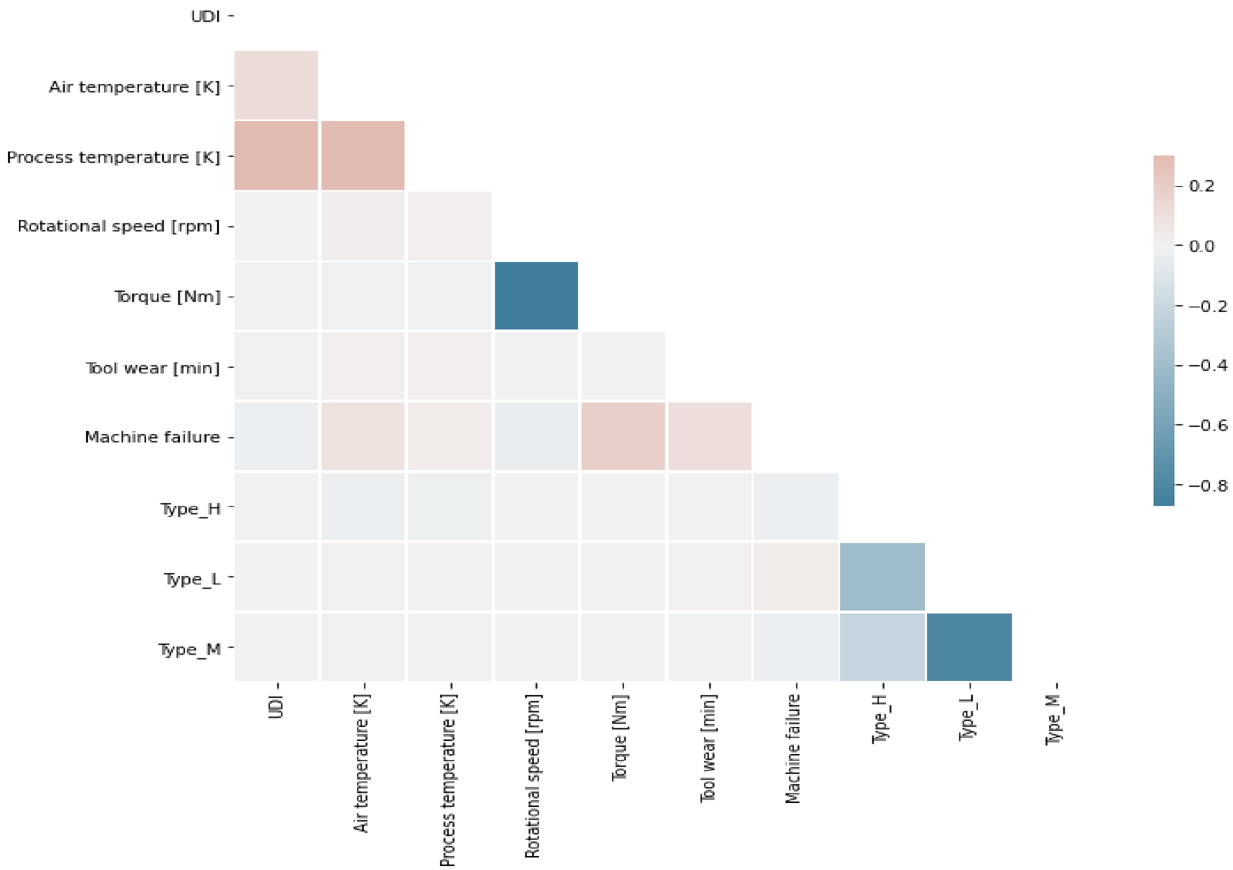
We have now gained some valuable insights on the data from machine failure. We will now try to train few machines learning models like logistic regression, decision tree and few black box models like random forest and neural network. The goal of these models is to be able to predict in future probability of machine failures based on the features that were provided.

Data Preparation

Looking at the data we see that Type column is categorical variable, so we dummy encode them and further append them as columns to the data. The Product Id are also unique and thus we have two primary key identifiers so we would be dropping the Product Id column and only keeping the UDI column. Thus, our final data table would consist of 3 additional column Type_L, Type_H and Type_M indicating whether the given product belongs to which type.

The data is split 70/30 for the purpose of training and testing the model performance.

The correlation matrix for the features are as follows:



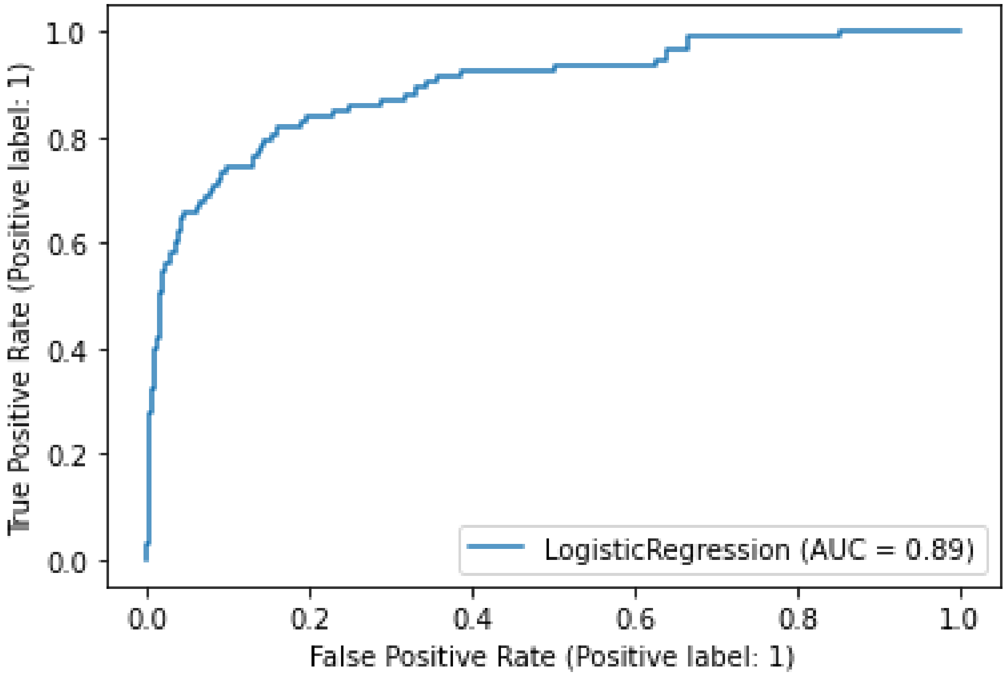
From the correlation matrix we can see that Process Temperature and Air Temperature seems to high correlation while torque and rotational speed seem to have high negative correlation. Type L and Type M also seems to share similar trend.

Model

Logistic Regression

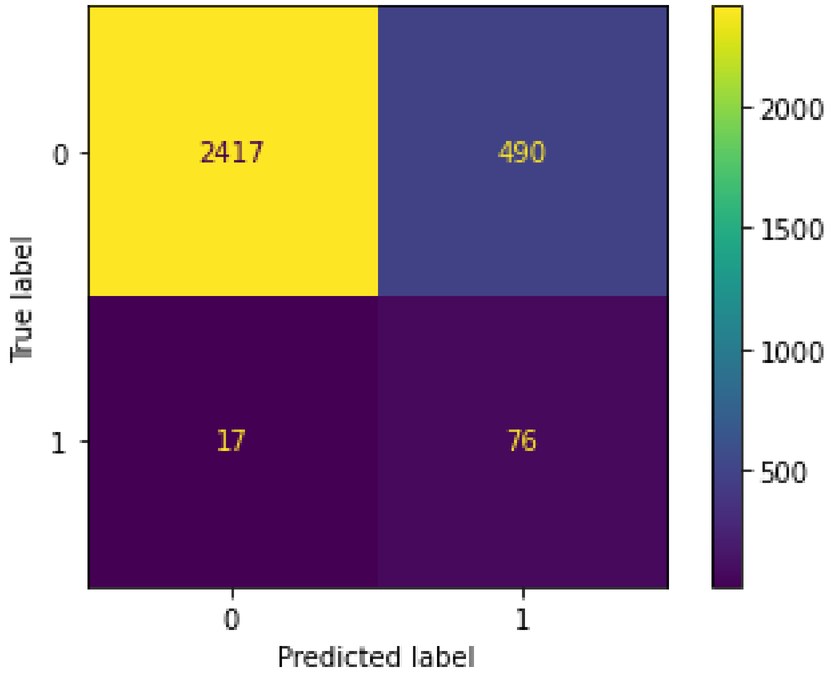
The model was trained with max_iter = 1000
The accuracy of train was 81.87% while the accuracy for the test was 83.1%.

The ROC curve for the model is as follows:



We have an AUC of 89% in this case thus signifying that we are able to cover 89% of our data using this model.

The confusion matrix for the same is as follows:

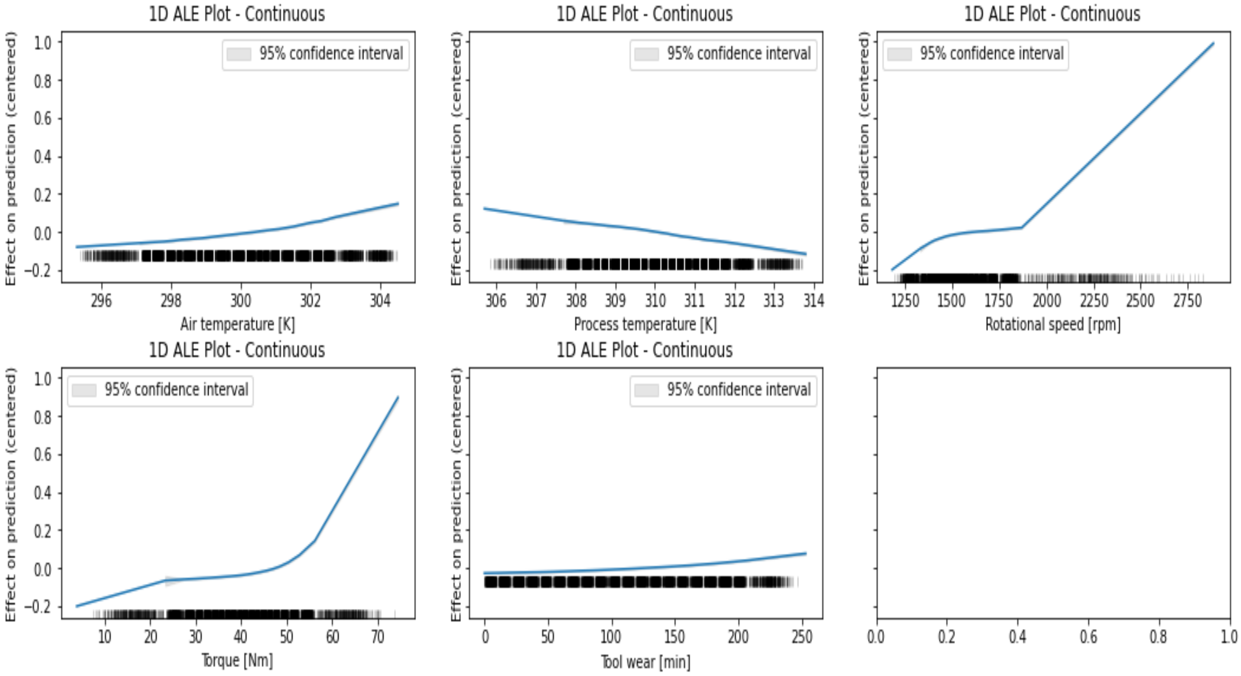


We can clearly see that we are able to accurately predict 76 True Positive and 2417 True Negative cases in the Test data set.

Let’s Interpret the results of Logistic Regression using ALE, ICE and PDP to evaluate the relationship of features with target variables.

ALE

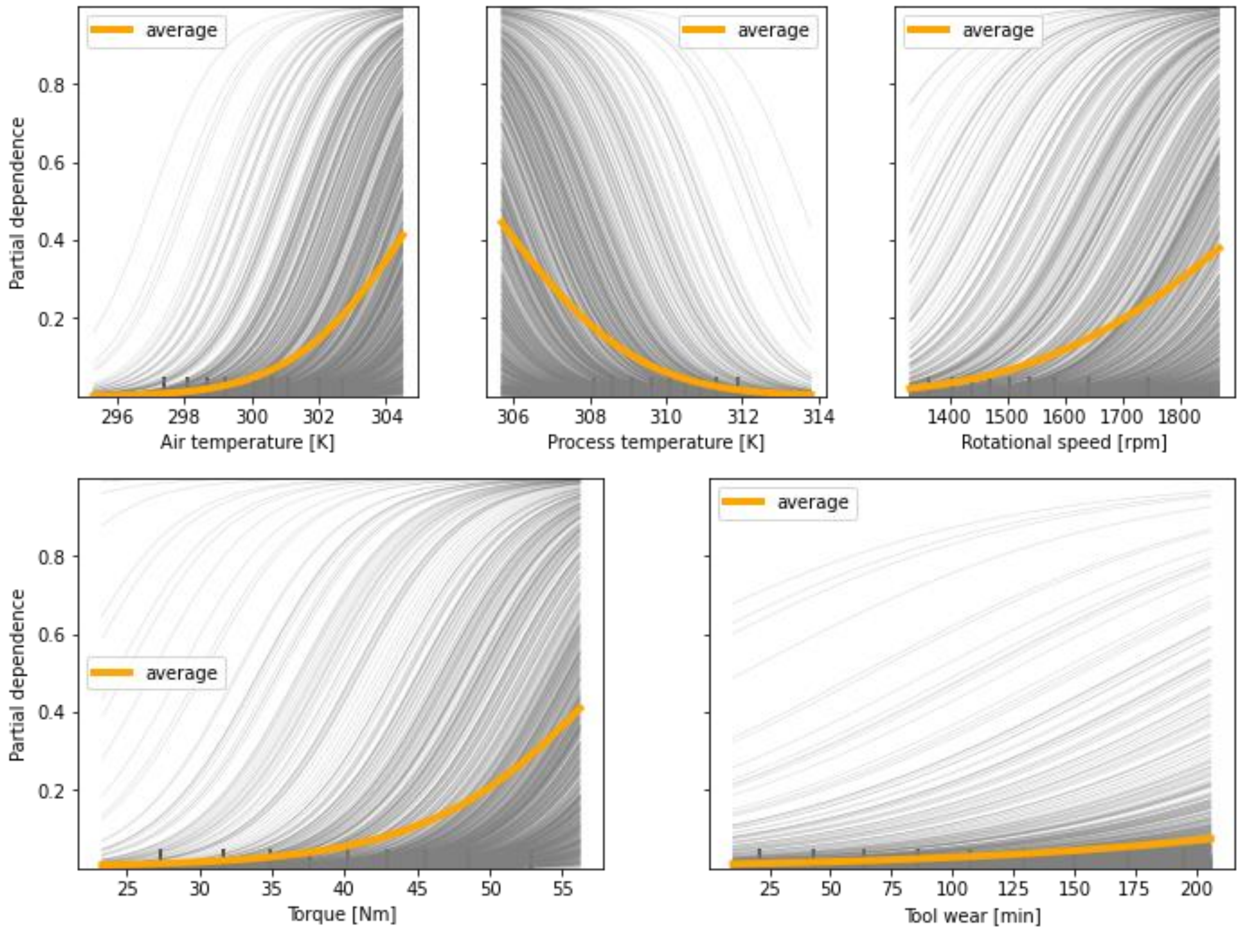
Accumulated local effects describe how features influence the prediction of a machine learning model on average



Looking at the ALE plot we can clearly see that Air Temperature, Rotational Speed, Torque and Tool wear causes failure as they increase while for process temperature, we see that as it decreases it causes more failures.

ICE

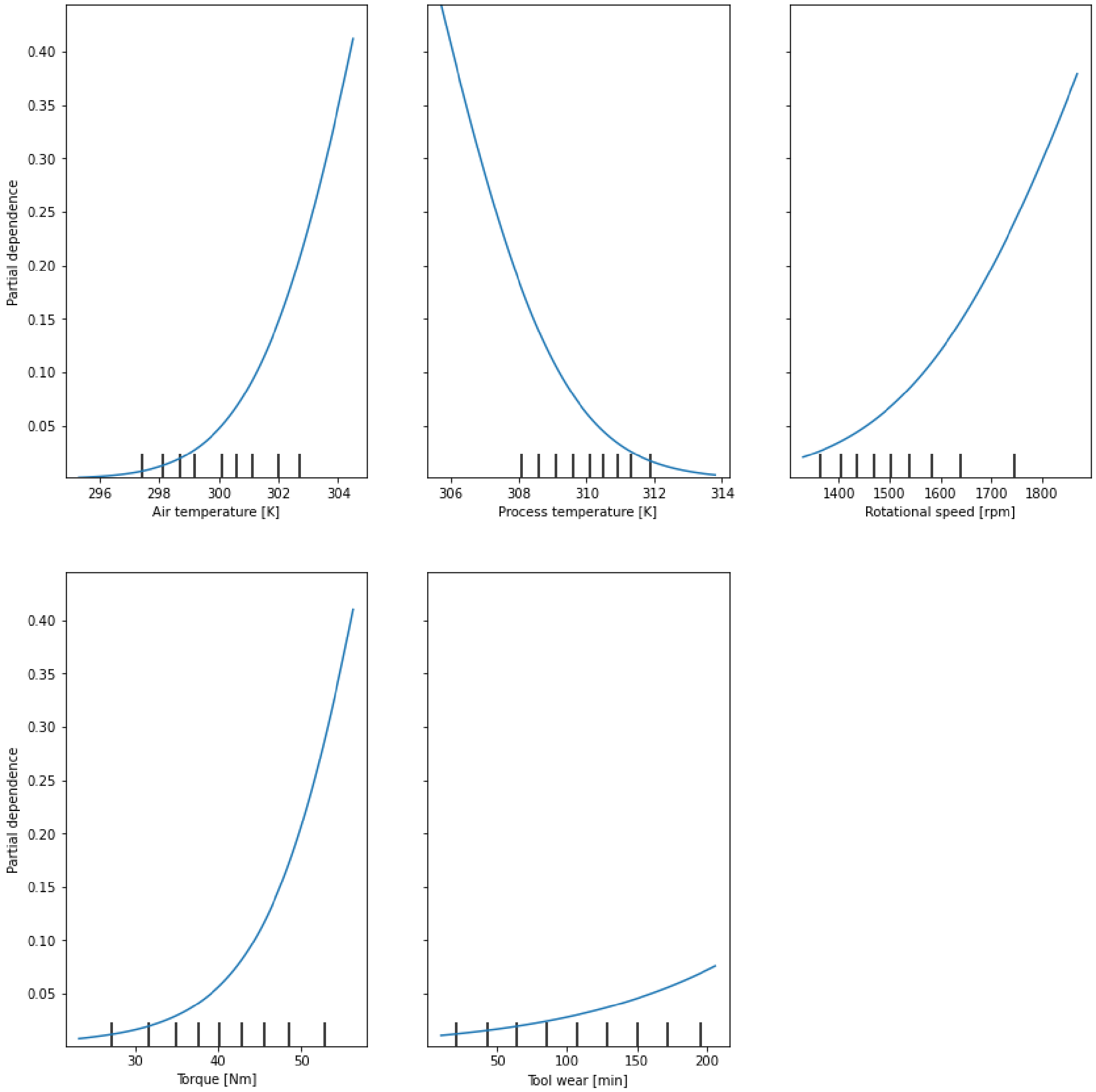
An individual conditional expectation (ICE) plot shows the dependence between the target function and a feature of interest.



Looking at the ICE plot we can clearly see that Air Temperature, Rotational Speed, Torque and Tool wear causes failure as they increase while for process temperature, we see that as it decreases it causes more failures.

PDP

Partial dependence plots show the dependence between the target function and a set of features of interest, marginalizing over the values of all other features.



Looking at the plot we can clearly see that Air Temperature, Rotational Speed, Torque and Tool wear causes failure as they increase while for process temperature, we see that as it decreases it causes more failures.

Conclusion

Looking at the plot all plots signify similar trends, and it is alignment with our initial hypothesis that increase in air temperature and Torque causes machine failure is also captured by the model.

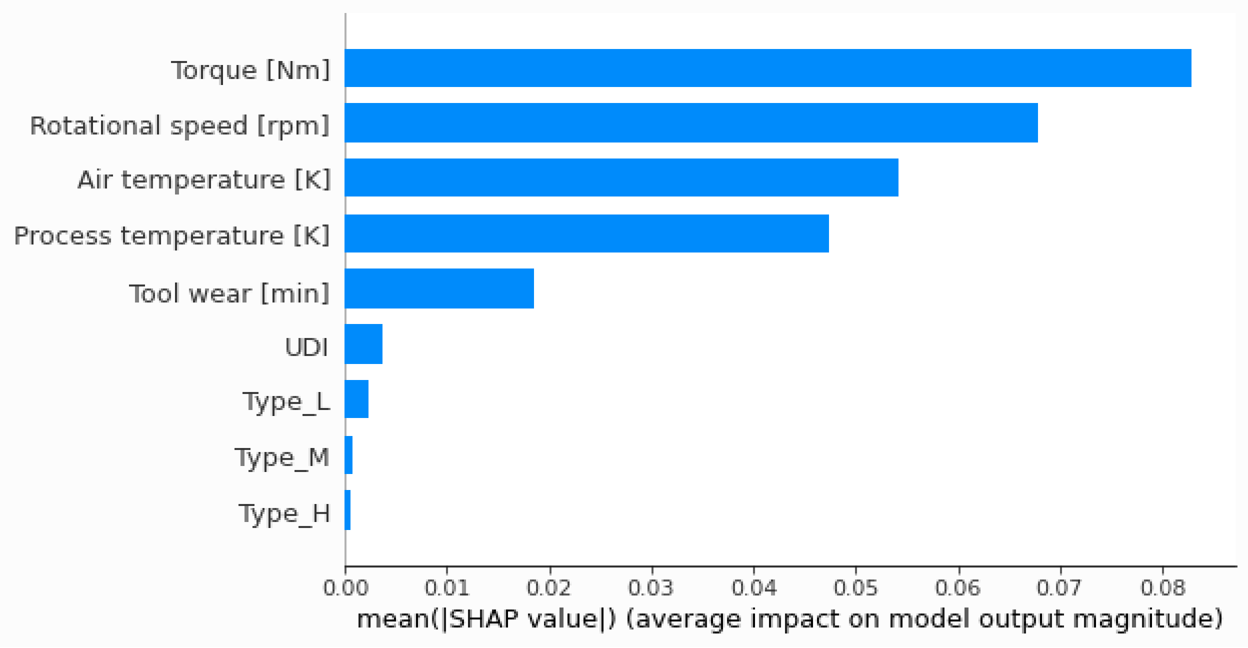
Data Subset

We will further try to evaluate our using Shapley Feature Selection or rather sub setting the data for important features.

We will be using Shapley values for feature selection.

Shapley

Shapley values are a widely used approach from cooperative game theory that come with desirable properties.

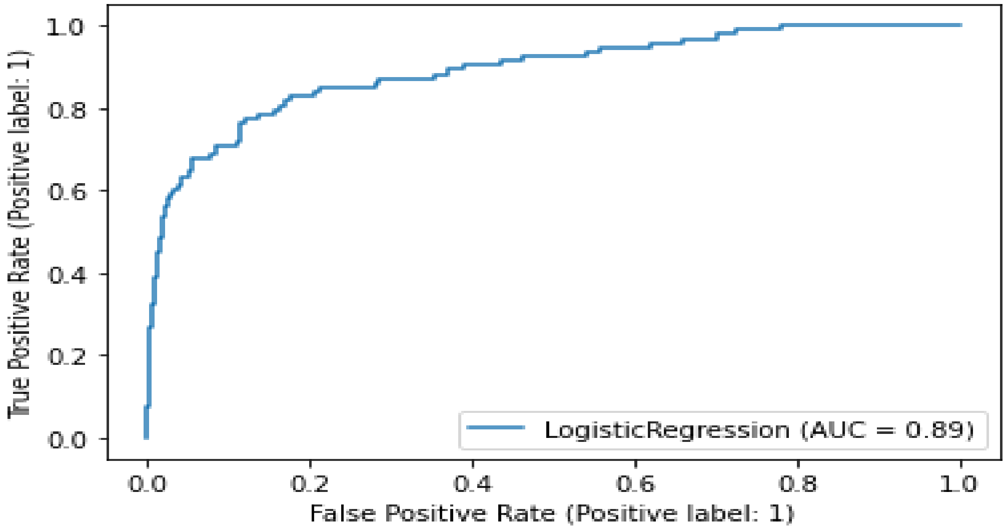


Based on the shapley values we see that Torque, Rotational Speed, Air Temperature, Process Temperature and Tool Wear are important features.

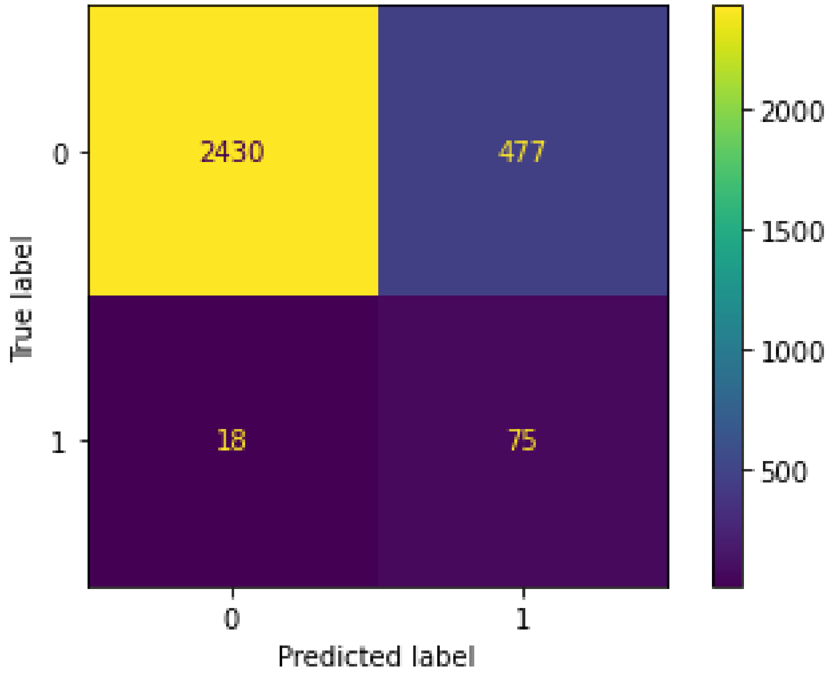
We will be using these features and retraining our model.

The accuracy of the model improves slightly in this case 81.5% for the train set and 83.5 for the test set.

The ROC curve remains the same



The confusion matrix

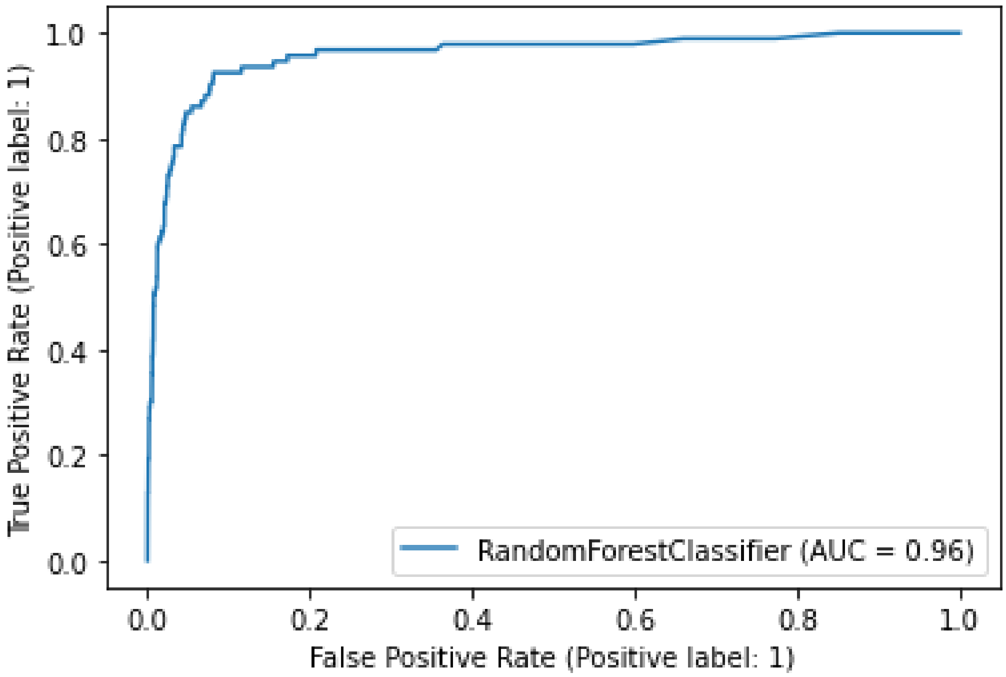


Thus, we can see that there is an improvement in predicting True False cases compared to the previous Model.

Random Forrest

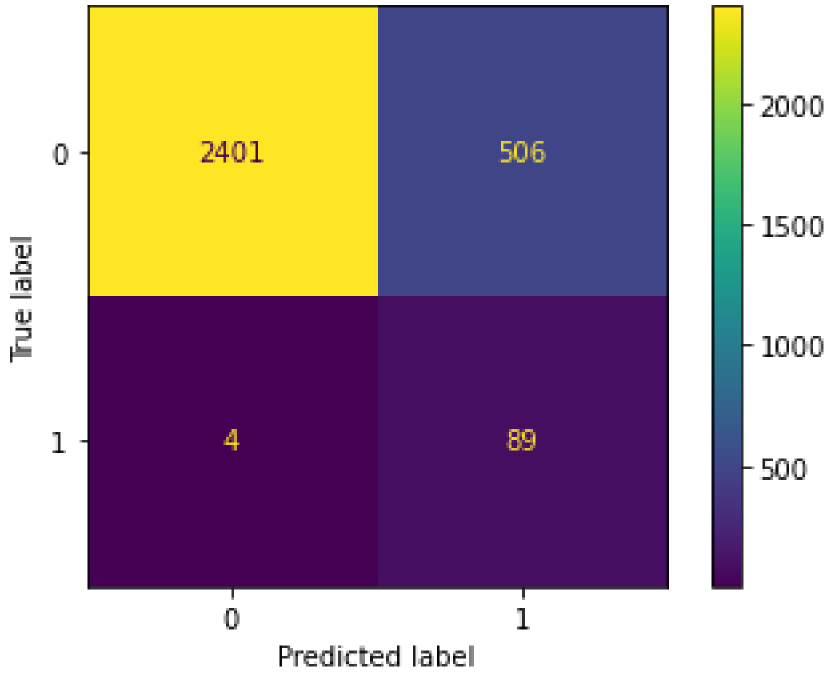
The model was trained with n_estimators=1000, max_depth=8, n_jobs = 100, max_samples = 182, random_state=42.
The accuracy of train was 83.11% while the accuracy for the test was 83%.

The ROC curve for the model is as follows:



We have an AUC of 96% in this case thus signifying that we are able to cover 96% of our data using this model.

The confusion matrix for the same is as follows:

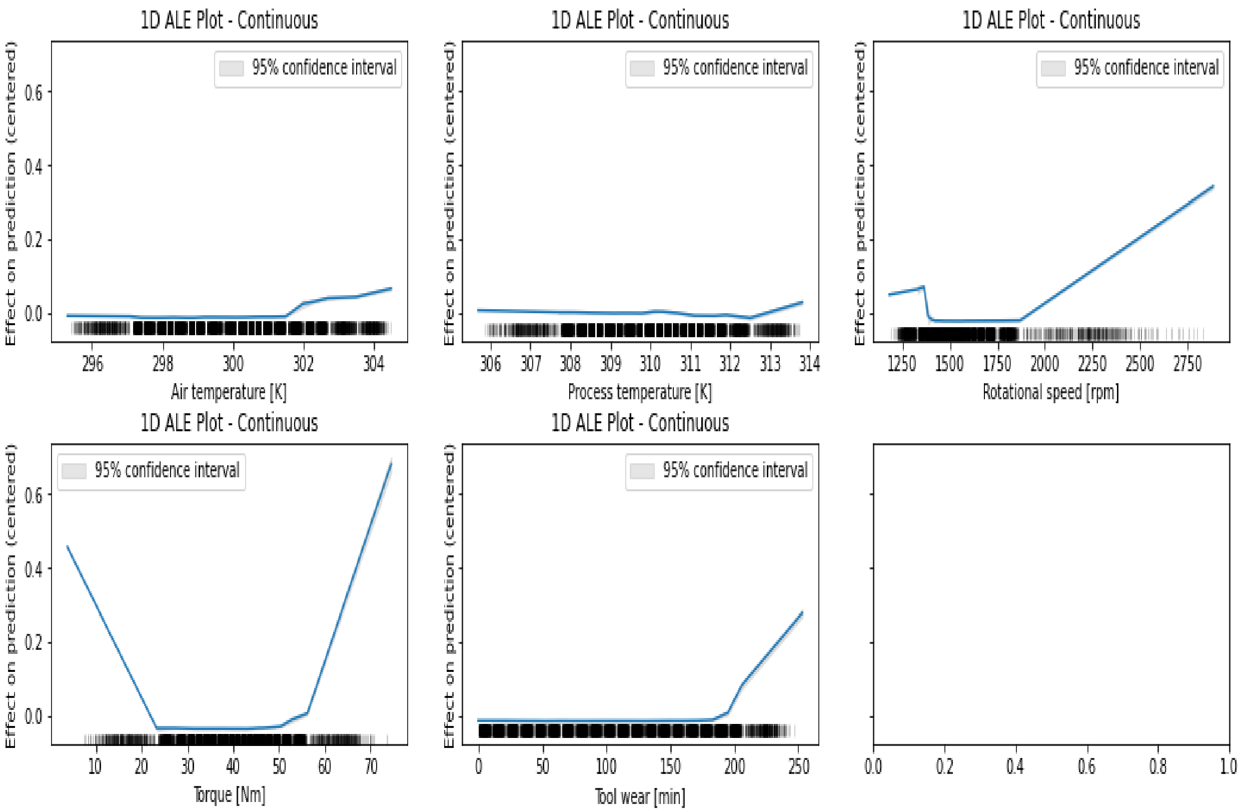


We can clearly see that we are able to accurately predict 89 True Positive and 2401 True Negative cases in the Test data set.

Let’s Interpret the results of Random Forrest using ALE, ICE and PDP to evaluate the relationship of features with target variables.

ALE

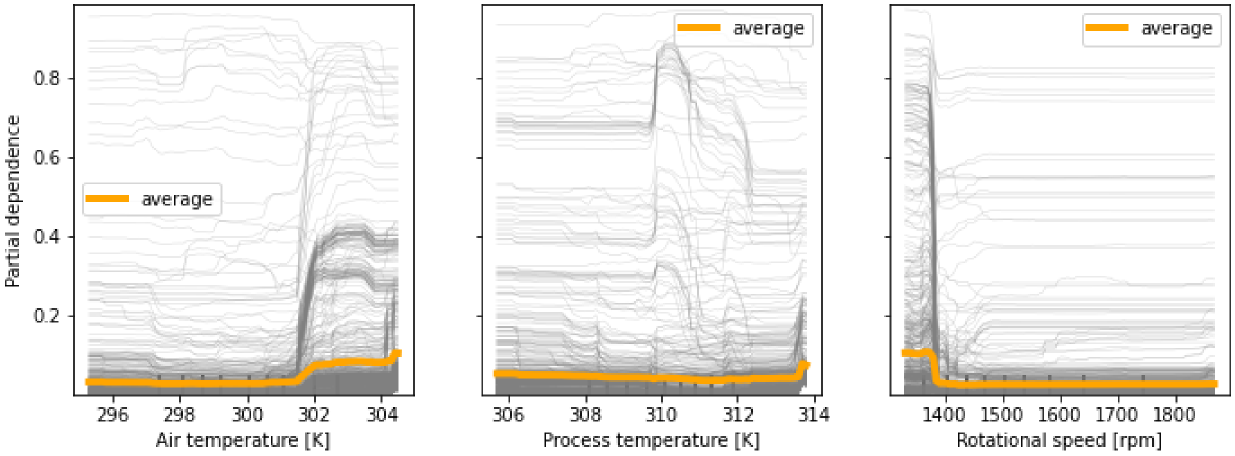
Accumulated local effects33 describe how features influence the prediction of a machine learning model on average

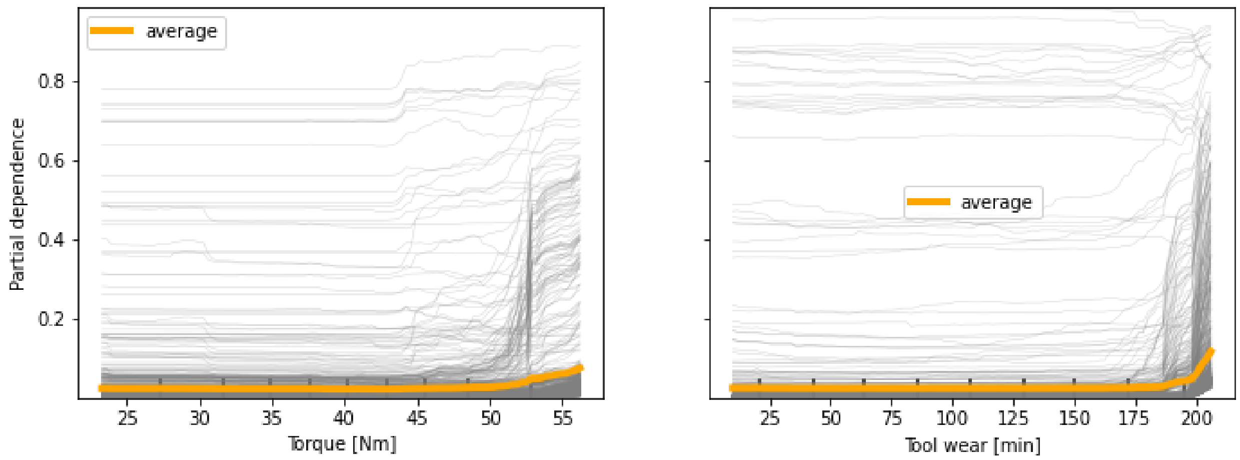


Looking at the ALE plot we can clearly see that Air Temperature above 301 K and Rotational speed above 1800 rpm Tool wear above 180 and Torque above 55 and below 22 increases the probability of machine failure.

ICE

An individual conditional expectation (ICE) plot shows the dependence between the target function and a feature of interest.

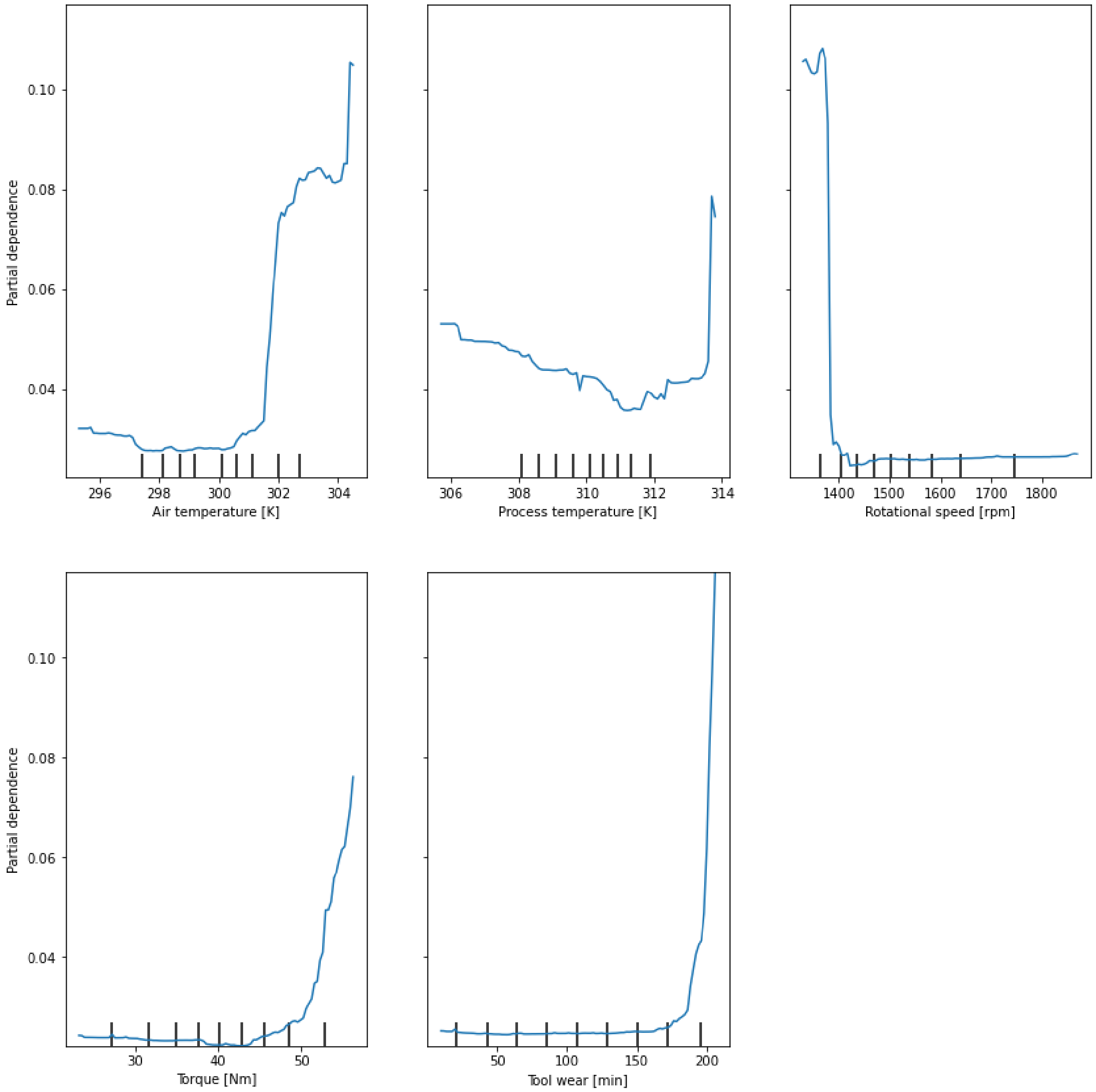




Looking at the ICE plot we can clearly see that Air Temperature, Torque and Tool wear, process temperature causes failure as they increase above a threshold

PDP

Partial dependence plots show the dependence between the target function and a set of features of interest, marginalizing over the values of all other features.



Looking at the plot we can clearly see that Air Temperature, Torque and Tool wear, process temperature causes failure as they increase.

Conclusion

Looking at the plot all plots signify similar trends, and it is alignment with our initial hypothesis that increase in air temperature and Torque above a threshold causes machine failure is also captured by the model.

Data Subset

We will further try to evaluate our using the same subset of features that we used for logistic regression.

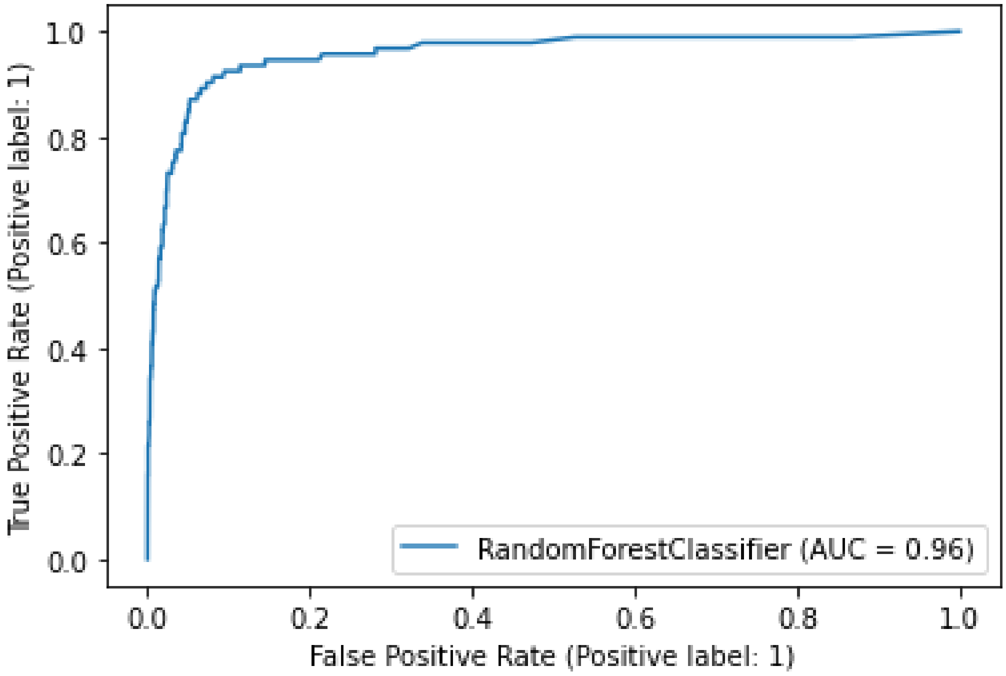
s

Torque, Rotational Speed, Air Temperature, Process Temperature and Tool Wear are important features.

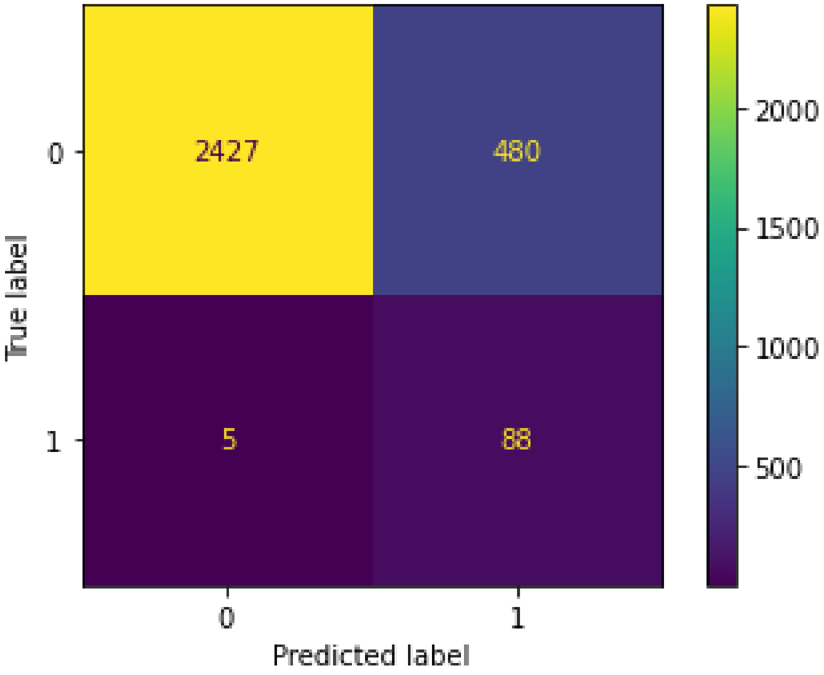
We will be using these features and retraining our model.

The accuracy of the model improves slightly in this case 83.5% for the train set and 83.8 for the test set.

The ROC curve remains the same



The confusion matrix



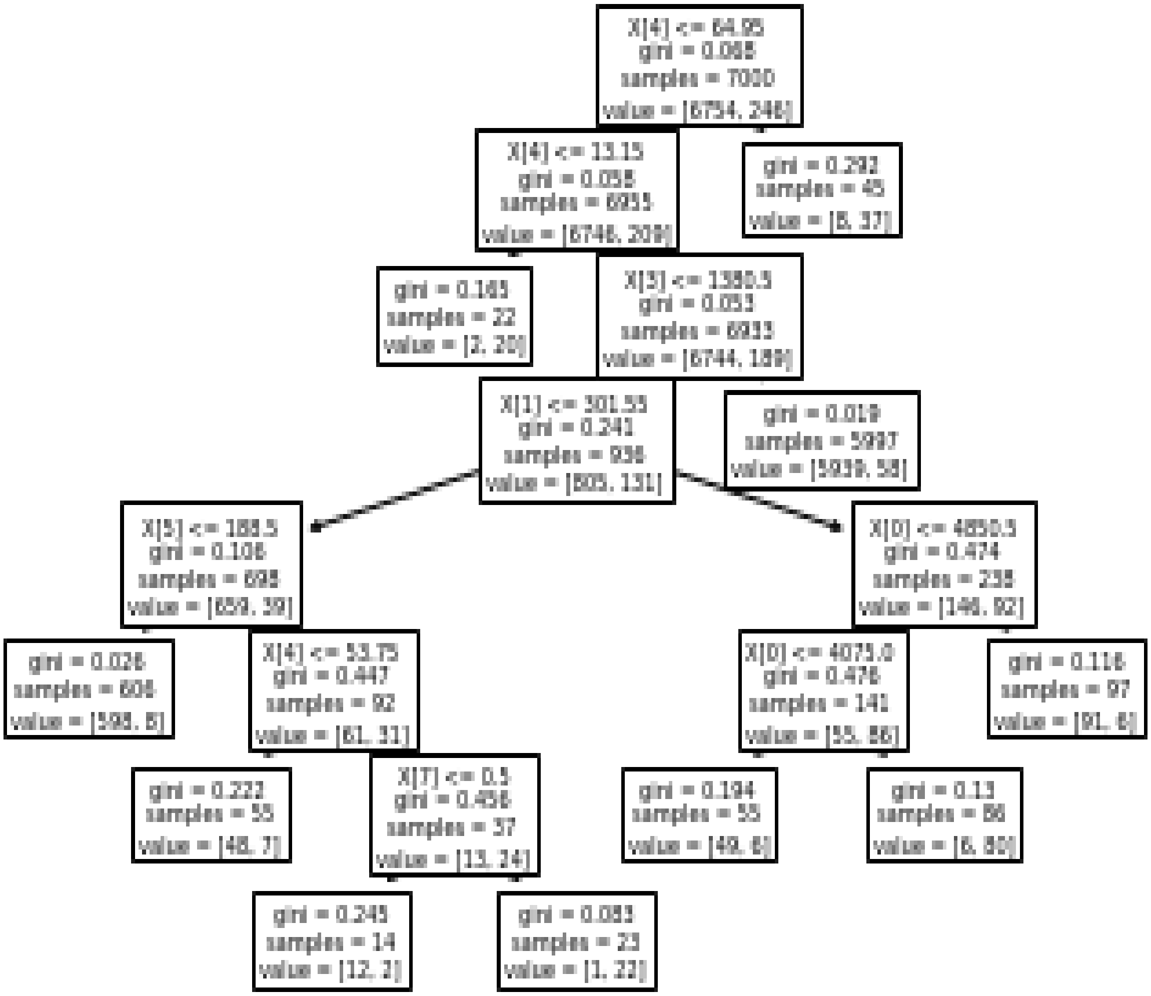
Thus, we can see that there is an improvement in predicting True False cases compared to the previous Model.

Decision Tree

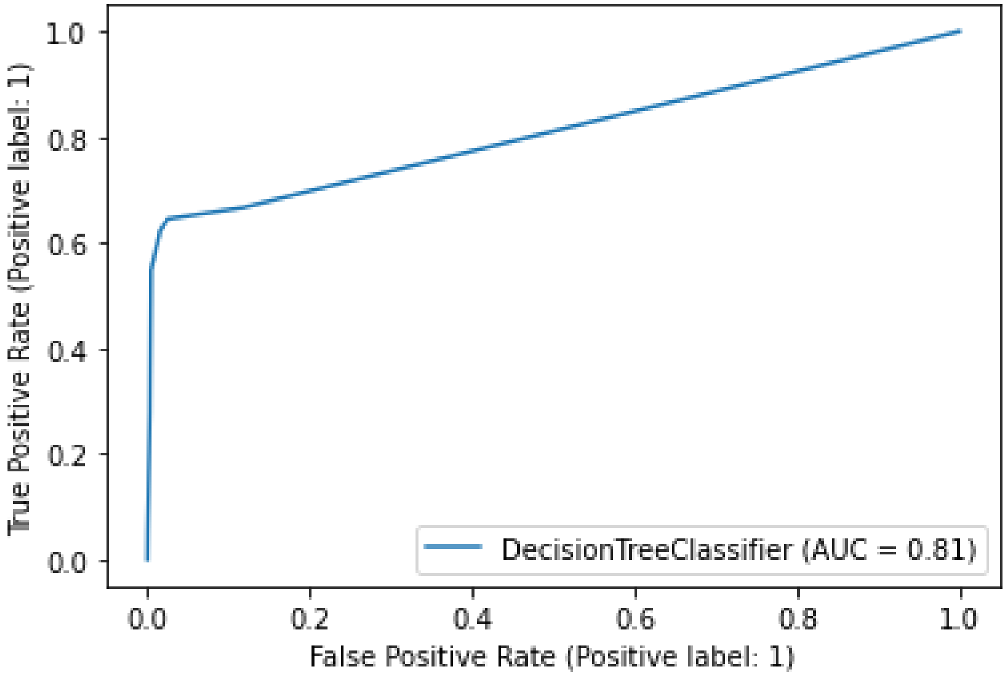
The model was trained with max_depth = 15, max_leaf_nodes = 10, random_state=42 .

The accuracy of train was 95.95% while the accuracy for the test was 96.46%.

The decision Tree classifier for the same is as follows:

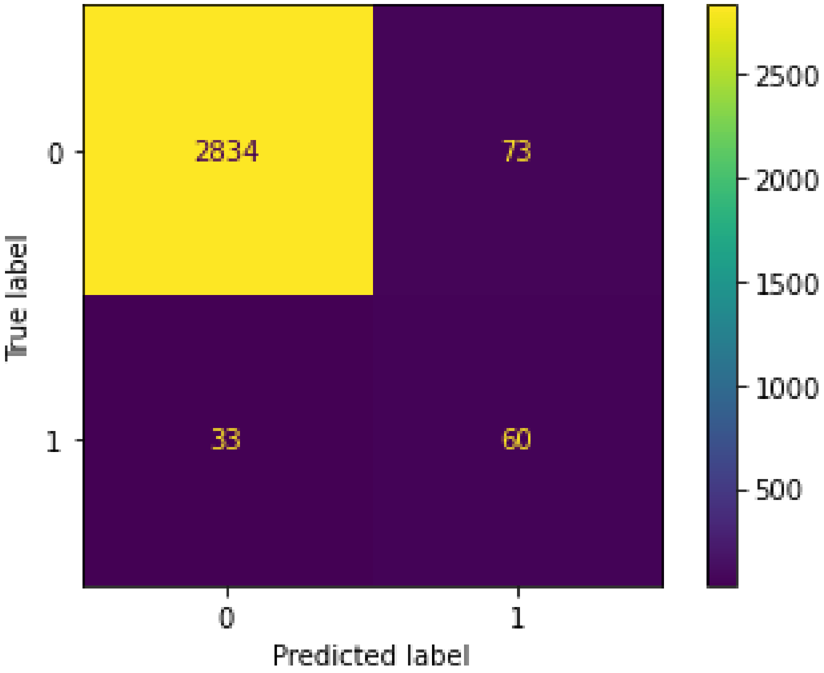


The ROC curve for the model is as follows:



We have an AUC of 81% in this case thus signifying that we are able to cover 81% of our data using this model.

The confusion matrix for the same is as follows:

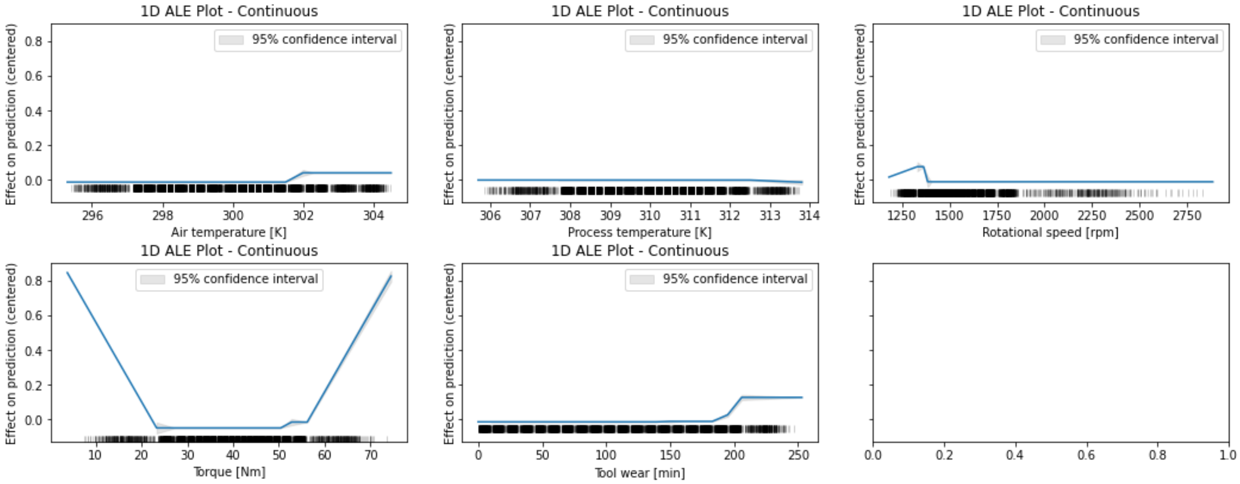


We can clearly see that we are able to accurately predict 60 True Positive and 2834 True Negative cases in the Test data set.

Let’s Interpret the results of Random Forrest using ALE, ICE and PDP to evaluate the relationship of features with target variables.

ALE

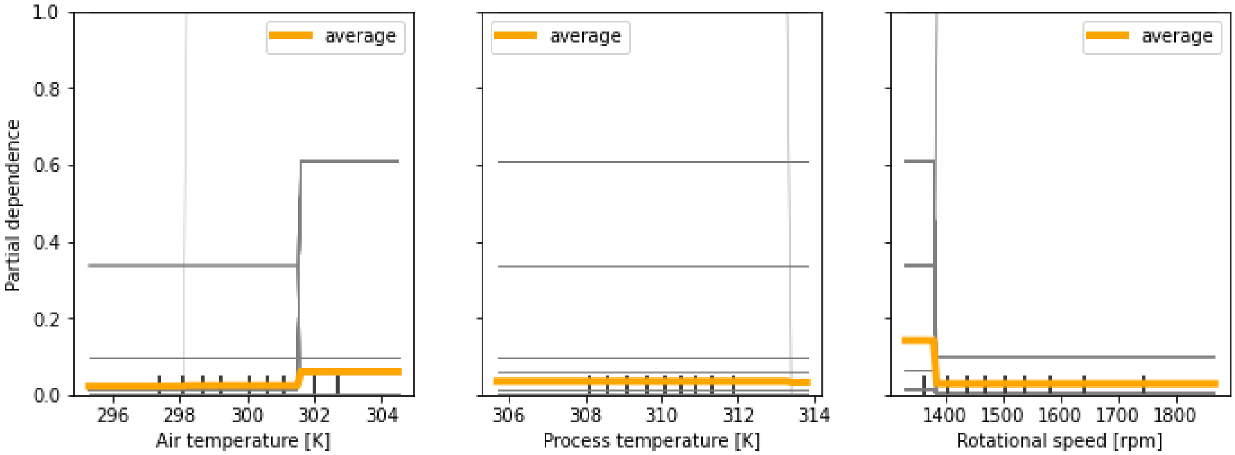
Accumulated local effects33 describe how features influence the prediction of a machine learning model on average

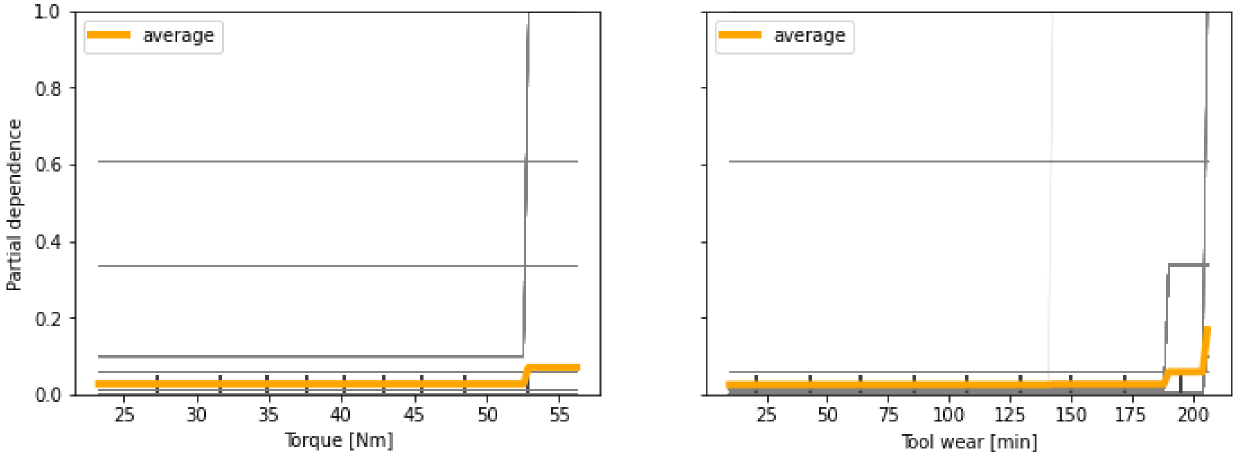


Looking at the ALE plot we can clearly see that Air Temperature above 301 K Tool wear above 180 and Torque above 55 and below 22 increases the probability of machine failure.

ICE

An individual conditional expectation (ICE) plot shows the dependence between the target function and a feature of interest.

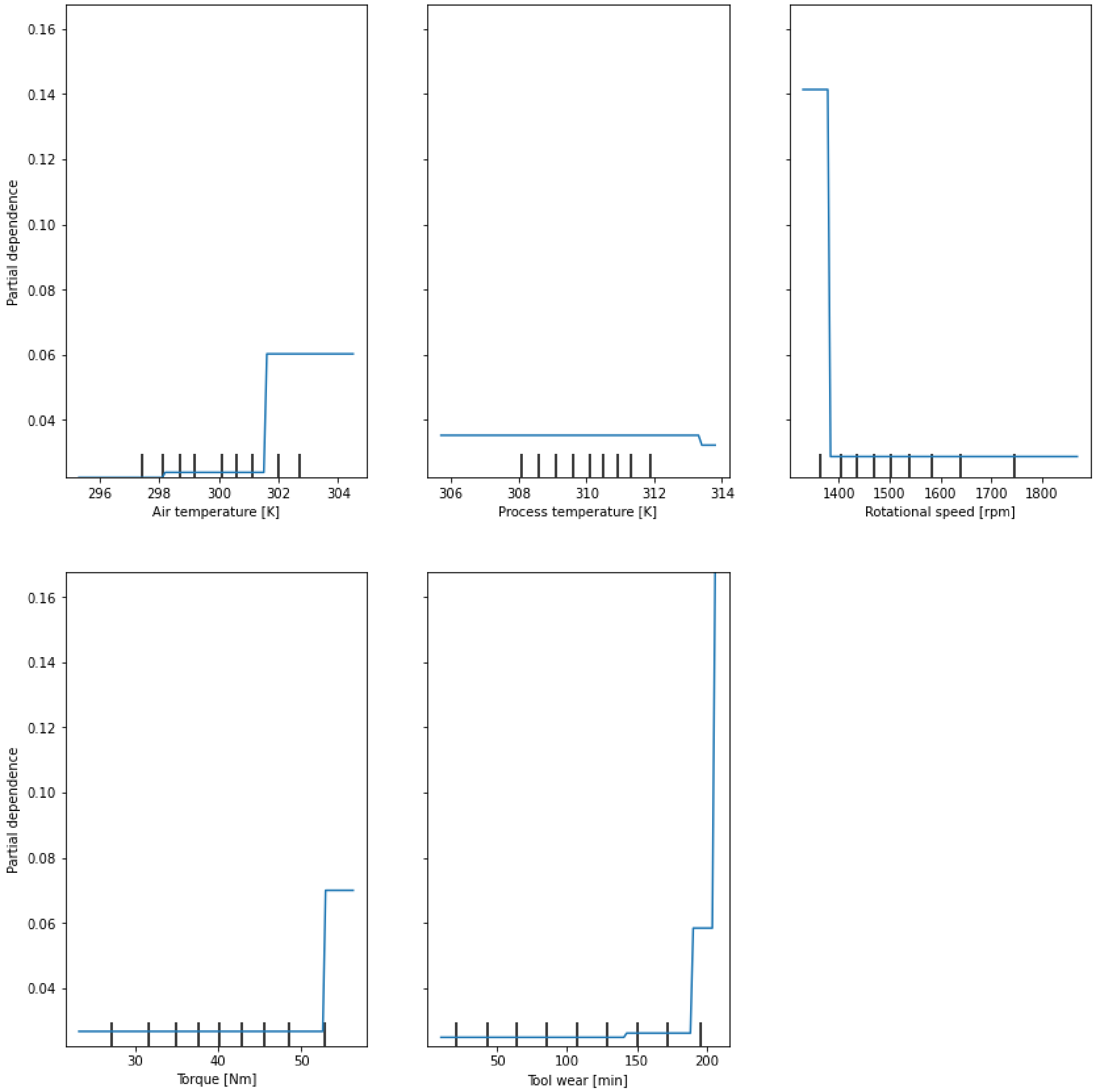




Looking at the ICE plot we can clearly see that Air Temperature, Torque and Tool wear causes failure as they increase above a threshold

PDP

Partial dependence plots show the dependence between the target function and a set of features of interest, marginalizing over the values of all other features.



Looking at the plot we can clearly see that Air Temperature, Torque and Tool wear, process temperature causes failure as they increase.

Conclusion

Looking at the plot all plots signify similar trends, and it is alignment with our initial hypothesis that increase in air temperature and Torque above a threshold causes machine failure is also captured by the model.

Conclusion

Random Forrest model seems to predicting the True Positive with high accuracy and thus would be our primary choice for the model. Looking at the results from PDP, ALE, and ICE for Random Forrest model are able to see that model also captures the data pattern we discussed earlier in Exploratory Data Analysis. Looking at the at the ALE plot we can clearly see that Air Temperature above 301 K and Rotational speed above 1800 rpm Tool wear above 180 and Torque above 55 and below 22 increases the probability of machine failure which is in line with our primary investigation results; We also notice that Air Temperature, Torque and Tool wear, process temperature causes failure as they increase above a threshold. All these findings are in correlation with our primary analysis we performed. Although Decision Tree also provides us similar insights as random Forrest with respect to ALE, ICE and PDP plots the accuracy of predicting True Positive is lower compared to Random Forrest. In case of logistic regression although model has better True positive accuracy the model is not able to capture the thresholds with respect to ALE, ICE and PDP plots and thus implying a direct or inverse relation with the target variable which in this case is the machine failure. It is successful in telling us our prior assumptions as air temperature increases failure increases, as torque increases failure increases and as rotational speed increases failure increases is in line with our findings.