Machine Learning Nanodegree – Capstone Project Proposal

Sanjay Soni,

August 22nd, 2019

PREDICTIVE MAINTENANCE USING MACHINE LEARNING

1. Domain Background

In any asset heavy industry like manufacturing, utilities, energy generation and distribution, maintaining the optimal health of the assets and machines is a critical aspect of the operations. If the assets are not in healthy state it not only has impact on the performance and productivity, but also may lead to serious HSEC issues (health, safety, environment and compliance)

Maintenance professionals have always faced this conundrum of creating effective maintenance strategy and planning with a target to **maximize the uptime** and **minimize the downtime** whilst **maintain the healthy state** of machine. It can be quite difficult to determine how often and when a machine should be taken offline to be services as well as weight the risk of machine downtime while attempting to maximize the uptime through early maintenance and replacements

Traditionally they have employed various techniques – quantitative and qualitative in order to identify the fault modes and plan the maintenance of the various machines in the manufacturing facilities.

Components of maintenance program often fall in one of the four categories below:

- Reactive Maintenance
- Planned Maintenance
- Proactive Maintenance
- Predictive Maintenance

Table below provides a comparative overview of the approaches

	Benefits	Challenges
Reactive	• Maximum utilization of tooling or machine components	• Potentially greater damage to machine beyond failed part • Unplanned downtime • Higher maintenance costs
Planned	• less likelihood of broken machinery • less unplanned downtime • more cost effective then reactive	• increased replacement costs over time • need for additional spare parts inventory • increased planned downtime
Proactive	• longer lifespan of equipment • decreased downtime – planned and unplanned • more cost effective than run-to-failure or planned maintenance • lower spare parts inventory	• ongoing maintenance and monitoring • need for organizational changes • increase training

The **Predictive Maintenance (PdM)** while not a new concept, has now become possible given the vast number of smart and connected technologies in use in the manufacturing industry that unite the digital and physical assets.

In PdM, data gathered from connected, smart machines and equipment can predict when and where failures could occur, potentially maximizing parts' efficiency and minimizing unnecessary downtime.

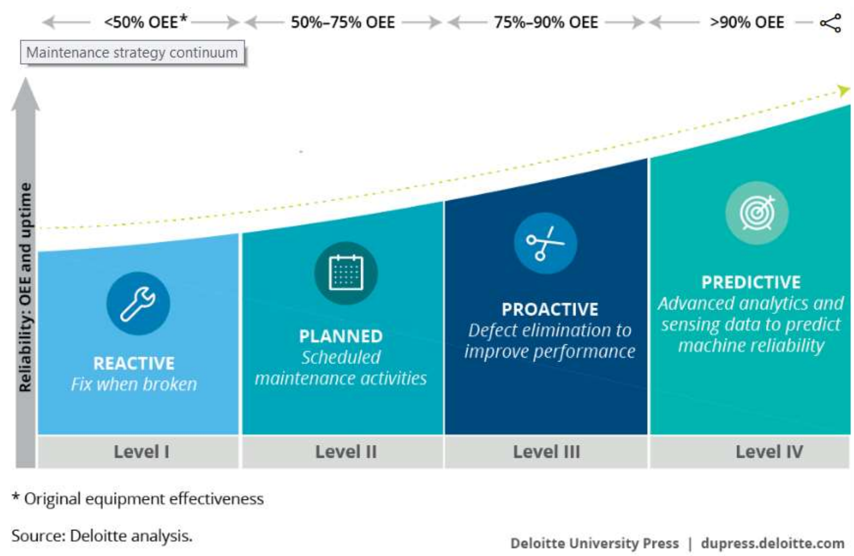
Unplanned outages impact the BOTTOMLINE of a Company

A recent study by Wall Street and Emerson found that unplanned downtime costs industrial manufacturers $50 billion / year. Equipment failure is the cause of 42% of this unplanned downtime

Australian Institute Research found that in 2018 Gas and Coal Power plants broke down 100 times

Considering these statistics and the fact that PdM has potential to predict the machine failures we can say that:

PdM is often the most efficient maintenance strategy available—a gold standard for which to aim.



https://www2.deloitte.com/insights/us/en/focus/industry-4-0/using-predictive-technologies-for-asset-maintenance.html

In this capstone I will look at different ways to approach the Predictive Maintenance modelling.

2. Problem Statement

As part of this capstone we will look at how to start and evolve on this journey of the Predictive Maintenance.

From machine learning problem statement point of view, we will formulate multiple problem statements which taken together will cater for the holistic predictive maintenance journey as explained below.

Data driven prognostic (predicting failures) faces the challenge of lack of data for run-to-failure events. Most of the time the data contains the fault signature for growing fault pattern, but no or very little data capture of fault evolution until failure.

So, as a starting point machine learning based predictive maintenance procedure should cater for these scenario’s by clustering the sensor data from machines into different operation groups - Normal, Warning and Critical.

Problem Statement 1:

To cluster the sensor data into different operation groups – Normal, Warning and Critical where,

- Normal: no maintenance required and machines are running in healthy state
- Warning: machine is still operational and would require maintenance in near future
- Critical: machine is operational but in critical state and require maintenance urgently

We will address this problem as unsupervised learning problem to cluster the data into three groups.

NOTE: We will use the dataset by ignoring the machine fault information to perform this clustering. We will then overlay the machine fault information to assess how effective the clustering has been.

In the scenarios where the data for fault evolution all the way to the occurrence of the fault is available, we can extend the approach above with the supervised machine learning. The problem can be framed either as a regression machine learning problem, or as a classification problem (binary, as well as a multi-class classification)

Problem Statement 2:

Predict the time-to-failure (TTF) or remaining-useful-life (RUL) for a machine based on the historical data available for TTF / RUL. The measure TTF / RUL can be any units for e.g. days, weeks or number of cycles etc.

We will address this problem as supervised regression machine learning problem.

Problem Statement 3:

a) Predict whether a machine will fail within certain time frame (e.g. days).
b) Predict whether a machine will fail in different time windows, for e.g.,
    - Fails in window [1, w0]
    - Fails in window [w0+1, w1]
    - Not fail within w1 days

Problem 3a) will be addressed a binary classification problem, and
Problem 3b) will be addressed as a multi-class classification problem

3. Datasets and Inputs

For this project we will use the Turbofan Engine Degradation Simulation Data Set, provided by NASA's Prognostics Center

http://ti.arc.nasa.gov/project/prognostic-data-repository

Turbofan Engine Degradation Simulation Data Set

Description	Engine degradation simulation was carried out using C-MAPSS. Four different were sets simulated under different combinations of operational conditions and fault modes. Records several sensor channels to characterize fault evolution. The data set was provided by the Prognostics CoE at NASA Ames.
Format	The set is in text format and has been zipped including a readme file.
Datasets	+ Download Turbofan Engine Degradation Simulation Data Set (38919 downloads)
Dataset Citation	A. Saxena and K. Goebel (2008). "Turbofan Engine Degradation Simulation Data Set", NASA Ames Prognostics Data Repository (http://ti.arc.nasa.gov/project/prognostic-data-repository), NASA Ames Research Center, Moffett Field, CA

From the dataset we will use the files for FD001 data set. This data set consists of training set and test set.

Each of these files is a text file with 26 columns of numbers, separated by spaces. Each row is a snapshot of data taken during a single operational cycle, each column is a different variable. The columns correspond to:

1) unit number
2) time, in cycles
3) operational setting 1
4) operational setting 2
5) operational setting 3
6) sensor measurement 1
7) sensor measurement 2

...

26) sensor measurement 21

In the data series the engine is operating normally at the start of each time series, and develops a fault at some point during the series.

- In training set, the fault grows in magnitude until the system fails.
- In test set, the time series ends sometime prior to the system failure.
    - Additionally, a vector of true Remaining Useful Life (RUL) is provided for the test data as a separate file.

The objective is to predict the number of remaining operational cycles before the failure in the test data set (note that this corresponds to problem statement 2)

For the binary and multi class classification problem we will compute the target variables (label) from the Remaining Useful Life (RUL) data as per following:

For some parameter period (time window) to be passed dynamically

- label_bc =
    - 1 if RUL is < parameter period (i.e. classified as maintenance needed)
    - else 0 (i.e. classified as maintenance not required)
- label_mc =
    - 2 if RUL < 0.5 * parameter period (Critical – immediate maintenance needed)
    - 1 if RUL < parameter period (Warning – maintenance required)
    - 0 if RUL > parameter period (Normal – normal operation, no maintenance needed)

4. Solution Statement

Solution will be split into three main sections

1) **Clustering of the Flight Data:**

First, we will do the clustering of the sensor data into three groups corresponding to Normal, Warning and Critical mode of operation of the turbine.

I will also perform the Principal Component Analysis to reduce the dimensionality of the data and visualize the three groups on the reduced dimensioned data in either 2D or 3D

We will then overlay the multi class classification labels (derived from RUL) and see how the three groups (clusters) are aligned to the clusters created by the unsupervised machine learning algorithm

2) **Regression**
For predicting the remaining-useful-life, I will apply the regression algorithms as a baseline and compare it with Decision Tree and Random Forest algorithms.

I will also generate time domain features and assess how much the model efficiency is improved for selecting the final model

3) **Classification**
For predicting whether the turbine will fail in a certain window or not, I will apply the logistics regression as a baseline and will then compare it with Decision Tree Classifier, Random Forest Classifier.

5. Benchmark Model

For regression model -> Linear Regression model will be used as a baseline model

For classification model -> Logistic Regression model will be used a baseline model

6. Evaluation Metrics

For regression model we will use multiple evaluation metrices – Root Mean Square Error, Mean Absolute Error and R^2

For classification we will use multiple evaluation metrices – Accuracy, Precision, Recall, F1 Score

As predictive maintenance may have different end goals depending on the operational context – I will also explain which metric will be suitable for which scenarios with some examples.

7. Project Design

At a high-level project will be divided into following sections:

- Data Wrangling and Data Pre-processing
- Data Exploratory Analysis
- Data Clustering
- Model Selection – Regression
- Model Selection – Binary Classification
- Model Selection – Multi Class Classification
- Summary and Next Steps

8. References

https://www2.deloitte.com/insights/us/en/focus/industry-4-0/using-predictive-technologies-for-asset-maintenance.html

https://www.sciencedirect.com/book/9780750675314/an-introduction-to-predictive-maintenance

https://www.maintworld.com/Asset-Management/Research-report-Predictive-Maintenance-4.0

http://ti.arc.nasa.gov/project/prognostic-data-repository