**Capstone Project**

**Machine Failure Prediction**

**By**

**Antony Joseph**

**Date: 04/06/2021**

Table of Contents

[Problem statement 3](#_Toc73712901)

[Industry/ domain 4](#_Toc73712902)

[Stakeholders 4](#_Toc73712903)

[Business question 4](#_Toc73712904)

[Data question 5](#_Toc73712905)

[Data 5](#_Toc73712906)

[Data science process 6](#_Toc73712907)

[Data analysis 6](#_Toc73712908)

[Modelling 12](#_Toc73712909)

[Outcomes 14](#_Toc73712910)

[Implementation 14](#_Toc73712911)

[Data answer 14](#_Toc73712912)

[Business answer 14](#_Toc73712913)

[Response to stakeholders 15](#_Toc73712914)

[End-to-end solution 15](#_Toc73712915)

[References 15](#_Toc73712916)

# Problem statement

Mechanical automated machines are used in the logistics industry to transport, store, retrieve and package products. Automated machines are made of moving parts that needs to be serviced (Preventive Maintenance) during planned downtime. It is an important task for the stability of machine operation. Preventive maintenance has evolved from reactive to pro-active. The next phase is predictive maintenance. Reactive maintenance is when the machine is fixed when it breakdown. Periodic maintenance is to service the equipment after certain time. Pro-active maintenance is to detect signs of failures and do servicing. Doing too many maintenances can result in overspending, wear and tear, and over utilising resources and inventory. On the other hand, doing too less preventive maintenance can results in unplanned breakdowns and can be too costly. Predicting failures accurately and doing maintenance at the right time will cut unnecessary downtime, optimise productivity and inventory, and save cost to the business.

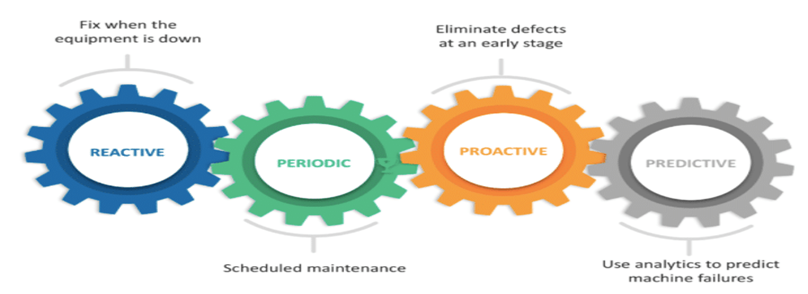


fig1: Maintenance evolution

The problem is applicable with a $600 million facility that uses highly automated machines to transport and store products from suppliers. It retrieves products based on demands from stores, packages it and deliver on time. The organisation uses a reactive/periodic maintenance-based approach. Machines still breakdown during production and causes unplanned downtime. The project is investigating the potential to predict machine failures accurately.

Currently, unplanned breakdowns are causing substantial revenue loss for the organisation. The organisation is also overspending on maintenance. Predicting failures accurately and optimising maintenance will result in value addition to the business.

fig2: Maintenance cost

Unplanned breakdowns occur during production which results in store not able to sell products to customers. This may result in losing customer trust and loyalty. The organisation wants to achieve equipment uptime and deliver products on time.

There are similar projects done with other industry and able to predict equipment failures.

# Industry/ domain

The organisation is well known in Logistics operations. It has distribution centres across Australia and supplies products to stores. It uses systems to track orders, delivery, production schedules, order completion etc. The overall value-chain is to procure products from suppliers and store. Based on store orders, retrieve products, package, and deliver to stores where consumers will buy products. The key concept of the industry is to minimize operational cost and deliver products on time. The project can be applied to other industries that uses automated machines to produce goods.

# Stakeholders

The key stakeholders for the project are Plant operation Manager and Maintenance manager. They are interested in reducing maintenance cost for the distribution centre. The stakeholders expect to see maintenance cost reduce by 10%.

# Business question

Business is always looking to reduce cost. The business question will be “How much value can be added by predicting equipment failure?” The business estimates a total of 10% reduction in unplanned downtime and achieve cost savings of ~ $4.2M.hat e the implications of false positives or false negatives?

# Data question

The data question for the problem will be “What data variables are needed to accurately predict machine failures?” Machines fulfil its objective by moving parts and utilising energy. Therefore, operating parameters of the machines would be essential. Every machine is different from the other in terms of age, service records and failures. The data that would assist in this project will be – Operating parameters, maintenance records, age, failures.

# Data

The data was sourced from Kaggle since the data from the machines were not available. The data features sourced has close resemblances with actual data.

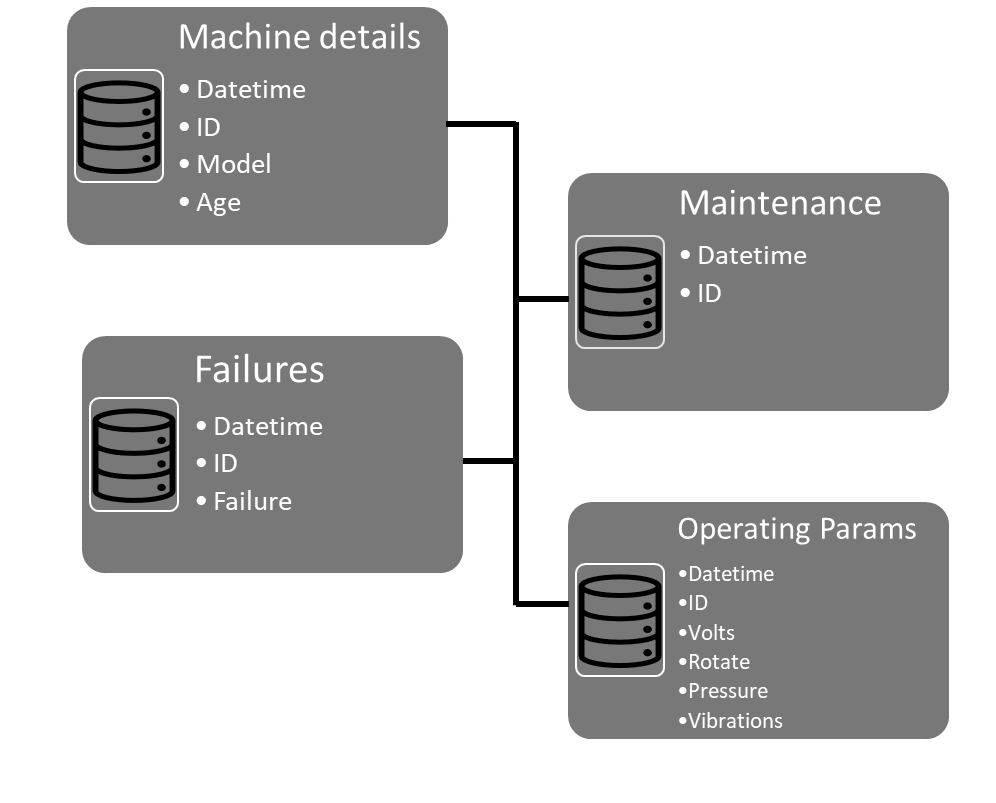


fig 3: data structure

Fig 3 shows the data structure sourced from Kaggle. The data consist of hourly data for 100 machines for 1 year. The data was verified by Kaggle. It is available to download. (https://www.kaggle.com/arnabbiswas1/microsoft-azure-predictive-maintenance)

# Data science process

## Data analysis

The datasets were 4 tables - Machine details, Operating parameters, Failure logs, Maintenance logs. Operating parameters were continuous hourly reading for 1 year. Machine details were static data. Failure logs and Maintenance logs were data captured when the event occurred (not continuous time frame). The tables were loaded into Jupyter notes as dataframes. All dataframe did not had null values. As the operating parameter tables data were checked for continuity. The tables were then merged based on the timeline.

The target variable – failures in the dataset was highly imbalanced. The target variable was less than 1% of the total dataset.

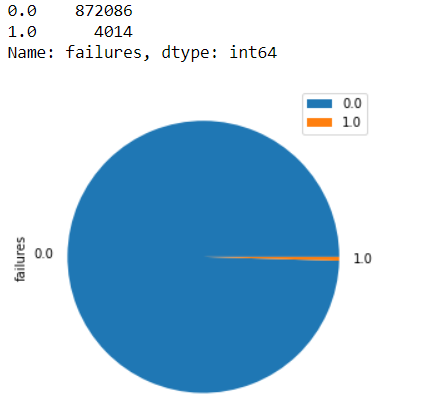


fig4: Unbalanced data

Age distribution of the machines ranged from 0 – 20 years.

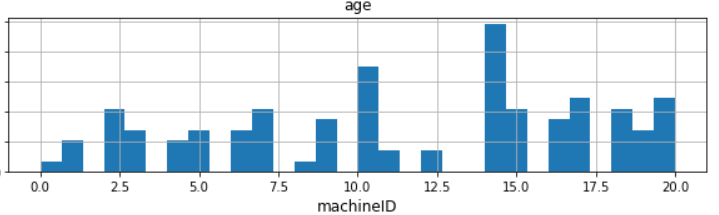


Fig: 5 Machine age

The operating parameters showed normal distribution.

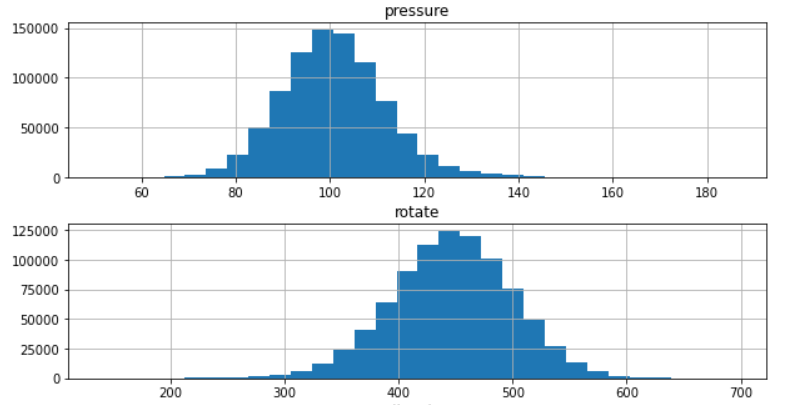


fig 6: Operating parameters

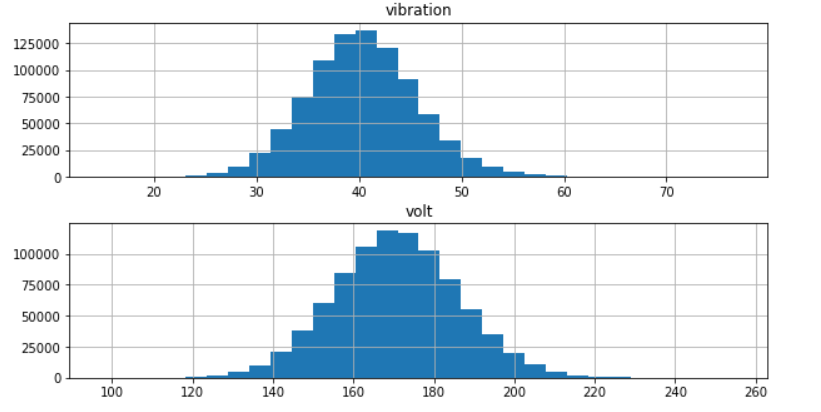


fig7: Operating parameters

Total number of failures for machines varied from 20 to 60.

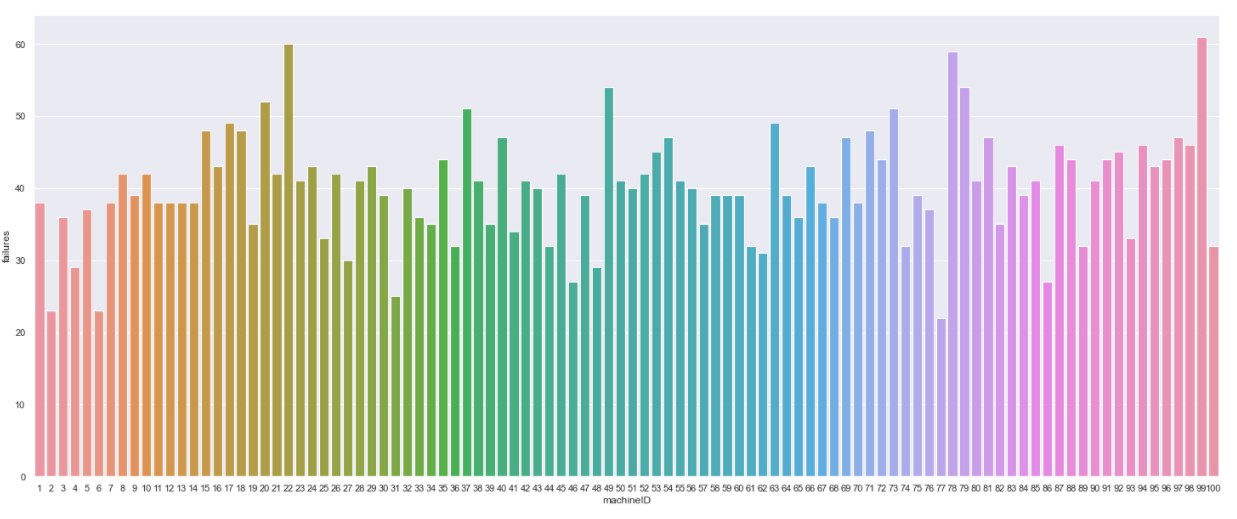


fig 8: Total failures based on machine ID

Maintenance activities for machines also varied.

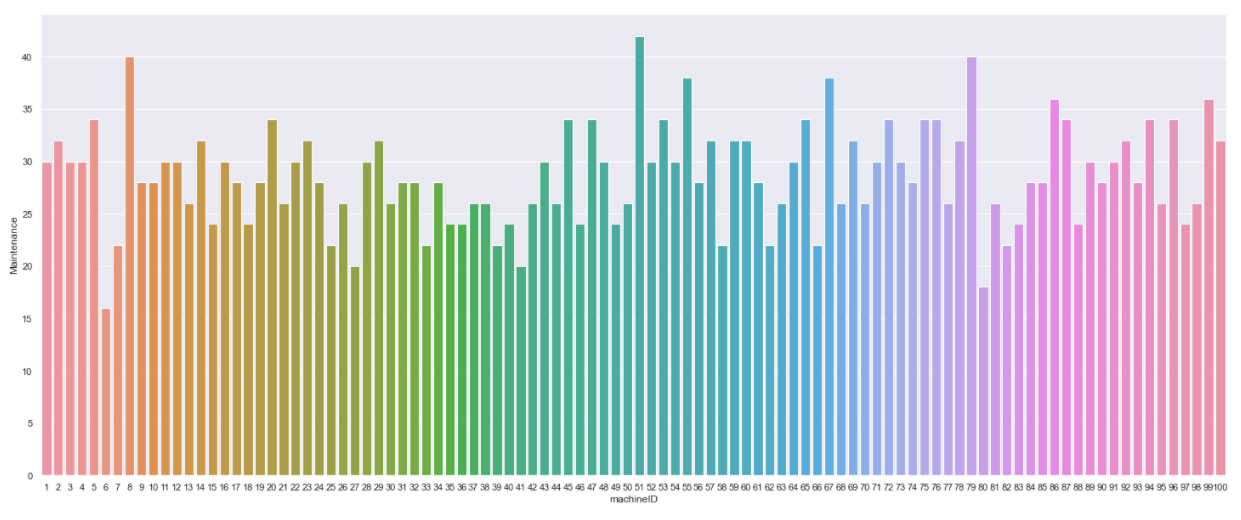


fig 9: Total maintenance activities based on machine ID

Since failures and other features of each machines varied greatly, machine number 1 was chosen for EDA. Target variable for Machine 1 was also unbalanced.

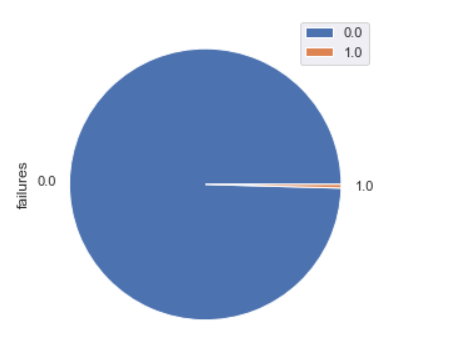


fig 10: machine #1 target variable

Plotted line graph comparing the failures and operating parameters. It showed some correlation with target variable.



fig 11: machine #1 failures and operating parameters

Checked if the failures occurred during particular month, day or day of week. No failures happened during certain month and days of month. It seems like the features were correlated to the target variable.

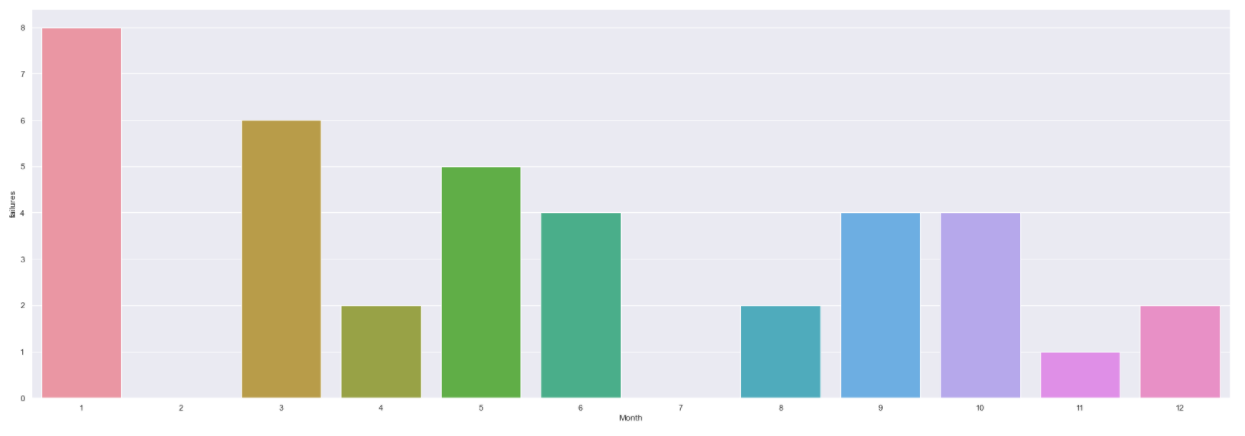


fig 12: machine #1 failures based on month

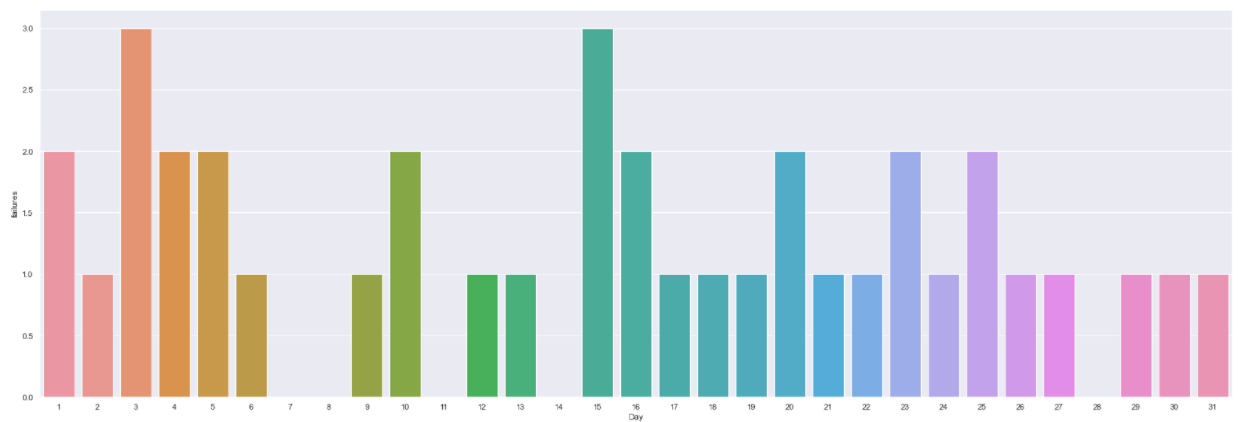


fig 13: machine #1 failures based on days of month.

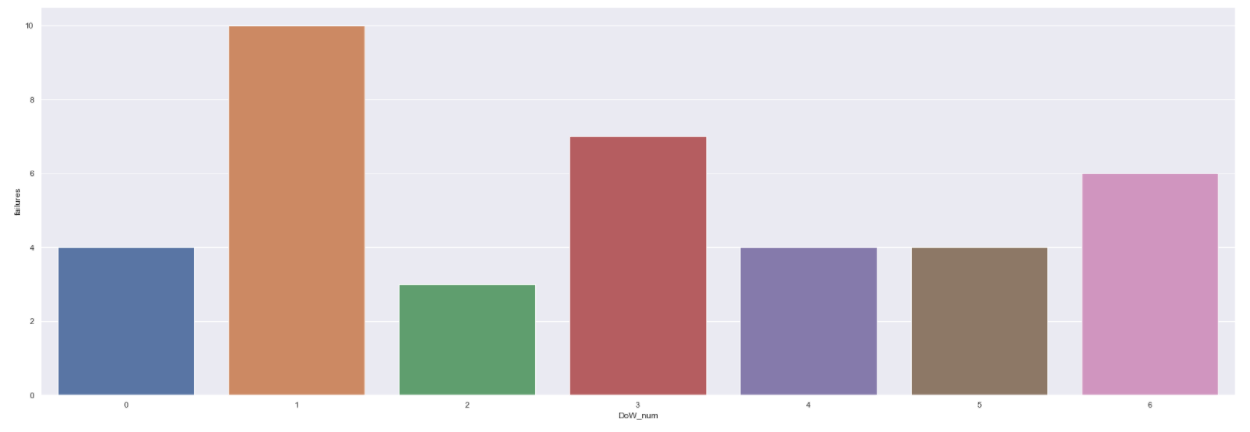


fig 14: machine #1 failures based on days of week.

Plotted Pearson correlation heatmap to find out the correlation with the features.

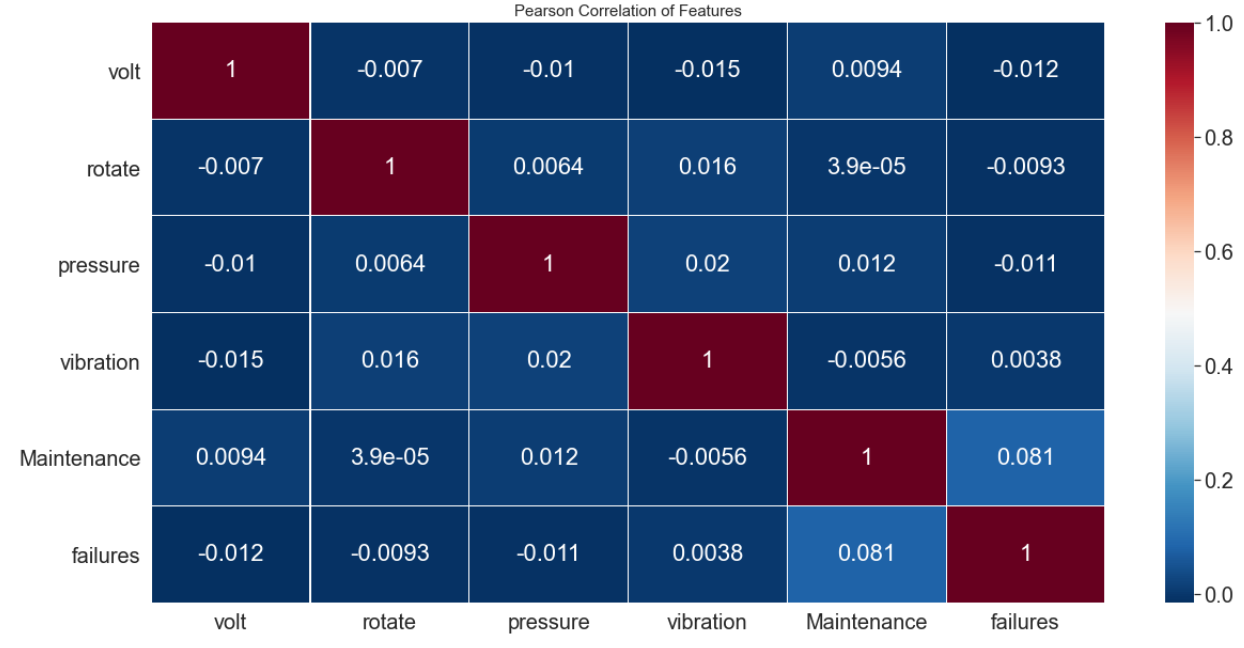


fig 15: machine #1 correlation heatmap.

Initial exploratory data analysis showed less correlation with the target variable(failures).

Performed feature engineering on the dataset. Added 14 days moving average variable of the operating parameters and maintenance to the target variable. A feature to estimate the time elapsed were also added to the dataset. In order to predict failure 7 days in advance the failure window was increased from 1 hours to 7 days. This was to give time for the business to plan and prepare resources for maintenance.

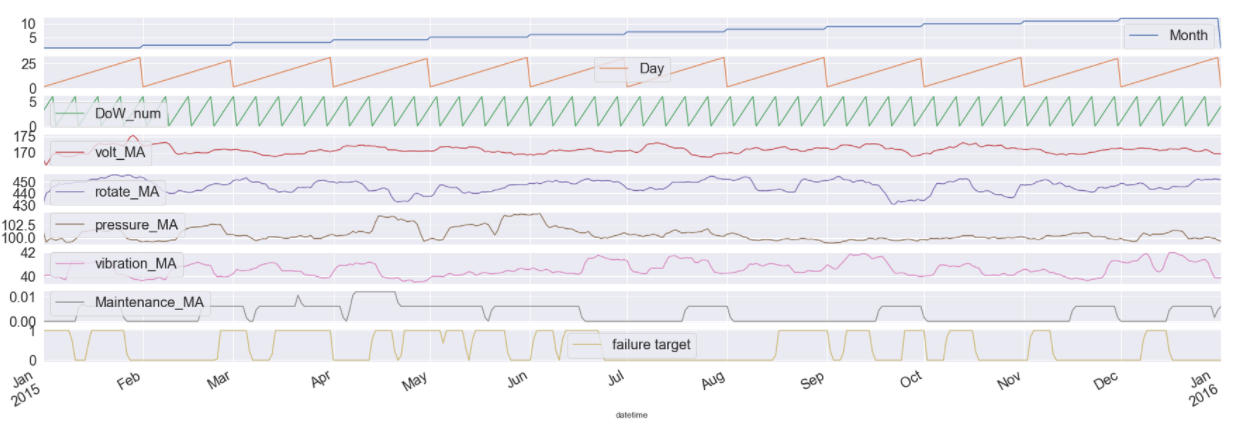


fig16: machine #1 failure vs feature engineered parameters.

A Pearson correlation map was plotted to see how the variables were correlated. Variables, time\_elapsed, Month, volt\_MA, rotate\_MA, pressure\_MA, vibration\_MA, Maintenance\_MA showed correlation with the target variable.

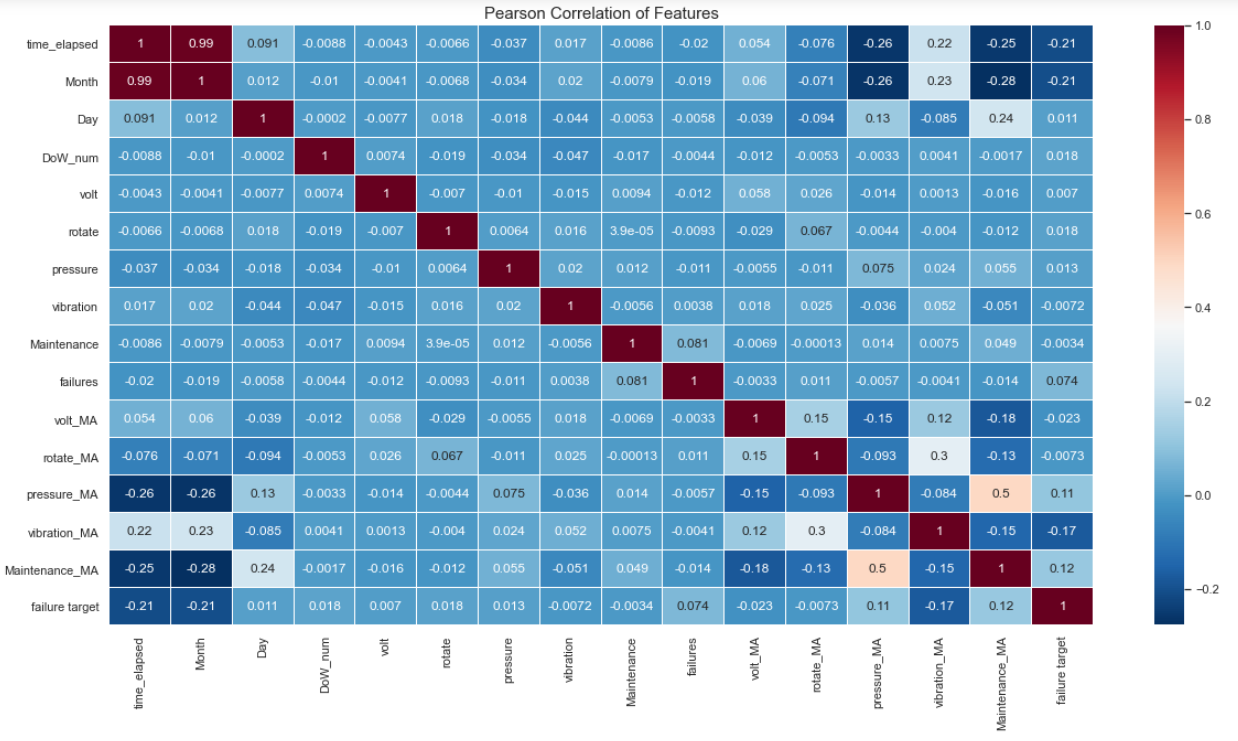


fig 17: machine #1 correlation map

Data science process was done in Jupyter notebooks. EDA process is able to reusable to process future data.

## 

## Modelling

The features selected EDA were Month, pressure\_MA, vibration\_MA, Maintenance\_MA. The target variable is failure\_target. Some of the features has more correlation with other features compared with the target variable. Features were selected using correlation map. 3 models were used to predict failures. The models were - Decision tree, Random Forest, Long short-term memory (RNN). To train the Decision tree and Random forest models took less time. LSTM model with 10 Epocs took ~2 min and 30secs. Initially the models did not perform as expected. Some of the features were eliminated by trial-and-error method. The features were - 'time\_elapsed', 'volt\_MA', 'rotate\_MA'.Recall. Precision, Accuracy, Recall, ROC AUC were used to determine the model performance. Below shows figures shows the models performance metrics.

Decision-Tree Classifier

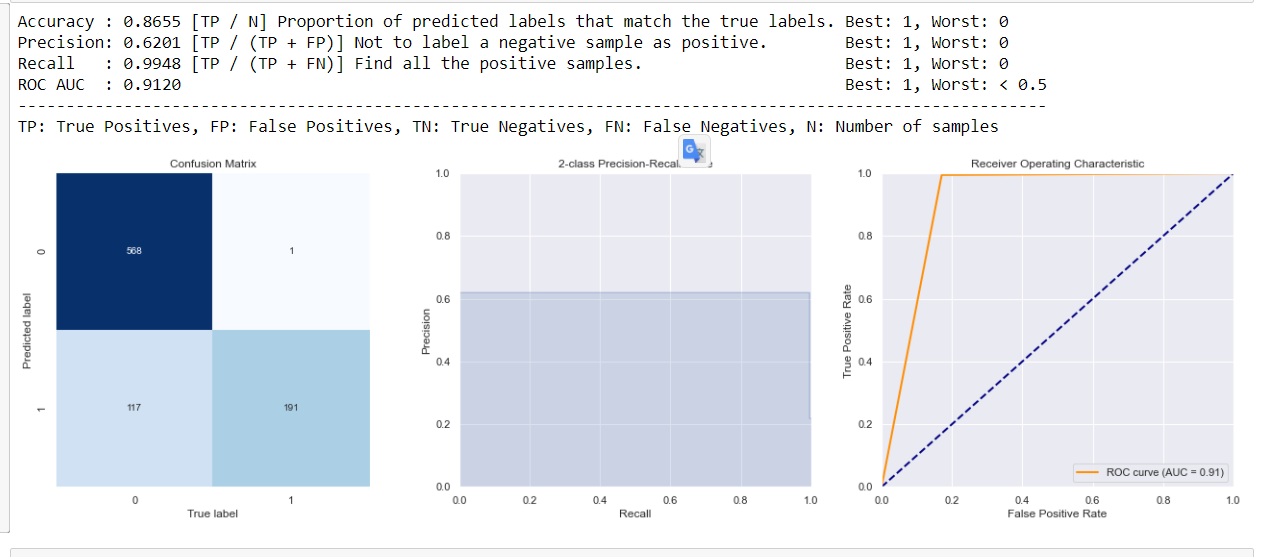


fig18: Decision tree classifier performance

Random Forest Classifier

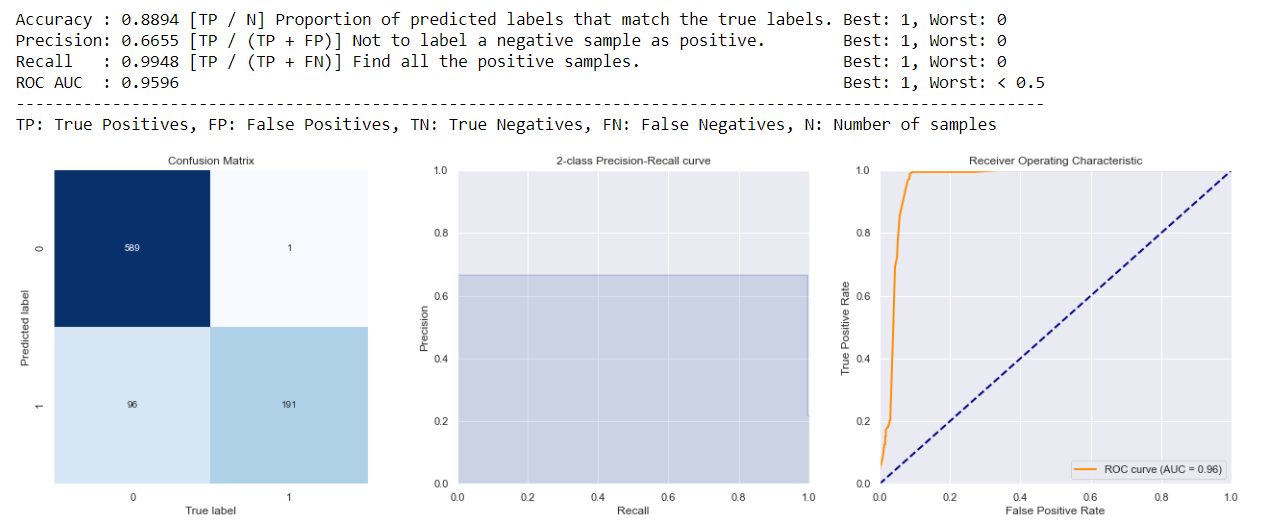


fig19: Random Forest classifier performance

Long Sort Term Memory(LSTM) RNN

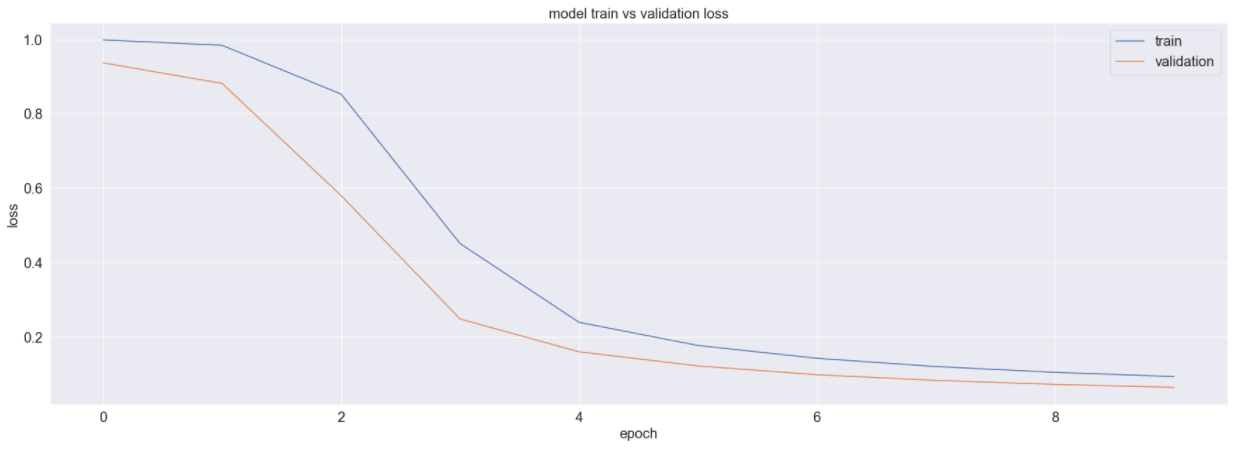


fig20: train vs validation loss

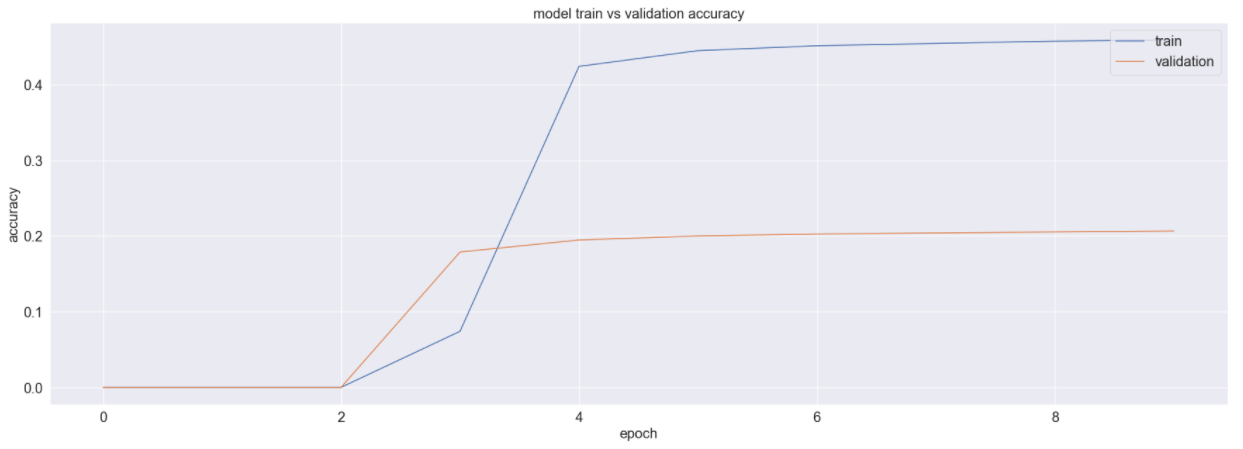


fig21: train vs validation accuracy

The loss and accuracy graphs show the model is not overfitted.

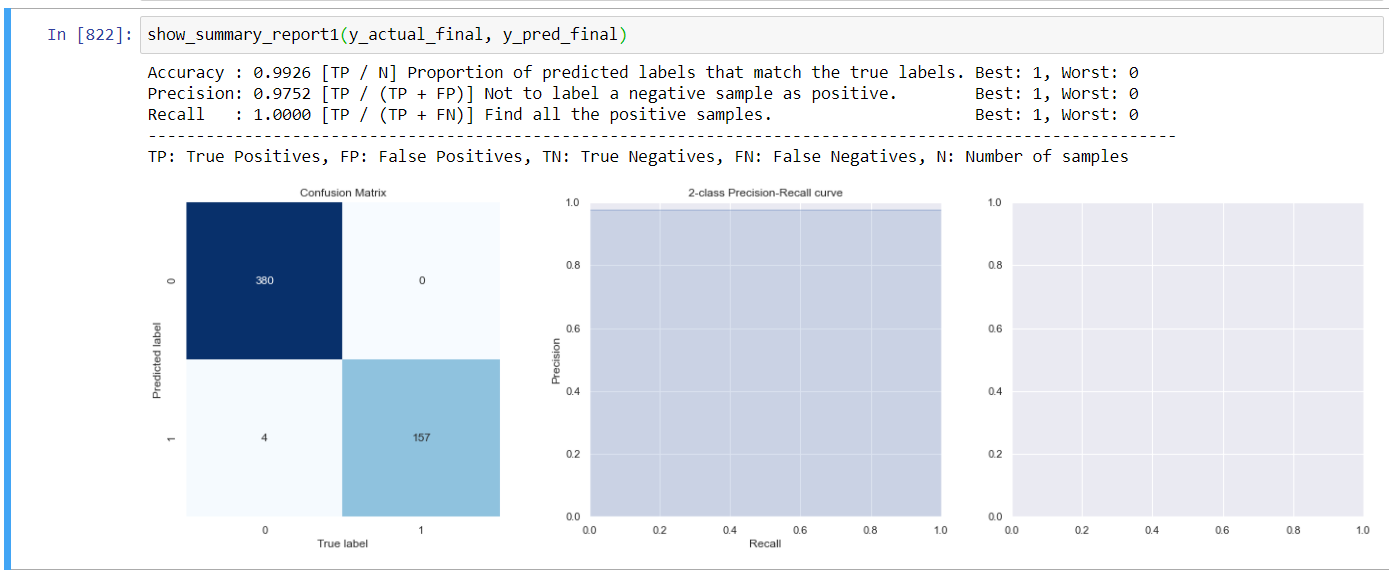


Fig20: LSTM RNN performance

LSTM model were selected as it has the best performance compared with other models with 99 % Accuracy, 97% precision and 100% recall.

## Outcomes

The EDA and feature engineering process led to identify correlated features. LSTM model gave better results than other models. The process confirms the possibility to predict machine failure with high accuracy.

## Implementation

The considerations to implement the solution will be to develop systems for data acquisition, storage, and program execution. Determine forecasting period with the business. Develop Interface for data visualisation.

# Data answer

The dataset had enough features and volume of data to accurately predict machine failures. Confidence level is greater than 90%

# Business answer

The business question was answered satisfactorily. The proposed solution provides over 90% accuracy to predict machine failure.

# Response to stakeholders

The project is promising and expected to achieve its objective in optimizing machine maintenance. It is recommended to proceed with implementing the project.

# End-to-end solution

To implement the end-to-end solution the organisation needs to have the necessary infrastructure for data acquisition and storage, systems to run the model and display the machine status.

# References

<https://www.kaggle.com/arnabbiswas1/microsoft-azure-predictive-maintenance>

<https://towardsdatascience.com/why-and-how-to-use-merge-with-pandas-in-python-548600f7e738>

<https://seaborn.pydata.org/introduction.html>

<https://matplotlib.org/stable/tutorials/introductory/pyplot.html>

<https://machinelearningmastery.com/>

<https://medium.com/swlh/machine-learning-for-equipment-failure-prediction-and-predictive-maintenance-pm-e72b1ce42da1>