A REPORT

ON

PREDICTING FAILURES IN ROTARY EQUIPMENT USING MACHINE LEARNING

BY

RAGHAV KHANNA (ID NO.: 2018B4A40914P)

AT

Bharat Petroleum Corporation Limited, Mumbai

A Practice School Station of

Birla Institute of Technology and Science, Pilani

June, 2020

A REPORT

ON

PREDICTING FAILURES IN ROTARY EQUIPMENT USING MACHINE LEARNING

BY

RAGHAV KHANNA ----------2018B4A40914P----------MSc. Mathematics and B.E Mechanical Engineering

Prepared in partial fulfilment of the Practice School -1 Courses

BITS C221/BITS C231/BITS C241

AT

Bharat Petroleum Corporation Limited, Mumbai

A Practice School Station of

Birla Institute of Technology and Science, Pilani

June, 2020

Acknowledgement

A lot of effort from many different people has gone in to make this project a success. It would not have been possible without the kind support and help of my mentors, Dr. Srikanta Dinda, Mr. MB Mate and Mr. K Pradeep and who have given their valuable time and energy for this project. I would also like to thank BITS Pilani for conducting the internship and giving the students a chance at industry experience and BPCL, Mumbai for accepting me and providing me with a chance to learn how a company and its various factions work. I would also like to thank the BPCL Learning Centre, especially Mr. Karrupasamy and Mr. Durairaj who put a lot of effort into making this internship possible for me. I would once again like to thank everyone who has supported me throughout this internship and for helping me learn through their immense knowledge and practical experience.

Birla Institute of Technology and Sciences, Pilani

Rajasthan

Practice School Division

Station: Bharat Petroleum Corporation Limited, Mumbai

Centre: Bharat Petroleum Corporation Limited, Mumbai

Duration: 17th May to 30th June

Start Date: 17th May

Date of Submission: 24th June

Title of the Project: Predicting Failures in Rotary Equipment Using Machine Learning

Name of Student: Raghav Khanna

ID No. : 2018B4A40914P

Discipline: MSc. Mathematics and B.E Mechanical Engineering

Name of Expert/Industry Mentor: Mr. MB Mate and Mr. K Pradeep

Designation of the Expert: Mr. MB Mate: Deputy General Manager (Maintenance), Mumbai Refinery and Mr. K Pradeep: Manager Maintenance (Rotary), Central Engineering Workshop, Mumbai Refinery

Name of PS Faculty: Dr. Srikanta Dinda

Key Words: Machine Learning, Wet Gas Compressor, Python, Libraries, Data Pre-Processing, Regression

Project Areas: Computer Science (Data Analytics and Machine Learning) and Mechanical Engineering

Abstract

Rotary equipments are critical, non-substitutable links in production chains of process industries and thus by extension the largest consumers of power. The state of functioning of and reliability of these equipments determines the ultimate production costs and capacity of these industries. Of these Rotary equipments, one of the most crucial in the refinery is the Wet Gas Compressor. This project aims to provide an insight into the predictive maintenance of the Wet Gas Compressor. The 4 main technical aspects of this project are: Data Collection, Data Cleaning, Forming a Predictive Machine Learning Model for Failure Prediction and Writing and Running the Algorithm which form the central pillar of this project. Through this project, I would be able to perform the key operations concerning rotating equipment that would enable me to understand the effect of process and environmental changes on the equipment operation, maintenance and reliability.

Raghav Khanna
23rd June, 2020

Mr. MB Mate
23rd June, 2020

Table of Contents

1. Cover Page 1
2. Title Page.....2
3. Acknowledgements.....3
4. Personal and Project Details and Abstract.....4
5. Table of Contents.....6
6. Preface.....7
7. Objectives and Goals of the Project.....8
8. Methodology Followed.....8
9. Overview of the Company.....9
10. About FCCU.....11
11. About Wet Gas Compressor.....13
12. Introduction.....17
13. Main Text- Computer Science.....18
    a. Data Collection.....18
    b. Libraries Used.....19
    c. Data Pre-Processing.....20
    d. Machine Learning.....22
    e. Framing Which Model To Use.....24
    f. Regression.....25
    g. Decision Tree Regression.....26
    h. Line Wise Explanation of the Code.....29
14. Conclusion.....32
15. List of Figures, Tables and Code Snippets.....33
16. Appendix.....34
17. References 35
18. Glossary.....36

Preface

This report has been prepared as a part of my PS-1(summer internship) project, as a second year student at BITS Pilani from 17th May to 24th June 2020. This report is made under the helpful guidance of, **Mr. Srikanta Dinda (BITS)**, **Mr. MB Mate** and **Mr. K. Pradeep(Mumbai Refinery).**

This report is about predicting failures in Rotary Equipment in the refinery with focus on the Wet Gas Compressor using Machine Learning. It involves the pre-processing of the data received, application of a suitable regression model and then predicting the Target Variable and hence the failures. Each component involves the theory regarding the topic followed by the application of that concept in my project in BPCL. Figures have been attached to supplement easier understanding.

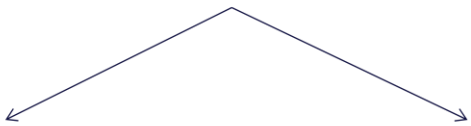
Completing this report and this internship has given me a lot of insight into the workings of a refinery and some of its important processes along with a lot of in-depth understanding of important Computer Science concepts. Through this report, I wish to summarise all my findings and learning of the past one and a half month.

Objectives and Goals of the Project

The main objective of the project is to be able to predict failures in the Wet Gas Compressor based on Historical Data using Machine Learning before they actually happen in real life. Other secondary objectives that will aid the main objective along the way include:

1. To be able to collect, clean and process data so that it can be compiled and executed in a program
2. Understand the fundamentals of Rotary equipment, including but not limited to Pumps and Compressors, and understand in detail the working of a Wet Gas Compressor
3. Learn how to implement Machine Learning and use to for Predictive Maintenance of the Wet Gas Compressor, i.e predicting failures of the Wet Gas Compressor based on a suitable model like Linear Regression
4. The exercise of working in a Chemical Plant would be futile if I do not at least understand the basic structure of a refinery and all its components

Methodology to be adopted



Mechanical Engineering part

- Get a general idea about Pumps and Compressors
- Learn about the basic working of a Wet Gas Compressor and the factors affecting its functioning
- List down all the factors that affect the performance of a Wet Gas Compressor since they will all be variables in the Computer Science part

Computer Science Engineering part

- Learn about various Data Science libraries in Python like the numpy and pandas library
- Learn the concept of Regression to apply Machine Learning
- Start to clean the data after receiving data from the company
- Write the Machine Learning algorithm after completing the Machine Learning Course
- Predict failures using my ML Algorithm and the data provided by the company

Overview of the Company

About Bharat Petroleum Corporation Limited (BPCL):

Bharat Petroleum Corporation Limited (BPCL) is an Indian state-controlled oil and gas company headquartered in Mumbai, Maharashtra. The Corporation operates two large refineries of the country located at Mumbai and Kochi. The company is ranked 275th on the Fortune Global 500 list of the world's biggest corporations as of 2019.

Bharat Petroleum operates the following refineries:

Mumbai Refinery: Located in Mumbai, Maharashtra. It has a capacity of 13 million metric tonnes per year.

Kochi Refinery: Located near Kochi, Kerala. It has a capacity of 15.5 million metric tonnes per year.

Bina Refinery: Located near Bina, Sagar district, Madhya Pradesh. It has a capacity of 6 million metric tonnes per year. This refinery is operated by Bharat Oman Refineries Limited, a joint venture between Bharat Petroleum and Oman Oil Company.

Numaligarh Refinery: Located near Numaligarh, Golaghat district, Assam. It has a capacity of 3 million metric tonnes per year.

The company business is divided in seven SBUs (Strategic Business Units), like Retail, Lubricants, Aviation, Refinery, Gas, I&C and LPG.

About the BPCL Mumbai Refinery

The Bharat Petroleum Mumbai Refinery (BPCL - MR) is one of the most versatile refineries in India and excels in all aspects like quality, technology, energy, human relations, safety, environmental friendliness and operating cost. .

With successful de-bottlenecking and implementation of various major projects, Mumbai

Refinery has a capacity to process 12 MMT of crude oil per annum. Mumbai Refinery has processed 93 different types of crude in five decades of its operations, making it one of the most flexible refineries in the country. Mumbai Refinery uses the latest microprocessor based Digital Distributed Control System (DDCS) and has been accredited with ISO 9001 (Quality Management System), ISO 14001 (Environment Management System) and OHSAS (Occupational Health and Safety Management System). Quality Assurance Laboratory has been accredited with a certification from, National Accreditation Board for Testing and Calibration Laboratories (NABL), an autonomous body under the aegis of Department of Science & Technology, Government of India, and is registered under the Societies Act 1860. Mumbai Refinery was one of the first refineries to have got accredited with ISO 50001 (Energy Management System) in the year 2014.

Mumbai Refinery has implemented a state of the art on-line monitoring tool, covering entire functions of the refinery, for disseminating information and decision making. Mumbai refinery stands tall among the peers for adhering to all quality and safety standards and also consistently meeting MOU targets set by MOP&NG.

Refinery Configuration

Mumbai refinery was commissioned in 1955 with a crude oil processing capacity of 2.2 MMTPA. The refining capacity has been augmented to the present level of 12 MMT through progressive revamps, the addition of various process units and the incorporation of advanced refining technologies.

Mumbai refinery processes various types of crudes which include Bombay High, East African, Petronas, Kuwait, Arab mix, Arab medium, Basrah, Arab extra light, Murban, Umm Shaif, Western Texas Intermediate, Bantulu, Kiduong etc. The main products are LPG, Naphtha, MS, Benzene, Toluene, Hexane, SBP, MTO, Kerosene, Jet Fuel, Diesel, Light Diesel Oil, Lubes, Fuel Oil, LSHS and Bitumen.

Commissioned Diesel Hydrotreating Unit (DHT), Naphtha Isomerization (ISOM) & Gasoline Treatment Unit (GTU) to meet the government mandate of producing 100% BS-VI fuels.

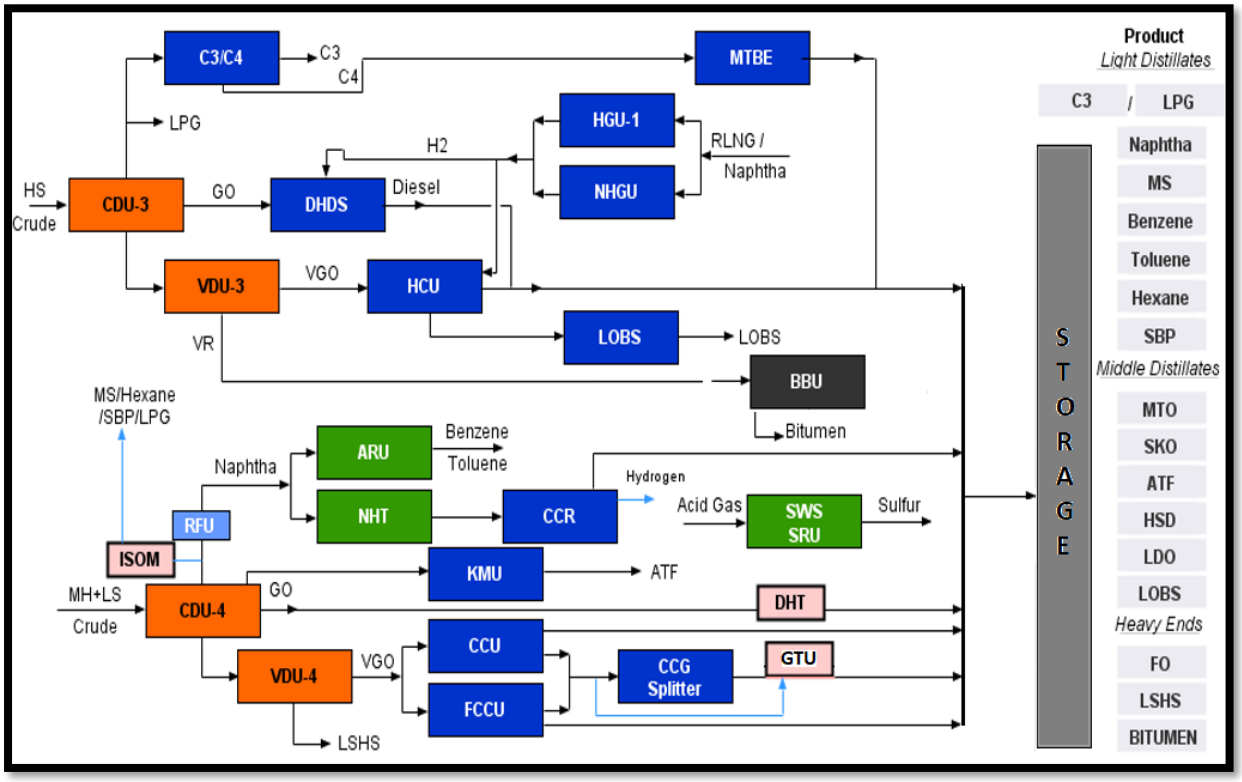


Figure 1: The process flow diagram of Mumbai Refinery

About FCCU

The distillation unit yields straight run products contained in the crude oil. However, some of these are not suitable in quantity and quality to meet the present requirements. For example, the quality of gasoline found naturally in crude oil does not satisfy car engine requirements. Also one needs higher yields of middle distillates, which the distillation unit alone cannot provide. These requirements of more middle distillates, better gasoline and more LPG, have resulted in the evolution of the Fluid Catalytic Cracking Units.

In these units, feedstock is charged to a reactor in which it is contacted with hot catalyst, made of Silica-Alumina that vaporizes this feedstock and at the same time brings about its chemical decomposition by cracking. The cracked vapours pass over to the Fractionator where they are separated into gas, gasoline, cycle oils and clarified oil.

During the cracking reaction, some carbon gets deposited on the surface of the catalyst, which is continuously removed by “burning” in the Regenerator. A Stripper that entrains and

separates hydrocarbons by stripping with steam reduces the load on the Regenerator. Hot regenerated catalyst is then returned to the Reactor to renew the cycle.

The catalyst, in the form of a fine powder moves between the three main vessels as a fluid. Cracking produces higher quality gasoline and other valuable products. Gas is burnt in the refinery furnaces. LPG is sold to domestic and industrial customers. Cycle oils are blended to diesel and Clarified oil is blended with Short Residue from the Feed Preparation Unit to produce furnace oil / LSHS.

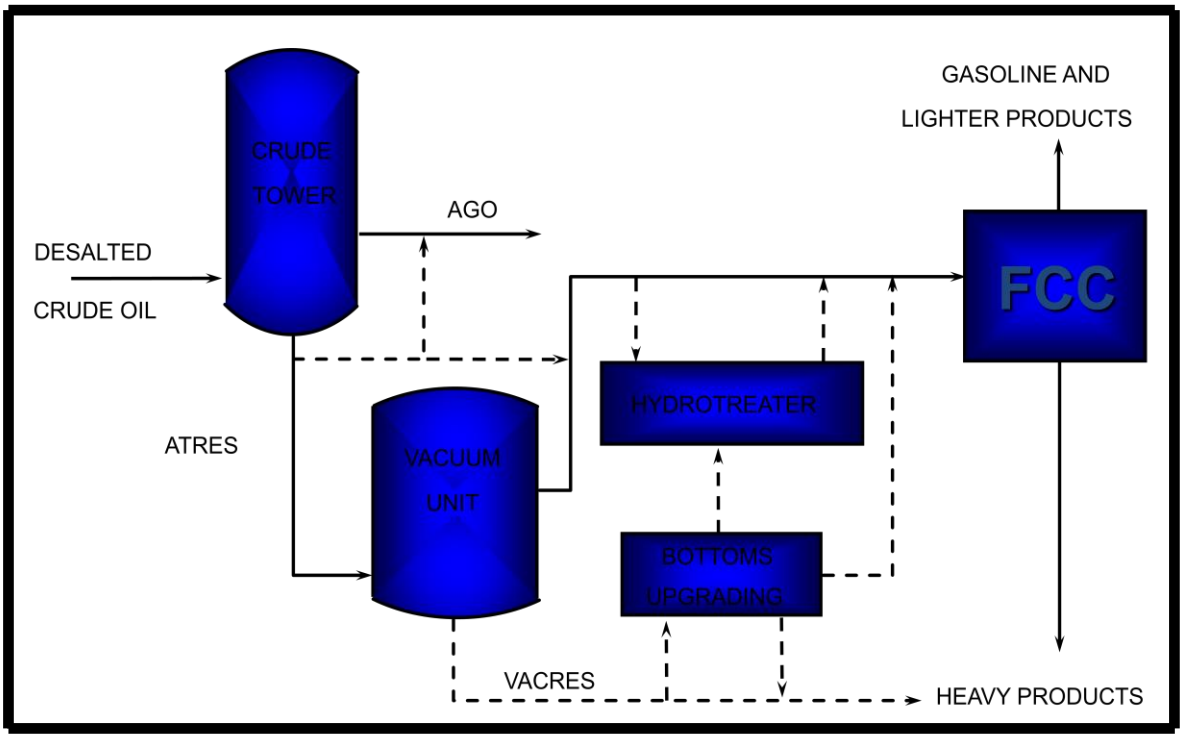


Figure 2: Flow Diagram of the FCCU

About Wet Gas Compressor

Function

The FCCU wet gas compressor's major function is reactor pressure control. The machine must compress gas from the main column overhead receiver to gas plant operating pressure while maintaining stable regenerator-reactor differential pressure. The wet gas compressor and its control system play a vital role in maintaining steady reactor operating pressure.

Working Principle

It uses rotating impellers to increase the pressure of a fluid. The fluid enters the compressor near the rotating axis, streaming into the rotating impeller. The impeller consists of a rotating disc with several vanes attached. The vanes normally slope backwards, away from the direction of rotation. When the fluid enters the impeller at a certain velocity due to the suction system, it is captured by the rotating impeller vanes. The fluid is accelerated by pulse transmission while following the curvature of the impeller vanes from the impeller center (eye) outwards. It reaches its maximum velocity at the impeller's outer diameter and leaves the impeller into a diffuser or volute chamber.

So the centrifugal force assists accelerating the fluid particles because the radius at which the particles enter is smaller than the radius at which the individual particles leave the impeller. Now the fluid's energy is converted into static pressure, assisted by the shape of the diffuser. The process of energy conversation in fluids mechanics follows the Bernoulli principle which states that the sum of all forms of energy along a streamline is the same on two points of the path. The total head energy in a pump system is the sum of potential head energy, static pressure head energy and velocity head energy.

As the velocity of the fluid increases, it is essentially a velocity machine. After the fluid has left the impeller, it flows at a higher velocity from a small area into a region of increasing area. So, the velocity is decreasing and so the pressure increases as described by Bernoulli's principle. This results in an increased pressure at the discharge side. As fluid is displaced at the discharge side, more fluid is sucked in to replace it at the suction side, causing flow.

Compressor Design: Wet Gas Compressor is a single shaft centrifugal compressor with Horizontal split casing. The mechanical energy is supplied by the drive is transmitted to the

process gas by the impellers of the centrifugal compressor. It uses six to eight impellers to compress gas from the main column overhead receiver to the gas plant operating pressure. Most have inter-stage condensing systems after the first three or four stages (low-stage) that cool the compressed gas, condense a small portion and separate the gas and liquid phases. Inter-stage receiver gas is then compressed in the last three or four stages (high-stage).

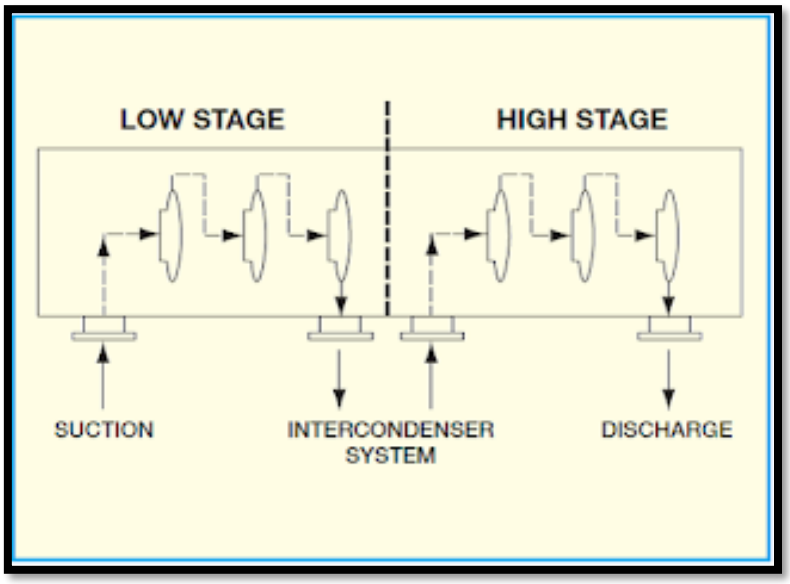


Figure 3: Schematic Diagram of a Wet Gas Compressor

Purpose of a Wet Gas Compressor in the BPCL Refinery

The combined function of the WGC System is to remove the bulk of the light ends (C1-C2) from the feed streams, while recovering the C3- and heavier material for separation into saleable products

Why use a Wet Gas Compressor over a normal compressor?

When compressing a wet gas, the required power increases due to the increased percentage of liquid than a normal gas and the maximum suction flow may be significantly reduced. To compress a Wet Gas using a normal compressor, we would have to first separate out the liquid from the gas before compressing the gas, so a wet gas compressor reduces the need for the extra tedious process.

Components of a Wet Gas System

- Oil system: lube-oil system, Control oil system, cooling oil system
- Live steam system
- Drainage system.
- Seal gas system.
- Process gas system.
- Machine Monitoring system
- Control system.
- Wash-water system.
- Leak off- steam system

Contains the sensor that monitors the level of lube oil in the Wet Gas Compressor that will be the factor affecting failure in the project

Major parts of a Wet Gas Compressor

1. **Casing** **2. Impellers** **3. Shaft**

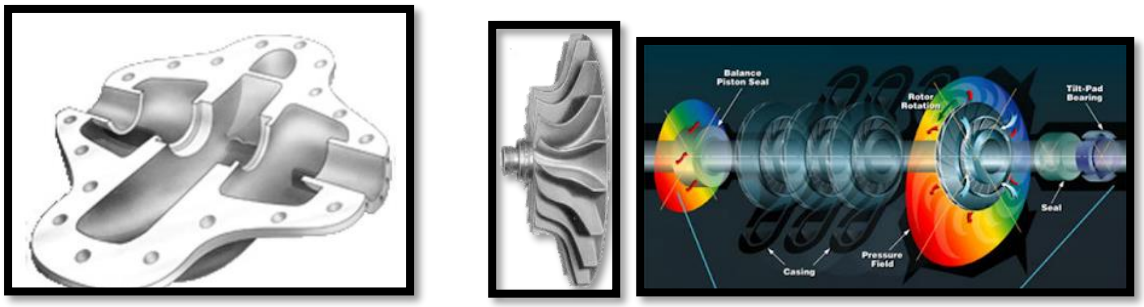


Figure 4: Major parts of a Wet Gas Compressor

Casing: The WGC's casing houses the hole assembly and protects is from harm as well as forces the fluid to discharge from the compressor and convert velocity into pressure. The casings design does not influence the total dynamic head but is important to reduce friction losses. It supports the shaft bearings and takes the centrifugal forces of the rotating impeller and axial loads caused by pressure thrust imbalance.

Impeller: An Impeller is a rotating component of the Wet Gas Compressor, which transfers energy from the motor that drives the pump to the fluid being pumped by accelerating the

fluid outwards from the centre of rotation. The velocity achieved by the impeller transfers into pressure when the outward movement of the fluid is confined by the casing.

Shaft: The shaft is the connection between impeller and drive unit which is in most cases an electric motor but can also be a gas turbine. It is mainly charged by a radial force caused by unbalanced pressure forces in the spiral casing and an axial force due to the pressure difference between front and backside of the impeller.

Specification of a Wet Gas Compressor in BPCL Mumbai Refinery:

Equipment No. = 106464
Gas Handled = Hydrocarbons
Manufacturer = BHEL
Model Number = 2MCL458
Feed = 2500 MT per day
Target/ Dependent/ to be predicted Variable for this project

The target variable, the variable which decides success or failure in my project is the **rate of loss of Lube Oil in the Wet Gas Compressor.** The data for the lube oil level is indicated by the column FCCGLI510.PV from the dataset.

The Wet Gas Compressor fails when the rate of drop of lube oil increases above a certain point. Therefore to create the Target Variable (which will be predicted/calculated in the end), we need to **find the slope at all points in the column where lube oil is reducing.**

RATE OF LUBE OIL LOSS = (x2-x1/5)*60*24
(at each point in litres per day)

Where:
x1 = Initial Lube Oil Level
x2 = Lube Oil Level after 5 minutes(x2<x1: since the rate of lube oil is dropping)

Introduction

Predictive Maintenance

Think about all the machines you use during a year, all of them, from the geyser or oven every morning to a flight every summer vacation. Now imagine that, from now on, one of them would fail every day. What impact would that have? The truth is that we are surrounded by machines that make our life easier, but we also get more and more dependent on them. Therefore, the quality of a machine is not only based on how useful and efficient it is, but also on how reliable it is. As is the case in our normal life, that is also the case for every company. Now **imagine if the machine in question is a huge Rotary machine which is the centrepiece of one of the critical process unit of a large refinery of the country**. The impact of its failure will be huge, resulting in not only production losses but also through put reduction of the refinery. This is where **predictive maintenance** i.e predicting the failures of the device before it actually stops functioning comes in.

When the impact of a failure cannot be afforded, such as a malfunctioning compressor or pump in a refinery, the machine is subjected to preventive maintenance which involves periodic inspection and repair. The challenge of proper scheduling grows with the complexity of machines: in a system with many components working together: how can we find the right moment when maintenance should be performed so that components are **not prematurely replaced (causing waste of money) but the whole system still stays functioning reliably (avoid machine from stopping)**? Providing an answer to this question is the aim of predictive maintenance and this project, where I seek to build a model that will quantify the risk of failure for the machine (WGC) in any moment in time and use this information to improve scheduling of maintenance

Rotary Equipments in BPCL

Primary purpose of rotary equipments is to enhance pressure energy by pumping fluids. Any failure or malfunctioning causes partial or complete halt of production which is highly undesirable. Modes of failure of rotary equipments are numerous; few of them are component specific while others are common. So it is daunting task to determine the exact cause of failures. BPCL’s Mumbai refinery, one of the most complex refineries in the country, is a Lube based refinery with one of the highest lube production capacity in India. It employs more than 5000 rotary equipments.

Main Text – ***Computer Science Part***

Each component has been divided into 2 parts- the general theory and the application in my project.

1. **Data Collection from the Refinery and Sorting the Data so that it can be fed into the program**

To build a failure model, we require historical data that allows us to capture information about events leading to failure. In addition to that, "static" features of the system provide valuable information, such as mechanical properties, average usage and operating conditions. Before deciding on a methodology and process, the following questions were kept in mind:

- **What are the types of failure that can occur? Which ones will we try to predict?**

  In BPCL's case, failure in the Wet Gas Compressor is caused when the lube oil level drops below a certain point. We try to predict when the rate of drop of lube oil level is increasing which then directly implies that the level of lube oil will drop more than the permissible amount. That is the only type of failure we will be trying to predict.

- **How does the "failure process" look like? Is it a slow degradation process or an acute one?**

  The failure process is not a slow degradation but an acute one as due to certain changes in its environment,

- **Which parts of the machine/system could be related to each type of failure? What can be measured about each of them that reflect their state? How often and with which accuracy do these measurements need to be performed?**

  Before the project started, I had aimed to get a Machine Learning model with over 90% accuracy which will be able to predict which factors affect failure how much. However, by the time of report submission, I was able to make my model upto 85% accurate. My efforts to achieve higher level of accuracy will be on post submission of report.

Since the life span of machines is usually in the order of years, it means that data has to be collected for an extended period of time in order to observe the system throughout its degradation process.

BPCL: In BPCL, all the data from the Rotary Equipment (major machines like Wet Gas Compressor in particular) is captured using various different sensors which have been placed inside the device, that calculate all the physical quantities and report to the mainframe every 5 minutes. All this data was collected and sorted and converted to a .csv file which can be read by a Python program.

DATE	FCCFFI232.PV	FCCFPC175.MV	FCCFPC175.PV	FCCGAI008.PV	FCCGDI501AM.PV	FCCGDI501M.PV	FCCGDI502AM.PV
09-01-2019 00:00	2337.908	82.5	1.55	1.2	0.0983451	0.134822	-0.000606519
09-01-2019 00:05	2344.052	82.4	1.55	1.2	0.0982615	0.134689	-0.000606519
09-01-2019 00:10	2347.076	82.7	1.55	1.2	0.098513	0.134888	-0.000606519
09-01-2019 00:15	2357.487	82.9	1.55	1.2	0.0988487	0.135151	-0.000606519
09-01-2019 00:20	2358.864	83.1	1.55	1.2	0.0987253	0.134935	-0.000606519
09-01-2019 00:25	2359.356	83.1	1.55	1.2	0.0989772	0.135187	-0.000606519
09-01-2019 00:30	2358.189	83.3	1.55	1.2	0.0988986	0.135115	-0.000606519
09-01-2019 00:35	2357.2	83.2	1.55	1.2	0.0991316	0.135257	-0.000606519
09-01-2019 00:40	2360.125	82.9	1.55	1.2	0.0993888	0.135455	-0.000606519
09-01-2019 00:45	2359.093	82.7	1.55	1.2	0.0997212	0.135699	-0.000606519
09-01-2019 00:50	2356.312	82.9	1.55	1.2	0.0994434	0.135388	-0.000606519
09-01-2019 00:55	2355.918	82.9	1.55	1.2	0.0990164	0.134995	-0.000606519
09-01-2019 01:00	2355.518	83.1	1.55	1.2	0.0992626	0.135153	-0.000606519
09-01-2019 01:05	2348.635	82.8	1.55	1.2	0.099293	0.135103	-0.000606519
09-01-2019 01:10	2340.751	82.7	1.55	1.2	0.0991848	0.134983	-0.000606519
09-01-2019 01:15	2345.468	82.5	1.55	1.2	0.0992287	0.134993	-0.000606519

Table 1: Snippet of data after being sorted

Libraries in Python

Before we can clean or pre-process the data, we need to include certain libraries in Python which will provide various in-built functions that will make our work much easier. There were 4 major libraries that I used in my project which were:

Numpy: The Library in Python used to perform various mathematical operations and handle large scale data using arrays.

Pandas: The Library in Python used for data manipulation and analysis

Sklearn: The library in Python which includes Machine Learning Regression and Classification algorithms

Matplotlib: The library in Python used to plot graphs and charts from the data provided

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split # used for splitting training and testing data
from sklearn import preprocessing
import matplotlib.pyplot as plt
from sklearn import linear_model
from sklearn.preprocessing import PolynomialFeatures
from sklearn import metrics #Import scikit-learn metrics module for accuracy calculation
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_error
from sklearn.tree import DecisionTreeClassifier # Import Decision Tree Classifier
```

Code Snippet 1: All the libraries imported and used in the project

2. Data Cleaning and Pre-Processing

Since sensors are machines too, there is a chance that they could've malfunctioned or switched off during certain intervals of time or have erroneous value, which would almost definitely lead to faulty readings or no readings at all. All this prevents our Machine Learning model from correctly predicting data due to these erroneous gaps in data. This is why our data needs to be cleaned.

a. **Missing values**: It is very much usual to have missing values in the dataset. It may have happened during data collection, but regardless missing values must be taken into consideration.

   Most common ways to deal with the missing values:

   Eliminate rows with missing data - Simple and sometimes effective strategy but it fails if many objects have missing values. If a feature has mostly missing values, then that feature itself can also be eliminated.

   Estimate missing values - If only a reasonable percentage of values are missing, then we can also run simple interpolation methods to fill in those values. However, most common method of dealing with missing values is by filling them in with the mean, median or mode value of the respective feature.

b. **Columns with no variance**(the physical quantity does not change throughout the time period) – these columns are generally removed from the data because they have no significance in predicting failures as the quantity is not changing irrespective of the fact that the machine has failed or not.

BPCL

A total of 10,000 data points were collected from the sensors in the refinery over the past year and 76 variables in total were monitored and measured. Out of these 10,000 data points, 12 data points were found to be empty, with no data, while there were 4 columns which were either empty or contained constant data throughout and hence were not useful for the project.

9625	10-04-2019 09:55	2604.303	81.9	1.55	1.3	0.132537	0.144687	-0.000606519
9626	10-04-2019 10:00	2611.068	82.1	1.55	1.3	0.132558	0.144639	-0.000606519
9627	10-04-2019 10:05	2611.284	83	1.55	1.3	0.131584	0.143569	-0.000606519
9628	10-04-2019 10:10	2614.502	21	0.39	0.3	0.033036	0.0360819	-0.000153652
9629	10-04-2019 10:15	2613.985	0	0	0	0	0	0
9630	10-04-2019 10:20	2613.442	0	0	0	0	0	0
9631	10-04-2019 10:25	2612.899	0	0	0	0	0	0
9632	10-04-2019 10:30	2612.356	0	0	Missing data		0	0
9633	10-04-2019 10:35	2611.814	0	0			0	0
9634	10-04-2019 10:40	2611.271	0	0	0	0	0	0
9635	10-04-2019 10:45	2610.771	0	0	0	0	0	0
9636	10-04-2019 10:50	463.135						
9637	10-04-2019 10:55	1217.029	81.6	1.55	1.2	0.129386	0.140497	-0.000606519
9638	10-04-2019 11:00	2604.337	81.6	1.55	1.2	0.129454	0.140641	-0.000606519

Table 2: Snippet of Data before being Pre-Processed

9625	10-04-2019 09:55	2604.303	81.9	1.55	1.3	0.132537	0.144687
9626	10-04-2019 10:00	2611.068	82.1	1.55	1.3	0.132558	0.144639
9627	10-04-2019 10:05	2611.284	83	1.55	1.3	0.131584	0.143569
9628	10-04-2019 10:10	2614.502	21	0.39	0.3	0.033036	0.0360819
9629	10-04-2019 10:55	1217.029	81.6	1.55	1.2	0.129386	0.140497
9630	10-04-2019 11:00	2604.337	81.6	1.55	1.2	0.129454	0.140641
9631	10-04-2019 11:05	2604.652	81.9	1.55	1.2	0.130442	0.141656
9632	10-04-2019 11:10	2600.769	82.1	1.55	1.2	0.131409	0.142544
9633	10-04-2019 11:15	2605.188	82.1	1.55	1.2	0.131826	0.142898

Table 3: Snippet of Data after being Pre-Processed

The missing data has been removed using the Python code which has also been attached ahead.

```
import numpy as np
import pandas as pd
from sklearn.impute import SimpleImputer # used for handling missing data
from sklearn.preprocessing import LabelEncoder, OneHotEncoder # used for encoding categorical data
from sklearn.preprocessing import StandardScaler  #for feature scaling
from sklearn.model_selection import train_test_split # used for splitting training and testing data
from sklearn.compose import ColumnTransformer
from sklearn import preprocessing

dataset = pd.read_csv('WGC_10000.csv') # to import the dataset into a variable

#dataset.replace('', np.nan, inplace=True)#set all empty cells as np.nan

dataset= dataset.dropna(axis=1, thresh= 1)
dataset = dataset.dropna(axis=0, thresh= 4)
#dataset_droppedrows = dataset

#only_na = dataset_droppedrows[~dataset_droppedrows.index.isin(dataset_droppedrows.index)]

#dataset_ip = dataset.fillna(dataset.interpolate())
#dataset_ip.to_csv("datasett_ip.csv") #change to mean/median/interpolate based on what you want
dataset = dataset.fillna(dataset.interpolate())

#dataset_ip = dataset_ip.drop(dataset_ip.std()[dataset_ip.std() == 0].index.values, axis=1)#drop columns with 0 variance
#dataset_ip.to_csv("datasett_ip.csv")

dataset = dataset.loc[:, (dataset!= 0).any(axis=0)]#drop the two 0 valued columns

dataset = dataset[dataset.astype('bool').mean(axis=1)>=0.25]#deletes all rows with 75% zeroes
dataset_ip = dataset

dataset.iloc[:,1:-1] = dataset.iloc[:,1:-1].apply(lambda x: ((x-x.min())/(x.max()-x.min())), axis=0) #normalise data between 0 and 1
Corr = dataset.corr()
```

Code Snippet 2: Python Code for Pre-Processing of Data

Explanation of each important line of the code:

- **Line 1-8 :** Import various Python libraries which will be required to build, compile and run the entire project
- **Line 9:** Importing and reading the .csv file WGC_10000 which has all the data sorted and compiled from the various sensors in the Wet Gas Compressor.
- **Line 14-15:** Remove the rows and columns which are empty or have a fixed number of empty cells in them, since they are part of erroneous data.
- **Line 27:** Drop the columns which have 0 variance in them, i.e those which are constant throughout.
- **Line 29:** Drop all the rows whose 75% of all values are 0(similar to the photos shown above)
- **Line 32:** Normalise all the data to between -1 and 1, so we can get a normal distribution of the data which will be required while writing the Machine Learning Algorithm.

Machine Learning

The most popular and technical definition of Machine Learning that I learnt in the course Machine Learning By Andrew Ng is : “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”

Machine Learning is generally categorized into three types: Supervised Learning, Unsupervised Learning and Reinforcement learning

Supervised Learning: In supervised learning the machine experiences the examples along with the labels or targets for each example. The labels in the data help the algorithm to correlate the features. Two of the most common supervised machine learning tasks are **classification** and **regression**.

In **classification** problems the machine predicts discrete values. It is similar to predicting True or False. That is, the machine must predict the most probable category, class, or label for new examples. Applications of classification include predicting whether a stock's price will rise or fall, or deciding if a news article belongs to the politics or leisure section.

In **regression** problems the machine must predict the value of a continuous response variable. Examples of regression problems include predicting the sales for a new product, or the salary for a job based on its description. **I have used a regression model for my project since the data provided to me was continuous in nature –** process elaborated further in next section

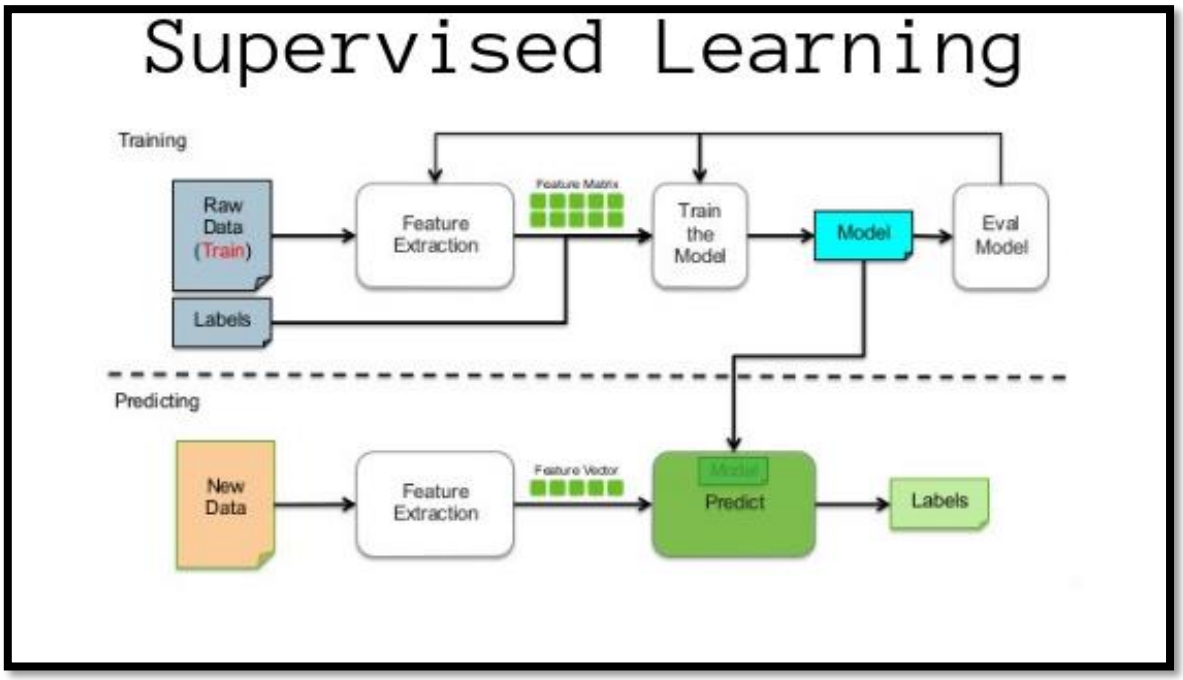


Figure 5: Schematic Diagram of Supervised Learning (used in the project)

Source: https://www.quora.com/What-is-supervised-learning/

```
X = dataset.loc[:, dataset.columns != 'Target']

y = dataset['Target']
```

Code snippet 3: Taking X as the independent variable (which will be used to predict) and y as target Variable (which will be predicted using X)

Unsupervised Learning: When we have unclassified and unlabeled data, the system attempts to uncover patterns from the data. There is no label or target given for the examples.

Reinforcement Learning: Reinforcement learning refers to goal-oriented algorithms, which learn how to attain a complex objective (goal) or maximize along a particular dimension over many steps. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance.

3. **Framing which model to use while predicting failure**

When thinking about how to frame a predictive maintenance model my mentor and I decided our model based on the following factors:

- What kind of output should the model give?
- Is enough historical data available or just static data?
- Is every recorded event labelled, i.e. which measurements correspond to good functioning and which ones correspond to failure? Or at least, is it known when each machine failed (if at all)?
- How long in advance should the model be able to indicate that a failure will occur?
- What are the performance targets that the model should be optimized for? High precision, high sensitivity/recall, high accuracy? What is the consequence of not predicting a failure or predicting a failure that will not happen?

BPCL

After taking into consideration 4 different models, we chose on the **Regression model** to predict failure in the equipment.

Strategy: Regression model

Data Characteristics: Static and historical data of the Wet Gas Compressor are available

Basic Assumptions:

- Based on static characteristics of the system and on how it behaves now, the remaining useful time can be predicted which implies that both static and historical data are required and that the degradation process is smooth.
- Just one type of "path to failure" is being modelled - **based on the level of Lube Oil in the Wet Gas Compressor; failure occurs if the level of Lube Oil falls below a certain level.**
- Labelled data is available and measurements were taken at different moments during the system's lifetime **(Time span of 1 year, data points were taken every 5 minutes)**

Regression

Regression can be defined as a method or an algorithm in Machine Learning that models a target value based on independent predictors. It is essentially a statistical tool used in finding out the relationship between a dependent variable an independent variable. This method comes to play in forecasting and finding out the cause and effect of the relationship between variables.

Regression techniques differ based on

1. The number of independent variables
2. The type of relationship between the independent and dependent variables.

Regression is basically performed when the dependent variable is of a continuous data type. The independent variables, however, could be of any data type — continuous, nominal/categorical etc.

Regression methods find the most accurate line describing the relationship between the dependent variable and predictors with least error. In regression, the dependent variable is the function of the independent variable and the coefficient and the error term.

BPCL

Since there are various different Regression Techniques, different techniques will give different results for the same dataset. Hence you need to find which Regression Model is most suitable for your model. I used 5 types of regression techniques and the R^2 values (accuracy) for each model along with a brief description of the model are listed below:

Linear regression: Linear regression attempts to model the relationship between two variables by fitting a linear equation to the observed data. **R^2 value = 0.70**

Lasso Regression: Lasso regression is a type of regression that uses shrinkage i.e where data values are shrunk towards a central point, like the mean. **R^2 value = 0.71**

Ridge Regression: Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. **R^2 value = 0.72**

Polynomial regression: Polynomial Regression is a form of regression in which the relationship between the independent variable x and dependent variable y is modelled as an nth degree polynomial. **R^2 value = 0.77**

Decision tree regression: Most Accurate Regression Model for my dataset – discussed in detail. **R^2 value = 0.85**

Decision Tree Regression

Decision trees are one of the most widely-used machine learning models, due to the fact that they work well with noisy or missing data and can easily be ensembled to form more robust predictors. Moreover, you can directly visual your model's learned logic, which means that it's an incredibly popular model for domains where model interpretability is important.

Various important terminologies regarding a Decision Tree Regression Algorithm have been given below:

1. **Root Node:** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
4. **Leaf/Terminal Node:** Nodes do not split is called Leaf or Terminal node.
5. **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning (Opposite of splitting)
6. **Branch / Sub-Tree:** A sub section of entire tree is called branch or sub-tree.
7. **Parent and Child Node:** A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.

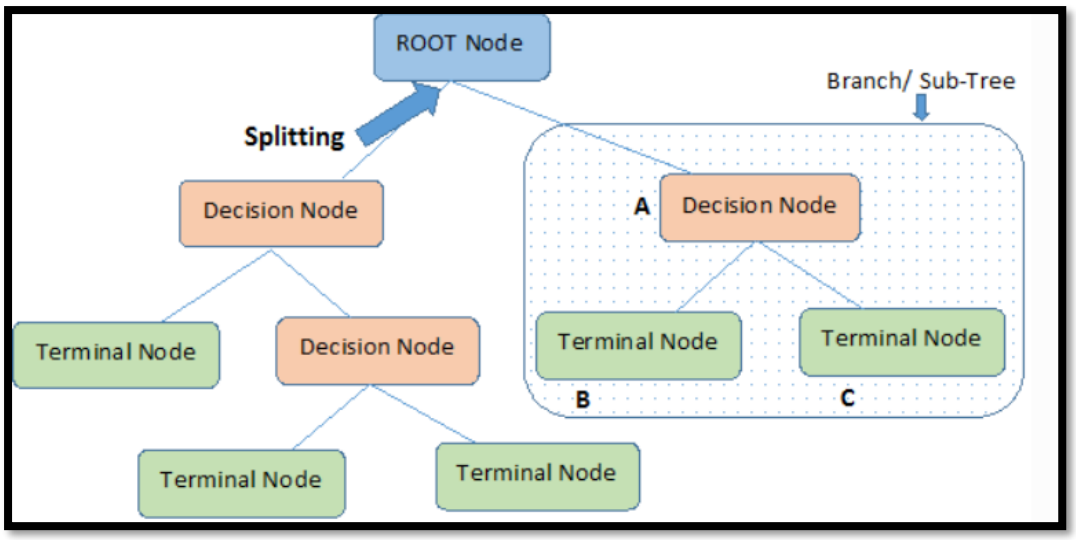


Figure 6: Schematic Diagram of a Decision Tree Regression Algorithm

Source: www.gdcoder.com

How does it work?

A decision tree is an algorithm which arrives at an estimate by asking a series of questions to the data, each question narrowing our possible values until the model get confident enough to make a single prediction.

Decision Tree Algorithm Pseudocode

1. Place the best attribute of the dataset at the **root** of the tree.
2. Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find **leaf nodes** in all the branches of the tree.

It is better explained through an example:

In the problem below, x1 and x2 are two features which allow us to make predictions for the target variable y by asking True/False questions.

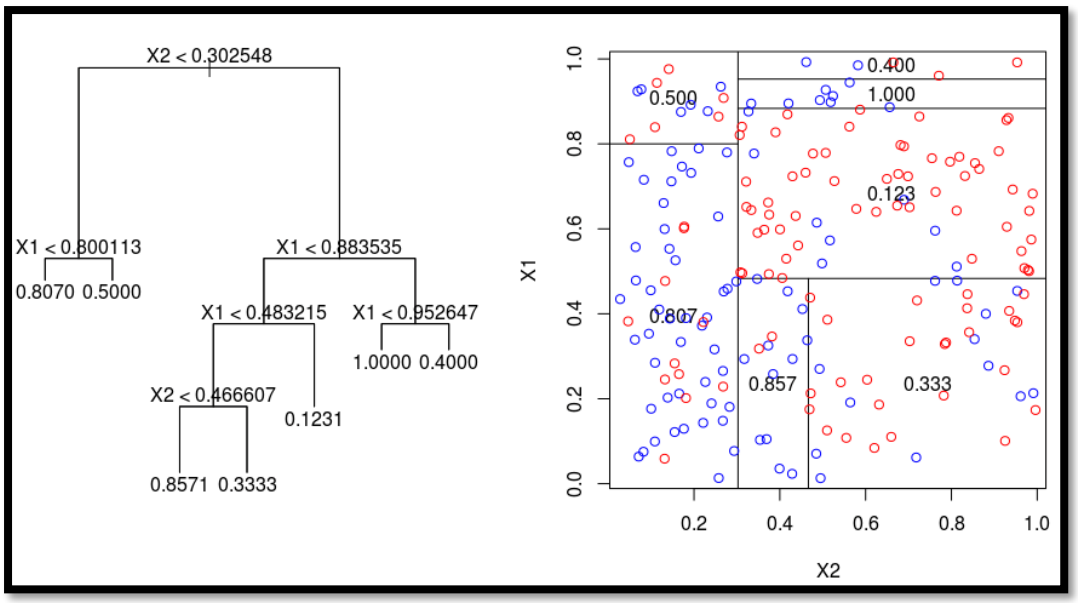


Figure 7: Decision Tree Regression Example (RHS shows the data while LHS shows the Decision tree for the same data)

For each True and False answer there are separate branches. No matter the answers to the questions, we eventually reach a prediction (leaf node). Start at the root node at the top and progress through the tree answering the questions along the way.

How does a Decision Tree Learn? And where do the values of X1 and X2 come from?

As a supervised machine learning model, a decision tree learns to map data to outputs in the training phase of model building. During training, the model is fitted with any historical data that is relevant to the problem domain and the true value we want the model to learn to predict. The model learns any relationships between the data and the target variable.

After the training phase, the decision tree produces a tree similar to the one shown above, calculating the best questions as well as their order to ask in order to make the most accurate estimates possible. When we want to make a prediction the same data format should be provided to the model in order to make a prediction. **The prediction will be an estimate based on the train data that it has been trained on.**

Code and Line-wise Explanation for the Decision Tree Regression Algorithm

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

Code Snippet 4: Splitting Data into Training Set and Test Set

Before we can apply the regression algorithm, we need to split the entire dataset into training set and test set. The Training Set is used to train the model i.e we give the Machine Learning model input and expected output for certain datapoints and allow it to "learn" and form an algorithm based on it.

```
from sklearn.tree import DecisionTreeRegressor
# create a regressor object

regressor = DecisionTreeRegressor(min_samples_split= 2, max_depth = 10, random_state = 80)
# fit the regressor with X and Y data
regressor.fit(X_train, y_train)
print('Variance Score Test:', regressor.score(X_test,y_test))

y_pred = regressor.predict(X_train)
```

```
Variance Score Test: 0.8507041364618664
```

Code Snippet 5: Creating a Decision tree Regression algorithm

We then create a variable regressor which uses the inbuilt DecisionTreeRegresor function from the sklearn library and 'fit' it on the Training Set. After the algorithm is formed the same algorithm is used on the Test set and the accuracy with which it is able to predict the Target variable/Dependent variable is noted (regressor.score()). The output of Variance Score indicates the accuracy (85%) of the model.

```
from sklearn.externals import joblib
joblib.dump(regressor,'Project_Model.pkl')

mse = mean_absolute_error(y_train,regressor.predict(X_train))
print("Training Set Mean Absolute Error : %.4f" %mse)

mse = mean_absolute_error(y_test,regressor.predict(X_test))
print("Test Set Mean Absolute Error : %.4f" %mse)
```

```
Training Set Mean Absolute Error : 11.3274
Test Set Mean Absolute Error : 16.7712
```

Code Snippet 6: Calculating Mean Absolute error of the model

We then check the net mean absolute error that the algorithm is giving (in litres/day) as compared to the actual data. We see that the net error(on the target variable – rate of loss of lube oil)) on the training set is close to 11.325 litres/day while on the test set is 16.7 litres/day.

At this stage, our prototype has been formed and now we need to use this prototype on another dataset to check validity.

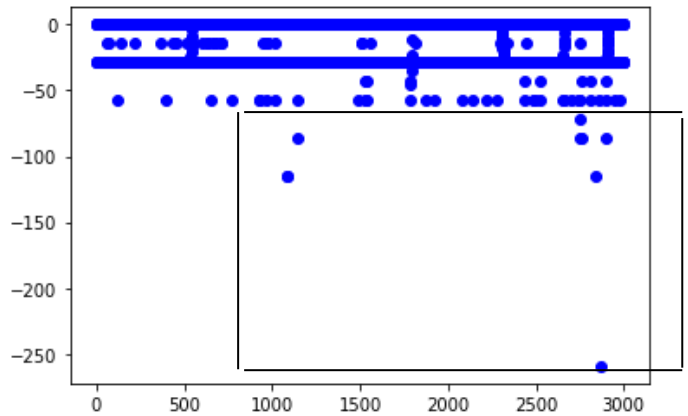
```
from sklearn.externals import joblib
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

model = joblib.load('Project_Model.pkl')

df = pd.read_csv('FinalData.csv')
df2 = pd.read_csv('FinalTargetData.csv')

predict_oil_loss = model.predict(df)
predict_oil_loss = predict_oil_loss/20
dp = pd.DataFrame(data=predict_oil_loss)
dp.to_csv("Prediction.csv")
```

Code Snippet 7: Using the model on a new data to check validity

We then export our Decision Tree Algorithm to a new file and use a new dataset to check validity of the model and plot the graph of the algorithm's prediction v/s the actual data

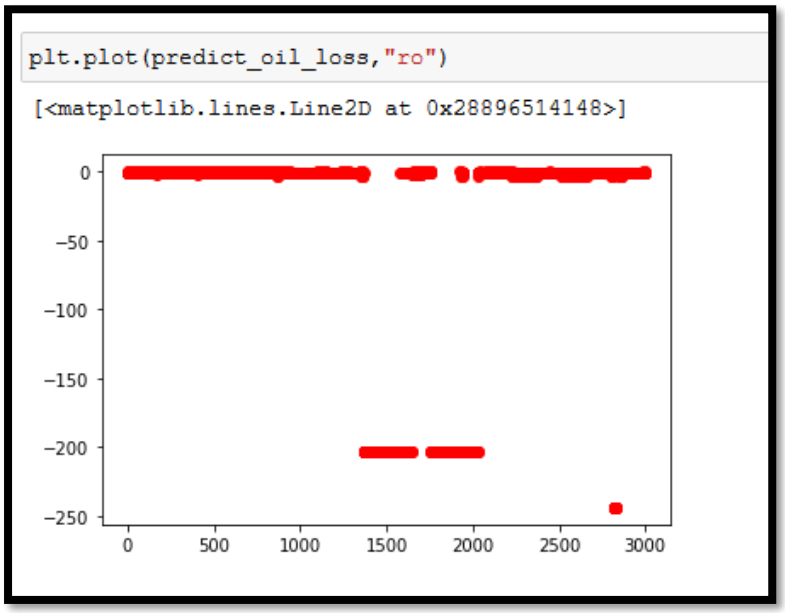
```
#plt.xticks(np.arange(0, 3000, 50.0))
#plt.plot(predict_oil_loss,"ro")
plt.plot(df2["Target"],"bo")
#plt.rcParams['figure.figsize'] = [400,100]
```

```
[<matplotlib.lines.Line2D at 0x288964abcc8>]
```



Code Snippet 8: Visualizing the target variable (Actual Data)

The graph in blue indicates the actual values of the Target Variable, while the box indicates the places where the WGC came close to failing or actually failed, since the rate of loss of lube oil increased tremendously. However, these blue dot are individual points as immediate action was taken to curb the failure.



Code Snippet 9: Visualising the target variable (Predicted Data)

The graph in red indicates where failure occurs according to the algorithm. Since we cannot factor in manmade intervention into the algorithm, the lines of failure are continuous. But the validity of the algorithm is confirmed as the point where definite failure is occurring (the bottom right corner of both the graphs) has been plotted extremely accurately by the model. While I haven't achieved all that I had hoped to achieve, I still plan on finding which parameter affects the Target Variable how much, I think I am close to my goal of predicting failures and being of help even a little using my project.

Conclusion

In an increasingly competitive world and market, every second of your time remains crucial. In such circumstances, especially for a large company, the failure of equipment pits them way behind their competitor. In that situation, predictive maintenance, the crux and core of my project come into picture.

Since I have not been able to visit the refinery during this pandemic, it has been difficult to completely visualise working of Wet Gas Compressor, where the sensors are placed and how each component/system of the WGC works, but with the theoretical knowledge about the parts, the methods undertaken in this project are industry intensive. They help us obtain fairly accurate inferences about the state of the system.

The main objective of the project remains to be able to predict when the Wet Gas Compressor will fail taking into account its Lube Oil Level as the primary dependent variable for the Machine Learning Algorithm, and while I have been able to predict certain failures, especially those which are large scale failure, I am yet to find which parameter affects the failure to what extent. I shall continue to work to find the impact each parameter has on the failure process even after the submission of this report.

With the use of Supervised Learning, i.e being able to predict when a device will fail according to Historical Data (when the device has failed in the past), we will be able to predict when the device will fail in the future. This will save the company enormous amount of time and energy, since they will know when the device is in need of maintenance before it stops functioning. An increasingly important project in the current condition, where each second tonnes of finished product is created, I hope I can help the company in any way using my project.

List of Figures, Tables and Code Snippets:

Figures:

Sr. No	Title	Page Number
1	Process flow diagram of Mumbai Refinery	11
2	Flow Diagram of FCCU	12
3	Schematic Diagram of a Wet Gas Compressor	13
4	Major Parts of a Wet Gas Compressor	14
5	Schematic Diagram of Supervised Learning (used in the project)	22
6	Schematic Diagram of Decision Tree Regression Algorithm	26
7	Decision tree Regression Example	27

Tables:

Sr. No	Title	Page Number
1	Snippet of data after being sorted	18
2	Snippet of data before pre-processing	20
3	Snippet of data after pre-processing	20

Code Snippets

Sr. No	Title	Page Number
1	List of all the libraries imported and used	18
2	Python Code for pre-processing data	20
3	Taking X as independent variable and y as Dependent variable	22
4	Splitting data into training data and test data	28
5	Creating a Decision tree algorithm	28
6	Calculating Mean Absolute error of the model	28
7	Using the model on new data to check validity	29
8	Visualising Target Variable(Actual Data)	29
9	Visualising Target variable(Predicted Data)	30

Appendix

All the code and supplementary material used in the entire project have and will be uploaded to my Github profile. The initial compiled dataset, the post cleaning dataet along with the Python code for the same has already been uploaded to the website.
https://github.com/raghav810/PS1

All the used Reference material, PDFs, reaserch papers as well as the material given by my instructor will also be uploaded on the same link**.**

References

1. Data Science with Python Specialization University of Michigan, Coursera
2. Machine Learning by Andrew Ng, Stanford University, Coursera
3. https://stackoverflow.com/questions/tagged/pandas
4. www.bigdatarepublic.nl
5. Manual on Wet Gas Compressor given by my mentor titles MAC-WGC
6. BPCL website for basic refinery processes
7. www.pandas.pydata.org
8. https://fcc-refinery-training-network.blogspot.com/2017/01/wet-gas-compressor.html
9. www.gdcoder.com

Glossary

Data Cleansing: The process of reviewing and revising data to delete duplicate entries, correct misspelling and other errors, add missing data and provide consistency.

Data Science: A discipline that incorporates statistics, data visualization, computer programming, data mining, machine learning and database engineering to solve complex problems.

Data Set: A collection of data, very often in tabular form.

Algorithm: A mathematical formula or statistical process used to perform analysis of data.

Predictive Analytics*:* Using statistical functions on one or more data sets to predict trends or future events. In big data predictive analytics, data scientists may use advanced techniques like data mining, machine learning and advanced statistical processes to study recent and historical data to make predictions about the future. It can be used to forecast weather, predict what people are likely to buy, visit, do or how they may behave in the near future.

Machine Learning: A method of designing systems that can learn, adjust and improve based on the data fed to them. Using predictive and statistical algorithms that are fed to these machines, they learn and continually zero in on "correct" behavior and insights and they keep improving as more data flows through the system.

Normal Distribution: The most important continuous probability distribution in statistics is the normal distribution (a.k.a. Gaussian distribution). The normal distribution is the familiar bell curve. Once m and s are specified, the entire curve is determined.

Predictive Modelling: The process of developing a model that will most likely predict a trend or outcome.

Variance: The average squared deviation for all values from the mean: