Accenture Applied Intelligence

Women in Data Science Accelerator

Project Information Pack

February 2019

Background

PRODCO – Production Company

PRODCO is a multinational Manufacturing Company operating in a European country. They produce high tech products for different industries.

PRODCO installed a new integrated production line last year for more than €100million that started production 9 months ago. However, since switching on the line, they have experienced high defect rates which prevented them reaching their production targets, causing massive additional costs.

PRODCO's engineering team have identified that the high defect rate is due to mechanical issues with the production process and not a failure of quality control on raw materials. They have been trying to fix the problems on their production line, but the general defect rate remains high.

PRODCO's engineering team have tried several approaches to reduce the number of finished items that need quality checks, but they were unable to detect defects in some shipped production batches before they left the factory. This resulted in high volumes of complaint claims from customers. As a result, PRODCO's Quality Control team is having to inspect every item that comes off the production line, resulting in additional personnel costs and backlogs on orders.

PRODCO have contracted your team to help them to reduce the impact these issues are having on their business. They have suggested two areas of focus for your team:

1) A root cause analysis as a Proof of Concept for a defect detection model, which will better enable their engineering team to pinpoint the location of the issues so that they can perform required maintenance and repairs. The engineering team have requested that the model be easily explained and 'analytics jargon free'.

2) A Proof of Concept analytics model to predict which items coming off the production line may be defective, and the type of defect that item is likely to have. They hope to reduce the cost of quality checks by only performing checks on items with a high probability of defect.

PRODCO are aware that you are only available for a short window of time and would prefer to see a comprehensive solution to one of their issues rather than an incomplete solution to both. They have therefore left it to the team's discretion as to whether they deliver a solution for one or both focus areas.

PRODCO's executive team are considering implementing similar models across their factories. The executive team are not strangers to analytical models, and while they would like the models to be easily explained to the Quality Control and Engineering teams, their primary goal is to use this Proof of Concept to showcase the potential uses and effectiveness of multivariate analyses and advanced analytics.

Current Defect Detection Process

PRODCO's manufacturing process allows engineers to track the path that individual items and component parts take through the production line. Sensors in different zones track environmental measures and the time taken for an item to move through the zone. PRODCO's current defect detection process is based on a visual inspection of the item at the end of the production line, which classifies each product into 'Pass' or one of four defect types. To perform maintenance or remediation work, Engineers use the combination of defect type and the data showing path, environment, and process durations for each item to try to identify the root cause of the defect.

Individual types of product (identified using Stock Keeping Unit (SKU) IDs) may show different defect rates due to slight differences in their physical manufacturing process, but all SKUs show some incidence of defects.

Project Requirements

The project has two equally important parts:

1) Data Analysis and Analytical Modelling
2) Business Value and Recommendations

In Week 6, you will be asked to present your projects to a judging panel and an audience of your peers. The judging panel will not have seen your project work week-on-week and may not be familiar with the exact problem that you are trying to solve. It is therefore recommended that you cover the following in your presentation:

- Client situation & focus area chosen by the team
- Analytical approach taken and why (model used? Rationale & logic for not using something else)
- High level model outputs and recommendations for how the client can use/implement

Your team will have 15 minutes including Q&A to present your solution. You are free to choose the medium that you use to present – Accenture will provide a screen and a microphone, anything else you need to bring yourselves.

However, we recommend using a digital format for any visualisations (PowerPoint slides, R-slides, Tableau or QlikView dashboards can be some examples), so that they are easily visible to your audience.

Remember! As a Data Scientist you are the translator of the data into business language and useful business insight.

Data

PRODCO keeps production data about each item that has gone through their manufacturing process in a single big table. This table contains: SKU ID, path (position & environment in zone) through the production line, time elapsed in and between zones, and pass/defect type recorded by Quality Control.

For the purposes of this POC, PRODCO have provided:

- Production Data
  - For Week 1, PRODCO have provided a sample dataset of 10,000 records for you to explore
  - In Week 2, PRODCO will provide the full 6-month dataset of 500,000 records
- Data dictionary
- Spreadsheet containing maintenance cost, product value, and QC man-hours

Support

As part of the Project Clinics each week, you will be assigned a Project Mentor and have access to Client Subject Matter Experts (SMEs) and Data Science SMEs.

The Project Mentor will act as project lead and help you determine your tasks and project plan for each week. You will have 10 minutes of time each week with a Client SME & Data Science SME (20 mins total per week)

The Client SMEs will be able to answer questions about the client background and production process, and help to clarify what the client is looking for from the solution you present. The Data Science SMEs will be able to answer questions about methodologies and suggest techniques. It is suggested that you prepare your questions before consulting with the SMEs as they will be splitting their time across teams.

Accenture Data Scientists are available for consultation during the Project Clinics to answer questions and work through problems and analyses with you.

In Week 5, Storytelling experts will be available during the Project Clinic to answer presentation questions.

Sample Project Flow

There are many different ways to deliver what the client requires. You may use Excel, Python, R or any analytics tool that you are comfortable with to perform your analysis (see some Excel, Python and R helpful tips below).

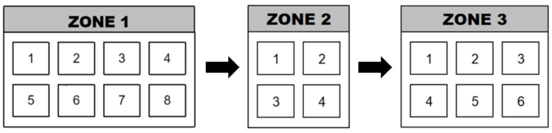
The project flow below is intended to provide some helpful pointers – it is not a complete or mandatory checklist

1. Assess the quality of the data, try to remove the redundant features and decide how to treat missing values if they exist.
2. Visualize the data and perform univariate, bivariate and multivariate analyses, as applicable.
3. Perform defect detection or prediction using advanced analytics techniques (examples of commonly used modelling techniques can be found in the Help Sheet).
4. Decide on the two or three key findings from your solution. You should be able to explain why you took the approach you did, and the benefits/savings to be gained by employing this model.
5. Talk about the next steps that can be taken by the company to improve their business such as:
    a. Are there any further insights not strictly relevant to the defect analysis which may be of interest to PRODCO? Suggest how these insights can be applied to other areas of the business.
    b. Consider what other data PRODCO, as a product manufacturer, collects. Surplus to the data already provided to Accenture by PRODCO, what other internal and external data sources could be utilised for any future analytical models, production or otherwise?

You will present results from the analysis in the final presentation to PRODCO. Remember that two or three key findings is better than a lot of irrelevant information. You may include graphs and/or statistics. Tell the story of the data and explore how variables may be relevant to each other, how the models and the data that PRODCO already collects can be utilized in solving the current problem and applied to future ones.

Production Process Description

Production Process



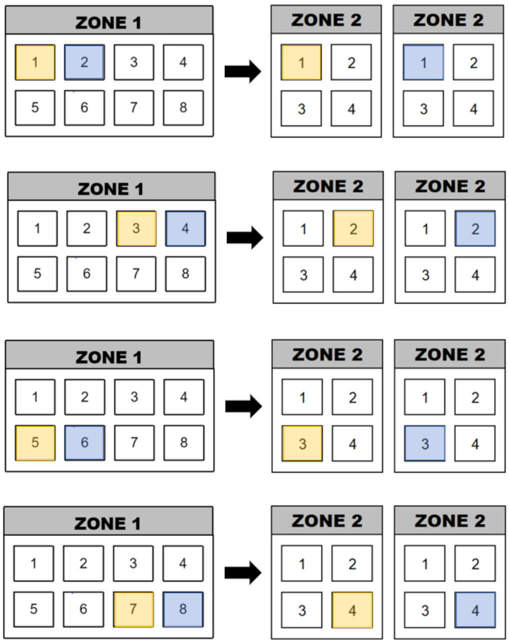
The production line spans across 3 separate Zones as shown above. Zone 1 is composed of 8 different cavities, Zone 2 is composed of 4 different cavities and Zone 3 is composed of 6 different cavities. Each cavity can hold one item. The cavities sit on a pallet as they move through the production process.

An item enters the manufacturing process in Zone 1. Here, it sits in one cavity until the Zone 1 manufacturing stage has been completed. Next, the item is transferred to a cavity on Zone 2. Again, it stays in this cavity until the Zone 2 manufacturing stage has been completed. Finally, the item is transferred to a cavity on Zone 3 and stays in this cavity until the Zone 3 manufacturing stage has been completed.

The item is then inspected for defects and classified into 'Pass' or one of four defect types.

An individual item can only take fixed paths through the production process which are outlined below.

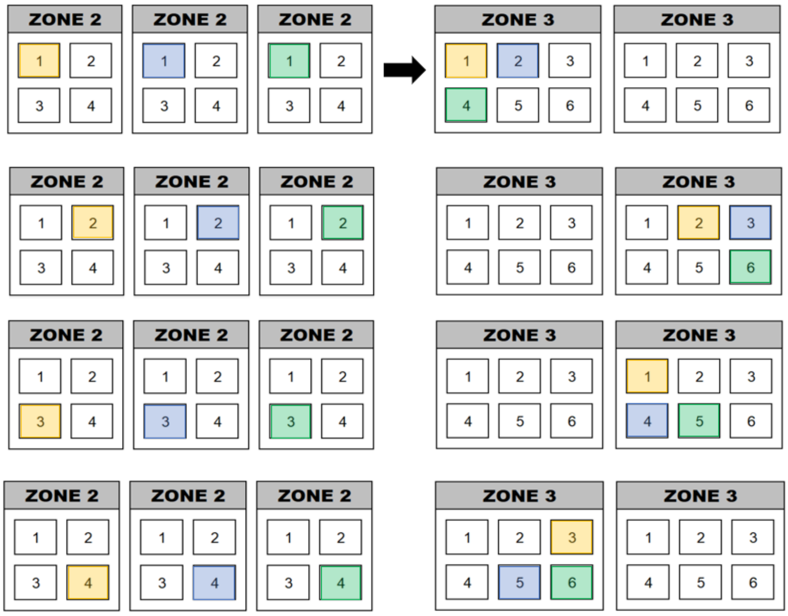
Moving From Zone 1 to Zone 2



An individual item can only take fixed paths between Zone 1 and Zone 2. There is a 2:1 relationship between Zone 1 and Zone 2 meaning that items coming from Zone 1 will fill 2 separate pallets in Zone 2 in the following manner:

If an item begins production in cavity 1 of Zone 1, then in Zone 2 it will sit in cavity 1. Similarly, if an item begins production in cavity 2 of Zone 1, then in Zone 2 it will sit in cavity 1 of a separate pallet in Zone 2. Each of these fixed paths that an item can take between Zone 1 and Zone 2 are shown above.

Moving from Zone 2 to Zone 3



Similarly, an individual item can only take fixed paths between Zone 2 and Zone 3. There is a 3:2 relationship between Zone 2 and Zone 3 meaning that 3 full pallets of items coming from Zone 2 are required to fill 2 separate pallets in Zone 3.

If an item begins production in cavity 1 of Zone 2, then in Zone 3 it will sit in either cavity 1, cavity 2 or cavity 4. Three pallets from Zone 2 are needed to fill the two pallets in Zone 3. Similarly, if an item begins production in cavity 2 of Zone 2, then in Zone 3 it will sit in either cavity 2, cavity 3 or cavity 6. Each of these fixed paths that an item can take between Zone 2 and Zone 3 are shown above.

Help Sheet

Excel Helpful Tips

Joining datasets: VLOOKUP

a) Begin with your primary dataset. You need to merge the additional data into this dataset using an **ID** field as the common variable.

b) Type =VLOOKUP into a cell.

c) Lookup_value: Select cell containing **ID** on current dataset
Table_array: Highlight all data on the dataset containing the data you want to bring in to your primary dataset
Col_index_num: Column you want to merge.
Range_lookup: Type “false”

d) Repeat for each applicable field.

Analysing the data: Pivot Table

a) To see some important statistics, Pivot table can be a very useful and fast tool.

b) Select a blank cell. Insert > Pivot Table. Select all data.

c) Play around with the pivot to get some insightful information.

d) By default, Pivot data shows SUM for numerical data and COUNT for string/categorical data. You can change this by right click menus.

e) If you decide to present a pivot table result, don’t forget to change the formatting into a more presentable one by Design and Analyse tabs that appear on the Ribbon when you select Pivot table.

R Helpful Tips

We advise to use RStudio, if you decide to use R for your task.

Environment Set Up/Data Read

- Set Working Directory
  - setwd(“file_path/file_path/…..”)
- Import CSV File
  - Data <- read.csv(“file_name.csv”, stringsAsFactors = F, strip.white = T, header = TRUE)
  - Note that file must be stored in the folder specified in your “setwd” statement, otherwise you will have to give the full file path.
- Export a Dataset as a CSV file
  - write.csv(Data, “file_name.csv”, row.names = F)

Data Preparation

- View data in the navigation window
  - View(Data)
- Check column names of a dataset
  - colnames(Data)
- Check variable formats of a dataset
  - lapply(Data, class)
- Look for the first row of data
  - head(data)
- Reformat a variable in a dataset
  - Data$variable_name <- as.Date(Data$variable_name)
  - as.Date used as an example. Can also be as.numeric, as.character, etc.
- Assign Column names to a dataset
  - names(Data) <- c(“name_1”, “name_2”, ……., “name_n”)
- Drop column numbers i to j from a dataset
  - Data <- Data[,-(i:j)]
- Get summary statistics for a dataset
  - install.packages(‘psych’)
  - library(psych)
  - describe(Data)
- Return the unique values of a variable in a dataset
  - unique(Data$variable_name)
- Remove duplicates from a dataset
  - Data_deduped <- unique(Data)
  - To de-dupe only based on certain variables use
    - Data_deduped <- Data[!duplicated(Data$var1, Data$var2, ..),]

Transform

- Using if else to create a new variable, based on the value of an original variable (similar to IF function in Excel)
  - Data$New_Var <- ifelse(condition, value if true, value if false)
- Merging data from two datasets
  - Data_merged <- merge(Dataset_1, Dataset_2, “common_key_variable”)
- For faster data transformations you can use dplyr package. Some of the most useful functions are:
  - select
  - filter
  - left_join

- inner_join
- case_when
- summarize
- mutate

Model

- Fit a linear regression model to data
    - Library(stats)
    - regmodel <- lm(Data$Dependent_Var ~ Data$Independent_Var1, ….., Data$Independent_VarN)
- fit a logistic regression to the data
    - logregmodel <- glm(Data$Dependent_Var ~ Data$Independent_Var1 + ….. , family = 'binomial')
- For more sophisticated algorithms you can install caret package.
    - A complete look into caret: http://topepo.github.io/caret/index.html
- For using ANNs you can use TensorFlow and Keras in R. As Keras is higher level API for fast prototyping. you are encouraged to use Keras for this case study.
    - https://keras.rstudio.com/
- If you decide to use a black box model you can use the package LIME:
    - https://www.r-bloggers.com/explain-explain-explain/
    - https://uc-r.github.io/lime
    - https://cran.r-project.org/web/packages/lime/vignettes/Understanding_lime.html

Visualisation

- Scatterplots
    - plot(Data$variable_name1, Data$variable_name2)
- Histograms
    - hist(Data$variable_name)
- Density Plots
    - dens <- density(Data$variable_name)
    - plot(dens)
- Dot Plots
    - dotchart(Data$variable_name)

For more presentable plot designs please see ggplot2.

http://r-statistics.co/Complete-Ggplot2-Tutorial-Part1-With-R-Code.html

https://www.statmethods.net/advgraphs/ggplot2.html

https://www.kaggle.com/andyxie/beginner-ggplot2-tutorial

Python Helpful Tips

Installing Python with Anaconda would help you install all packages needed for this case study. Jupyter Notebooks or Spyder can be considered useful environments.

Environment Set Up/Data Read

- Set Working Directory
  - Import os
  - os.chdir("file_path/file_path/…..")
- Import CSV File
  - import pandas as pd
  - Data <- pd.read_csv("file_name.csv", header = True)
  - Note that file must be stored in the folder specified in your "chdir" statement, otherwise you will have to give the full file path.
- Export a Dataset as a CSV file
  - To be able to use this function your data must a Pandas DataFrame
  - pd.to_csv(Data, "file_name.csv)

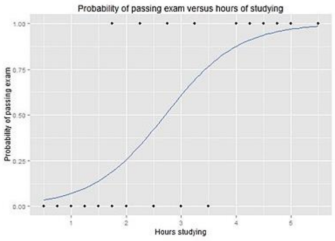
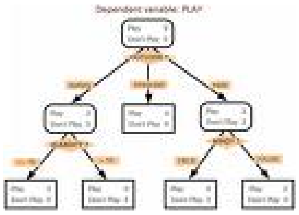
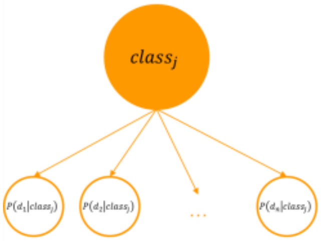
Data Preparation

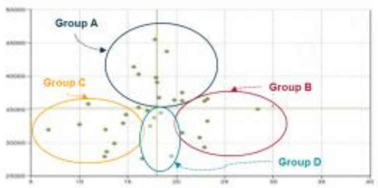
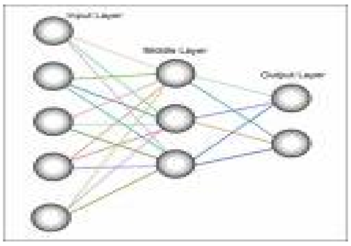
- Check column names of a dataset
  - Data.columns
- Check variable formats and number of missing values of a dataset
  - Data.info()
- Look for the first rows of data
  - data.head()
- See the basic statistics of each column
  - Data.describe()
- Reformat a variable in a dataset
  - Data.infer_objects()
  - Data['variable_name'] = Data['variable_name'].as_type('int64'|'category'|'string'…)
  - Data['variable_name'] = Data['variable_name'].to_datetime()
- Assign Column names to a dataset
  - Data.columns <- ["name_1", "name_2", ……., "name_n"]
- Drop column numbers i to j from a dataset
  - import numpy as np
  - import pandas as pd
  - Data <- Data.as_matrix.T
- Return the unique values of a variable in a dataset
  - Data['variable_name'].unique
- Remove duplicates from a dataset
  - Data.drop_duplicates()

Transform

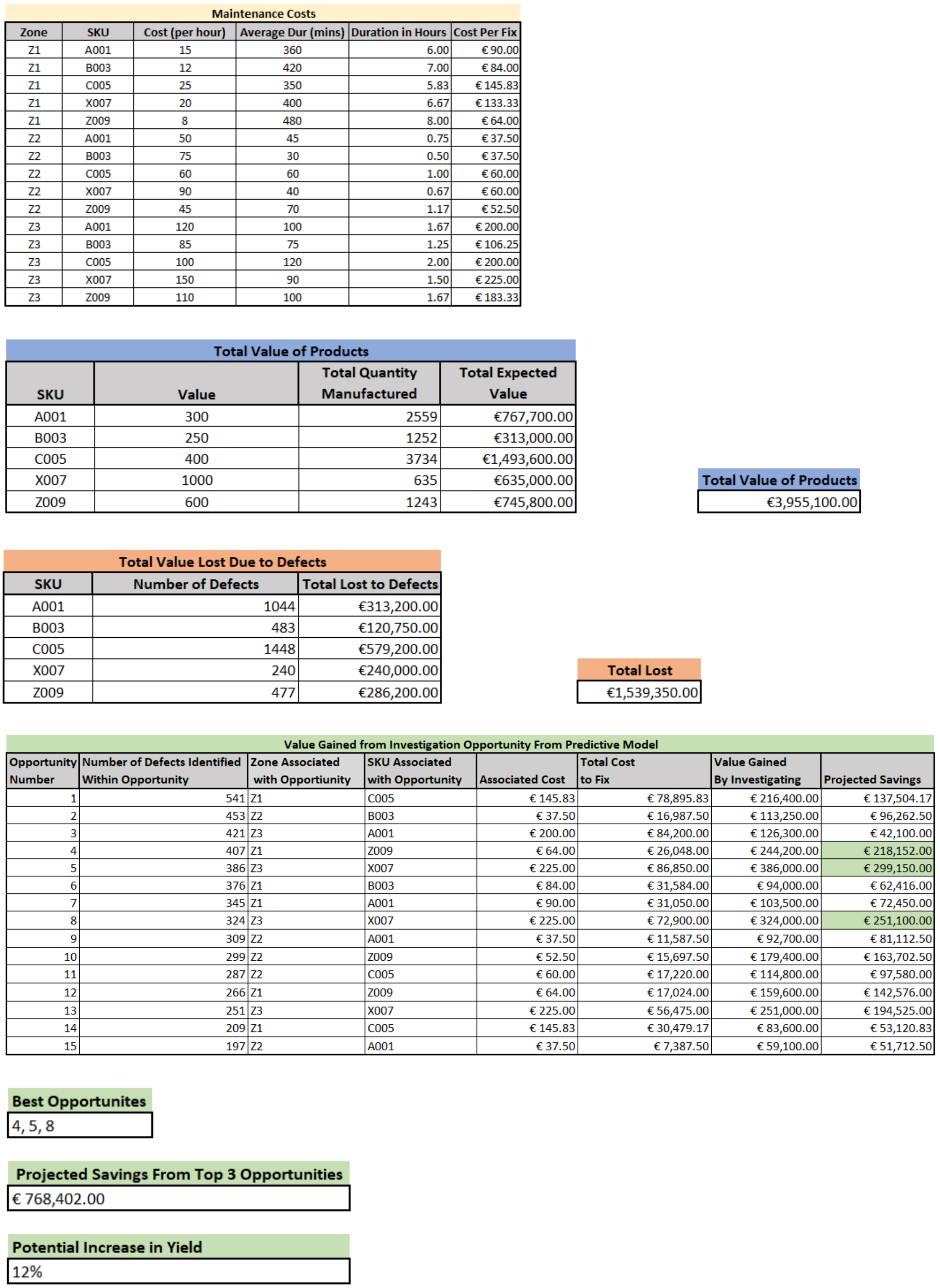
- Using if else to create a new variable, based on the value of an original variable
  - DataFrame.assign(new_variable_name = lambda x: some_transformation(x))
- Merging data from two datasets
  - DF_Merged = DF1.merge(DF2, on = 'common_variable', how = ('left', 'inner' etc.)
- You can generate Pivot tables from Python DataFrames by
  - DF.pivot_table()
  - For usage details: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.pivot_table.html

Models

Predictive Modelling Technique	Description	Main Considerations and Assumptions	Model Output
Regression Modelling	 This technique creates a mathematical relationship between different variables for predicting the likelihood of the target variable occurring.	Like many techniques, regression modelling can result in "overfitting", meaning the model becomes susceptible to fluctuations in the data. The output provided by this technique is typically accessible for non-Data Scientists.	This technique produces ranked scores for subjects, with larger numbers typically indicating a greater likelihood of the target variable. The output of this approach allows for traceability of variables leading to a high likelihood of the target variable.
Decision Trees	 Decision trees, such as Classification and Regression Trees (CART) and Chi-Square Automatic Interaction Detector (CHAID) help to classify possible predictive variables into the likelihood of the target variable occurring by iteratively dividing the data into partitions based on yes or no conditions. There are many different criteria for calculating how to divide into partitions, which should be tested on the data to determine which produces the best result.	This technique requires statisticians to "ask the right questions" and divide using the right conditions, as this can significantly impact the model results. The output provided by this technique is typically very accessible for non-Data Scientists.	The technique classifies subjects into a group rather than assign a specific ranked score. The output of this approach allows for traceability of variables leading to a high likelihood of the target variable.
Bayesian Classification	 $class_j$ $P(d_1\|class_j)$ $P(d_2\|class_j)$ … $P(d_n\|class_j)$ This technique, also known as naïve Bayesian, is based upon prior probabilities and an inherent assumption that possible predictive variables contribute independently to their likelihood of the target variable occurring. Even though this independence assumption may not be practical and accurate, this technique can often produce a very useful result.	This assumption of independence can decrease the usability and accuracy of the result. Furthermore, given that it is based on a complex statistical technique, it can be unintuitive.	The technique classifies subjects into a group (or class) based on most probable classification. Classes can be defined based on the modeller's goals (such as individual target variable class/criticality of outcome).

Predictive Modelling Technique	Description	Main Considerations and Assumptions	Model Output
Nearest Neighbours	 This technique provides another approach to classification, where subjects with the greatest similarity to each other (quantified as smallest "distance") and the greatest dissimilarity to others (largest distance) are grouped together. This technique can be implemented using a wide range of different approaches to quantify "distance".	Data Scientist must select the number of groups to be defined by the model – a larger number generally results in a more robust model but with less distinct differences between groups. Data Scientist must also reduce the amount of data attributes available to the model.	This technique classifies subjects into a group of similarly behaving groups of the likely target variable value. The output of this approach allows for traceability of variables leading to a high likelihood of a target variable.
Ensemble Models	This technique provides a formal structure to combine multiple techniques to produce a more robust, more accurate result. There are many commonly used combinations, generally focusing on either averaging the results across different models (known as averaging) or combining the results to address specific inaccuracies (known as boosting).	These techniques generally produce a more complex result. This can be unintuitive and can require more system resources to re-execute and refresh. While the results usually can be explained by a variable importance graph, they are not as interpretable as linear models.	This technique would typically be applied to produce ranked scores, with larger numbers indicating a greater likelihood of the target variable.
Artificial Neural Networks	 Neural networks uncover and model highly complex relationships between pieces of data. At a very high level, these models identify layers of neurons (representing characteristics or behaviours of a subject), quantified weights to indicate the importance of these neurons, synapses (representing interconnection of weighted neurons), learning rules to adjust weights and synapses, and an activation function for interpreting the result.	Widely acknowledged as one of the most complex techniques, neural network models produce a complex, unintuitive result, which most often must be treated like a "black-box". This can significantly reduce the usability of the result and the extent to which it supports investigations.	This technique produces ranked scores for subjects, with larger numbers typically indicating a greater likelihood of the target variable.

Sample Value Assessment



Maintenance Costs

Zone	SKU	Cost (per hour)	Average Dur (mins)	Duration in Hours	Cost Per Fix
Z1	A001	15	360	6.00	€ 90.00
Z1	B003	12	420	7.00	€ 84.00
Z1	C005	25	350	5.83	€ 145.83
Z1	X007	20	400	6.67	€ 133.33
Z1	Z009	8	480	8.00	€ 64.00
Z2	A001	50	45	0.75	€ 37.50
Z2	B003	75	30	0.50	€ 37.50
Z2	C005	60	60	1.00	€ 60.00
Z2	X007	90	40	0.67	€ 60.00
Z2	Z009	45	70	1.17	€ 52.50
Z3	A001	120	100	1.67	€ 200.00
Z3	B003	85	75	1.25	€ 106.25
Z3	C005	100	120	2.00	€ 200.00
Z3	X007	150	90	1.50	€ 225.00
Z3	Z009	110	100	1.67	€ 183.33

Total Value of Products

SKU	Value	Total Quantity Manufactured	Total Expected Value
A001	300	2559	€767,700.00
B003	250	1252	€313,000.00
C005	400	3734	€1,493,600.00
X007	1000	635	€635,000.00
Z009	600	1243	€745,800.00

Total Value of Products
€3,955,100.00

Total Value Lost Due to Defects

SKU	Number of Defects	Total Lost to Defects
A001	1044	€313,200.00
B003	483	€120,750.00
C005	1448	€579,200.00
X007	240	€240,000.00
Z009	477	€286,200.00

Total Lost
€1,539,350.00

Value Gained from Investigation Opportunity From Predictive Model

Opportunity Number	Number of Defects Identified Within Opportunity	Zone Associated with Opportiunity	SKU Associated with Opportunity	Associated Cost	Total Cost to Fix	Value Gained By Investigating	Projected Savings
1	541	Z1	C005	€ 145.83	€ 78,895.83	€ 216,400.00	€ 137,504.17
2	453	Z2	B003	€ 37.50	€ 16,987.50	€ 113,250.00	€ 96,262.50
3	421	Z3	A001	€ 200.00	€ 84,200.00	€ 126,300.00	€ 42,100.00
4	407	Z1	Z009	€ 64.00	€ 26,048.00	€ 244,200.00	€ 218,152.00
5	386	Z3	X007	€ 225.00	€ 86,850.00	€ 386,000.00	€ 299,150.00
6	376	Z1	B003	€ 84.00	€ 31,584.00	€ 94,000.00	€ 62,416.00
7	345	Z1	A001	€ 90.00	€ 31,050.00	€ 103,500.00	€ 72,450.00
8	324	Z3	X007	€ 225.00	€ 72,900.00	€ 324,000.00	€ 251,100.00
9	309	Z2	A001	€ 37.50	€ 11,587.50	€ 92,700.00	€ 81,112.50
10	299	Z2	Z009	€ 52.50	€ 15,697.50	€ 179,400.00	€ 163,702.50
11	287	Z2	C005	€ 60.00	€ 17,220.00	€ 114,800.00	€ 97,580.00
12	266	Z1	Z009	€ 64.00	€ 17,024.00	€ 159,600.00	€ 142,576.00
13	251	Z3	X007	€ 225.00	€ 56,475.00	€ 251,000.00	€ 194,525.00
14	209	Z1	C005	€ 145.83	€ 30,479.17	€ 83,600.00	€ 53,120.83
15	197	Z2	A001	€ 37.50	€ 7,387.50	€ 59,100.00	€ 51,712.50

Best Opportunites
4, 5, 8

Projected Savings From Top 3 Opportunities
€ 768,402.00

Potential Increase in Yield
12%