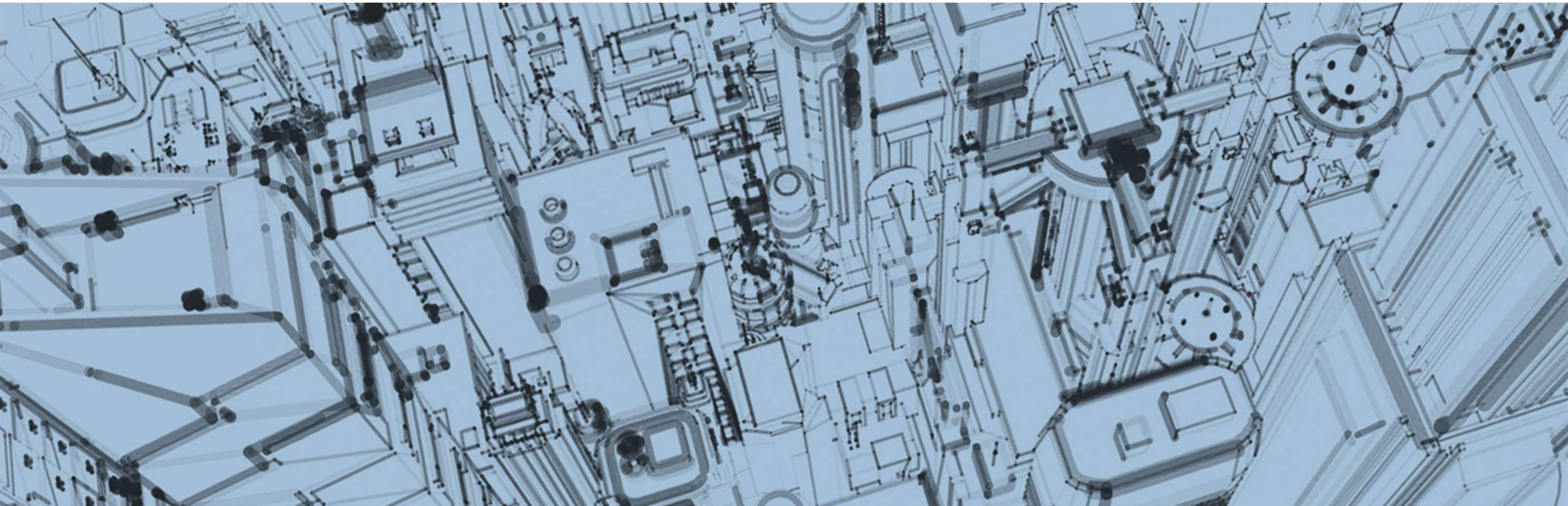

Chemical Plant - Machine Learning

H Cardak – July 2020

Introduction

- Chemical detection platform composed of 14 temperature-modulated metal oxide semiconductor (MOX) gas sensors as in **Figure 1**.
- The sensors are exposed to dynamic mixtures of carbon monoxide (CO) and humid synthetic air in a gas chamber.
- Data collected with features based on input from 14 MOX sensors (MOhm), CO (ppm), Humidity (%r.h.), Temperature (ºC), Flow rate (mL/min), Heater voltage (V).
- **Step_1** is "Modelling Process" to develop a tool, which will provide predictions for the presence of CO (ppm), given features described above within collected data [e.g. validation is also part of first step].
- **Step_2** is "Model Results" and recommendations on commissioning, testing, periodic monitoring and calibrating the solution deployed.
- **Step_3** is "Process Architecture" for the actual deployment of the model.

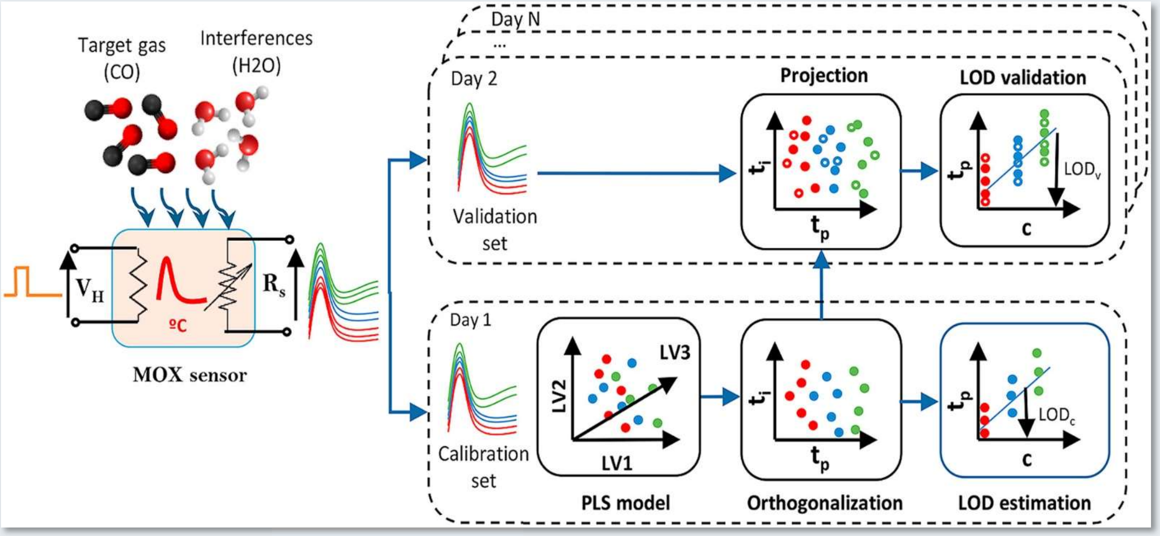

Figure 1: Experiment Setup

[**Ref:** https://www.sciencedirect.com/science/article/abs/pii/S0003267018303702]

Modelling Process

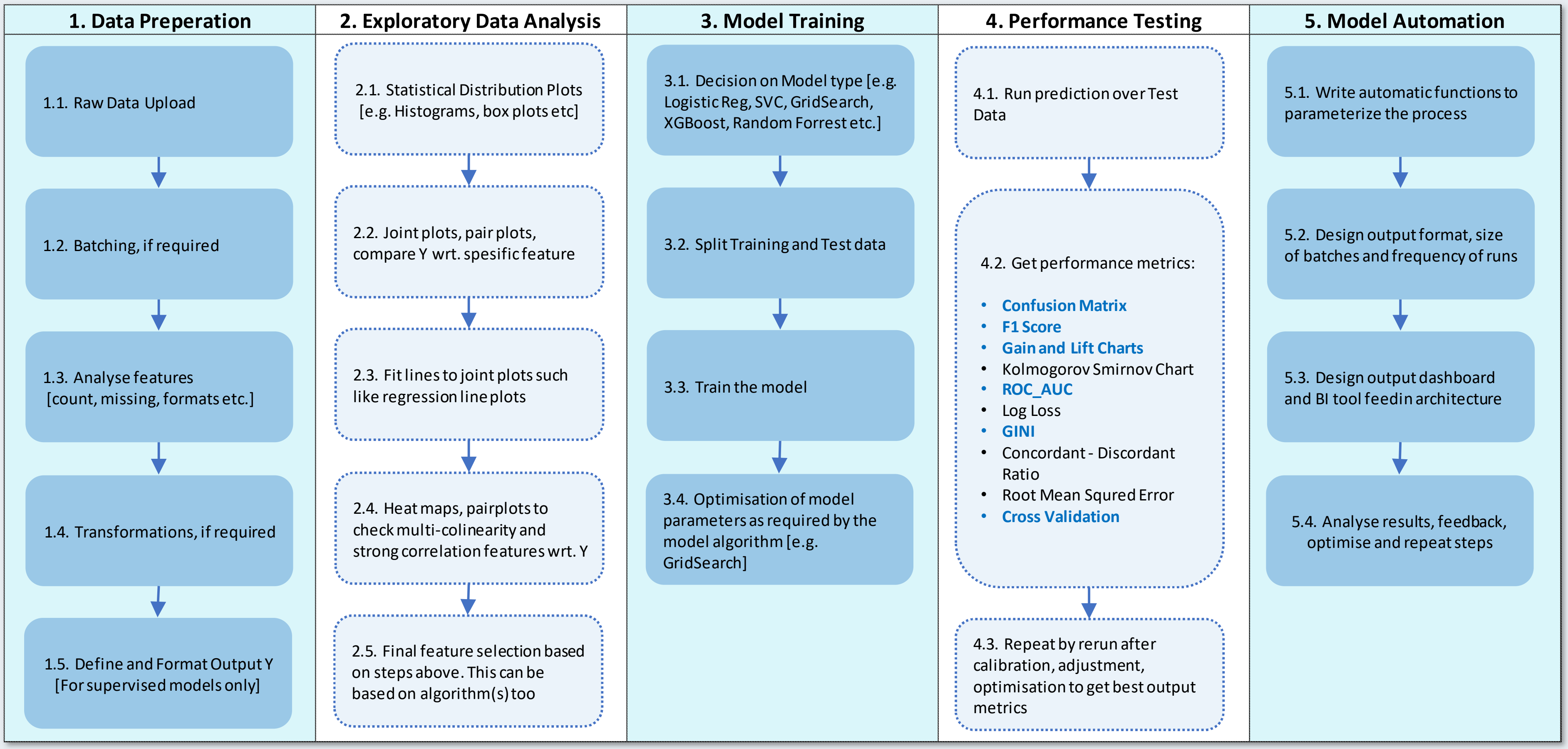

Figure 2: Machine Learning Logic Flow – Process Map

Modelling Process

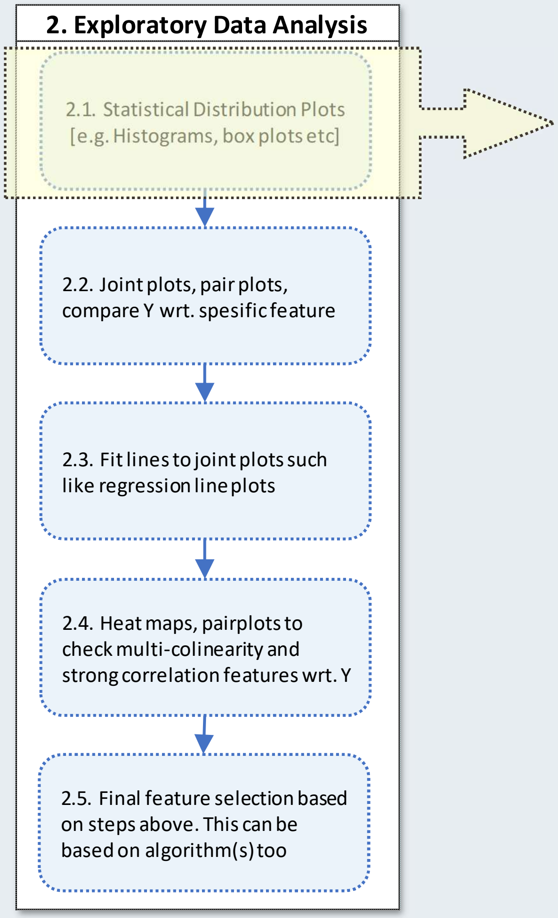

Figure 3: Exploratory Data Analysis Flow

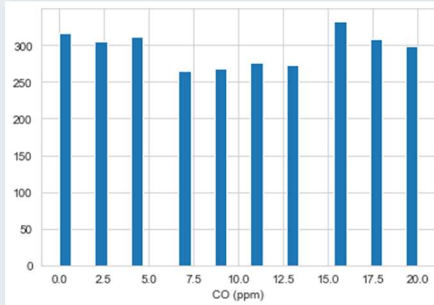

Chart 1: CO (ppm) Histogram

- CO (ppm) is what we are trying to predict, also called Y in ML theory.
- Vertical axis represents number of observations corresponding to the particular horizontal axis CO values.
- Has got uniform distribution, not normally distributed.

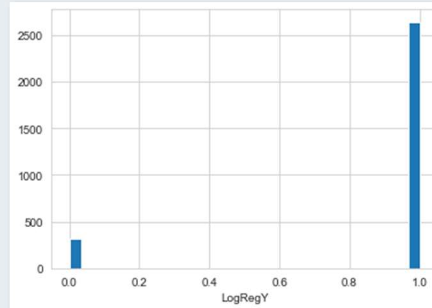

Chart 2: Binary CO (ppm) Histogram

- CO (ppm) is converted into binary Y, if CO >0 then LogRegY=1, else LogRegY=0.
- Histogram indicates around 90% of Y=1.

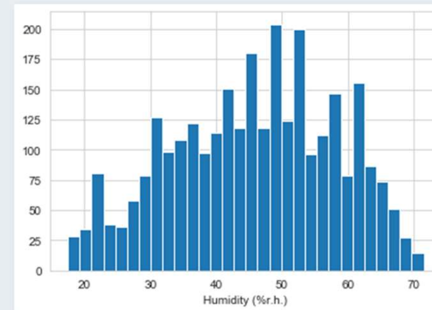

Chart 3: Humidity (%r.h.) Histogram

- Humidity distribution resembles normal distribution.
- Normal distribution found in nature and generally indicates a natural random input.
- Though left tail shows slight thickness.

Modelling Process

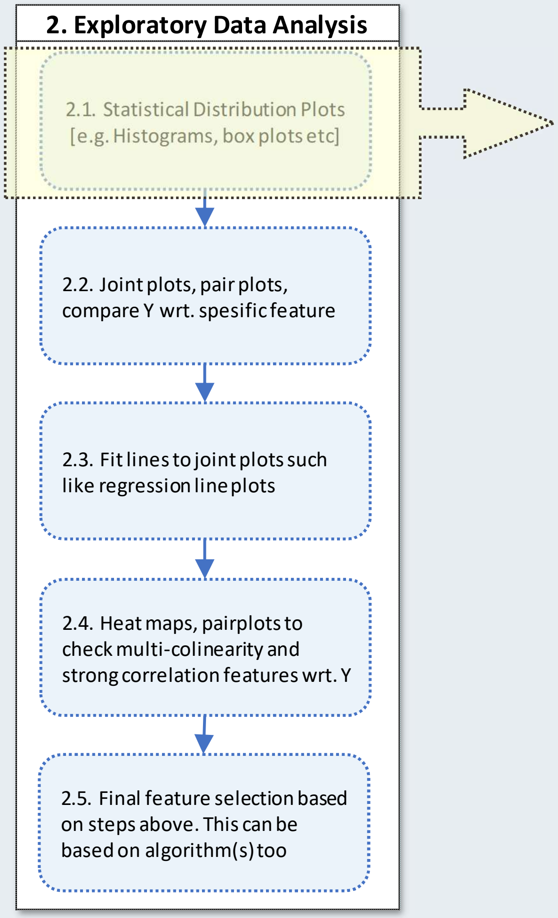

Figure 3: Exploratory Data Analysis Flow

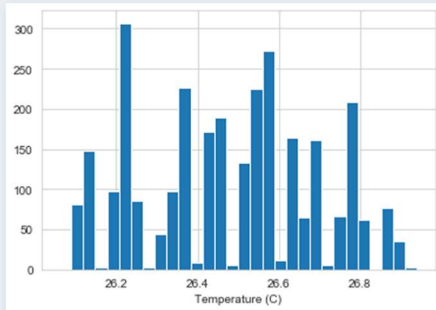

Chart 4: Temperature (C) Histogram

- Temperature input resembles not random.
- Vertical axis represents number of observations and temperature range is around 26C ± 1.
- Since most model algorithms assume natural random distribution [e.g. Normal], this variable may not be a good model feature candidate.

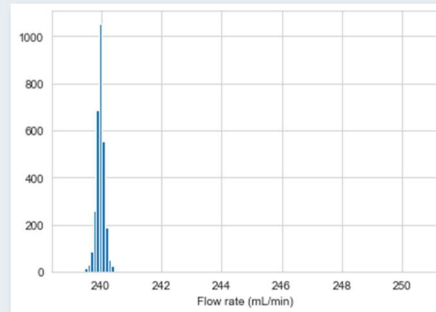

Chart 5: Flow rate (mL/min) Histogram

- Flow rate displays huge skewness to the right [can not be seen in the plot, but have ranges around 250 mL/min].
- In modelling process, outliers [which creates skewness] generally neglected or transformed [Cap and flooring].

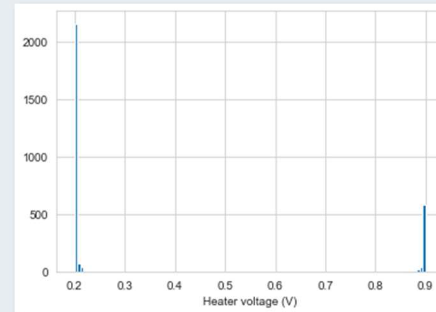

Chart 6: Heater voltage (V) Histogram

- Voltage not random and shows bi-modal [double peak or two waves] behaviour.
- This may be a good un-supervised cluster feature [e.g. used as a filter to separate two experiments or data samples].
- We can still throw into our ML algorithms, which will automatically eliminate if not predictive.

Modelling Process

2. Exploratory Data Analysis

- 2.1. Statistical Distribution Plots [e.g. Histograms, box plots etc]
- 2.2. Joint plots, pair plots, compare Y wrt. spesific feature
- 2.3. Fit lines to joint plots such like regression line plots
- 2.4. Heat maps, pairplots to check multi-colinearity and strong correlation features wrt. Y
- 2.5. Final feature selection based on steps above. This can be based on algorithm(s) too

Figure 3: Exploratory Data Analysis Flow

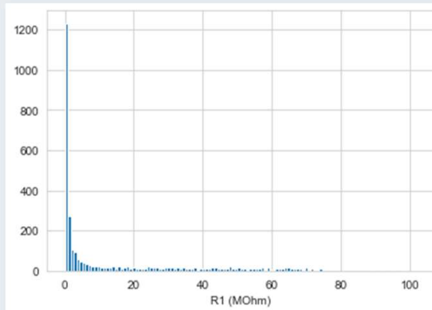

Chart 7: R1 (MOhm) Histogram

- R1 skewed to right with min value 0 and gradually increasing.
- Vertical axis represents number of observations
- Still a long tail and represent un-natural distribution here.

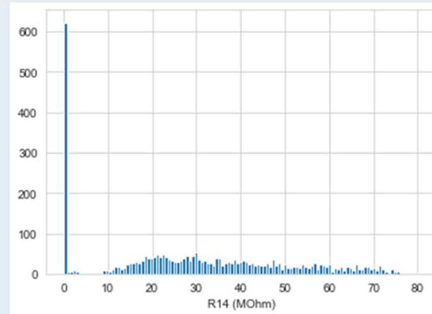

Chart 8: R14 (Mohm) Histogram

- R1 has lots of 0 set.
- But excluding 0 values, it may process a good predictive behaviour since as the rest of the part of the distribution behaves slightly Normal.

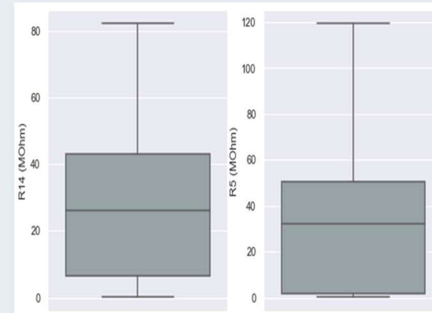

Chart 9: R14 vs R5 Boxplot

- Boxplots displayed here. Middle horizontal line is mode [50th percentile].
- R5 has main values concentrated around 0-50 whereas R14 is near 30.
- Boxplots also show min/max [hence range]. We can identify R14 max is around 85 vs R5 max around 120 Mohm
- Check **Appendix** for remaining feature plots

Modelling Process

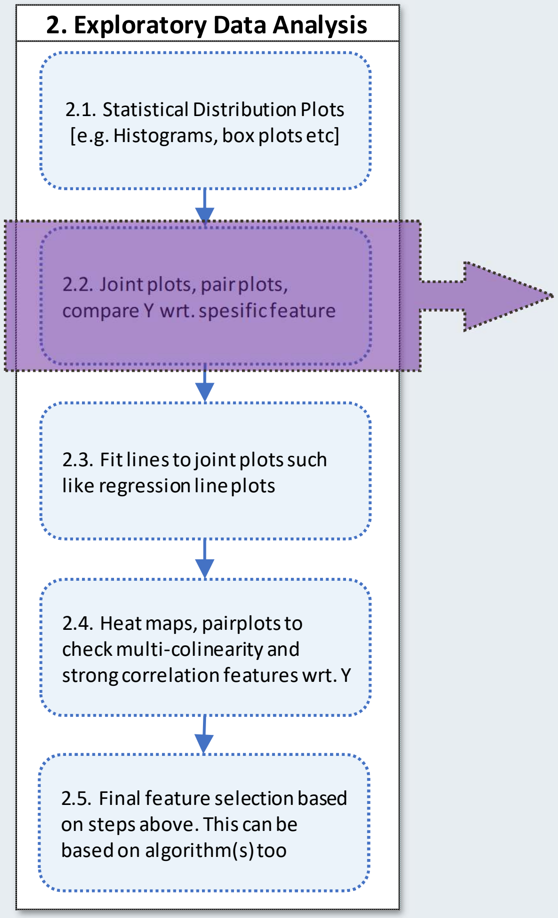

Figure 3: Exploratory Data Analysis Flow

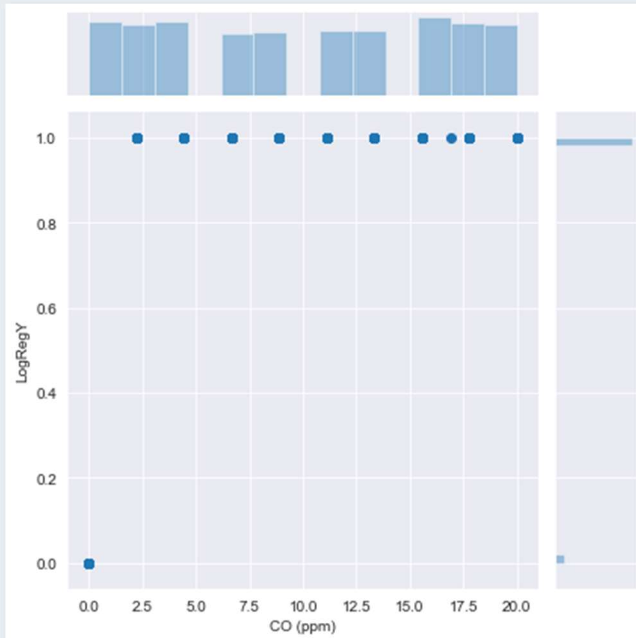

Chart 10: Pair plot CO(ppm) vs LogRegY

- This is to represent how to read the plot. LogRegY is derived directly from CO, so when LogRegY=1, it corresponds to scatter points of CO > 0 [This was how we defined it].
- Big chart box is scatters, small 2 chart boxes are histograms like previous slides

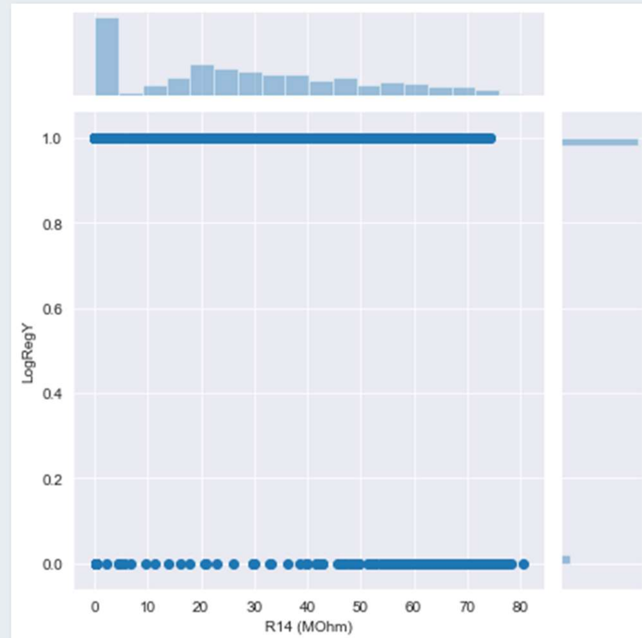

Chart 11: Pair plot R14 (MOhm) vs LogRegY

- As R14 increases, scatters of LogRegY=0 increases too, so there is definitely a correlation here [More on to this in coming slides].

Modelling Process

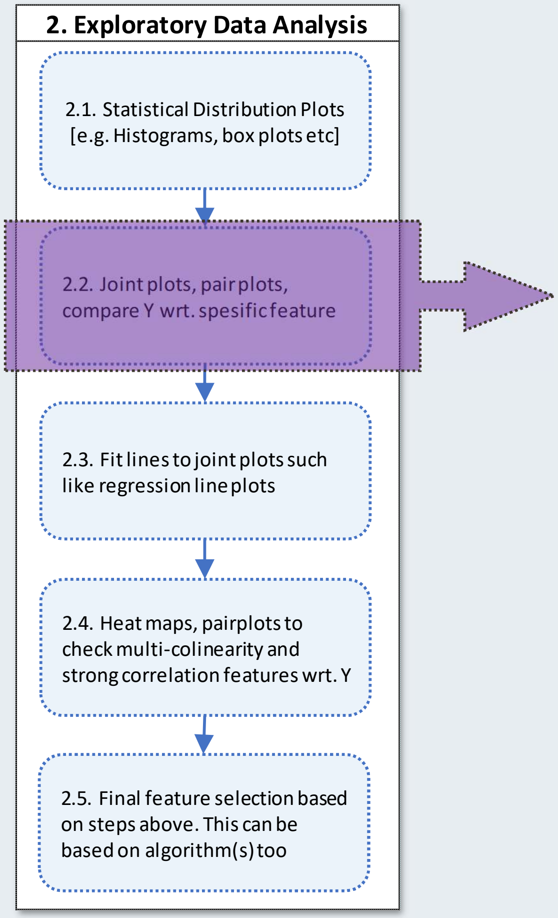

Figure 3: Exploratory Data Analysis Flow

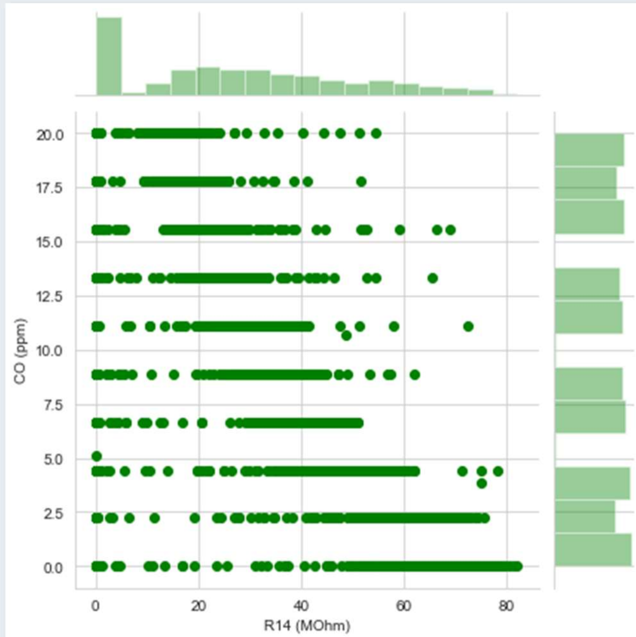

Chart 12: Pair plot R14(MOhm) vs CO (ppm)

- As R14 values increase, CO displays a noticeable increase in scatters, again an indication of correlation [more on this on next slides].

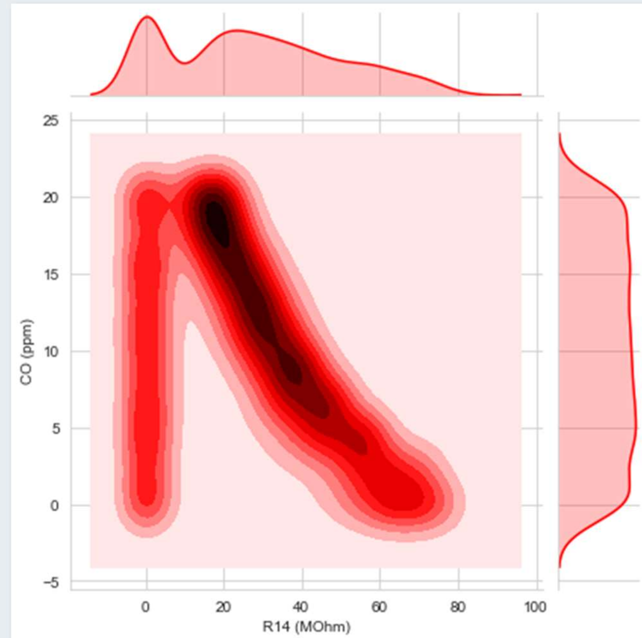

Chart 13: Pair plot R14 (MOhm) vs CO(ppm)

- Just another type of representation of the correlation where bolder red indicates more scatter points. Clearly seen the linear correlation here.

Modelling Process

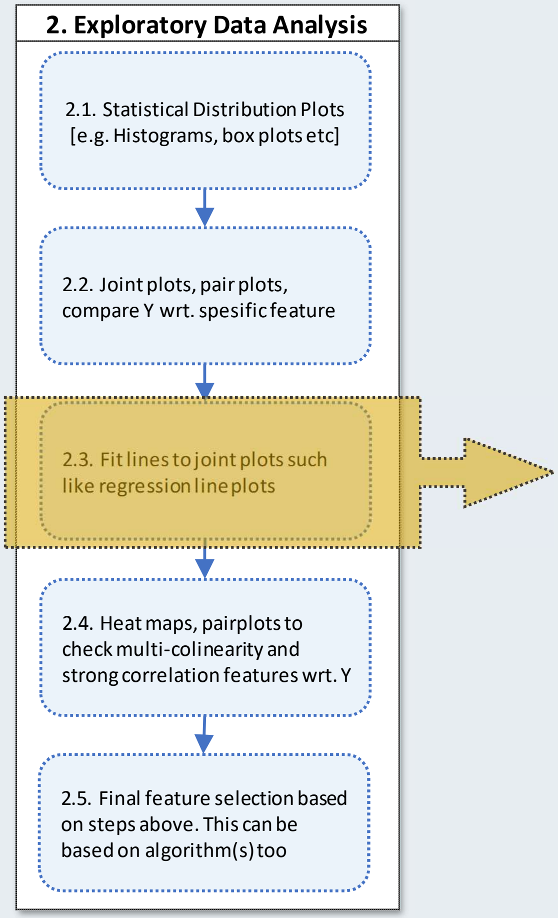

Figure 3: Exploratory Data Analysis Flow

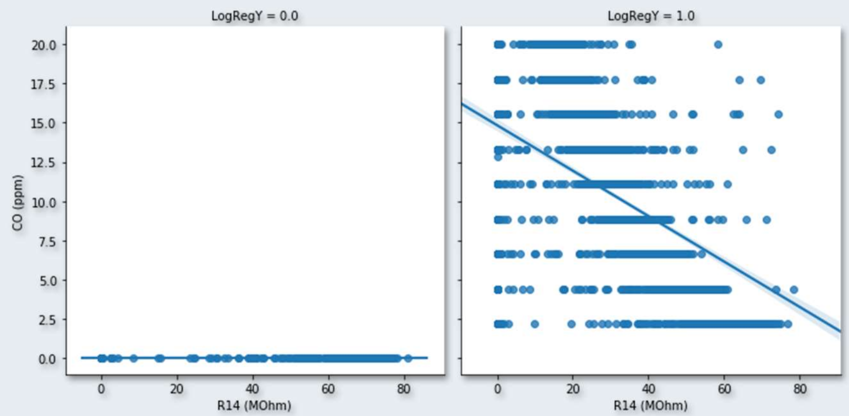

Chart 14: Regression fit plot, R14(MOhm) vs CO (ppm) vs LogRegY

- Left plot indicates the population where LegRegY=0, correlation between R14(MOhm) vs CO (ppm). Notice the increase in number of scatters, when the value of the R14 increases.
- Right plot indicates the population where LogRegY=1 (or CO(ppm)>0). There is clear regression fit line here. The steeper the slope, higher the correlation. The fit line indicates the best fit behaviour.

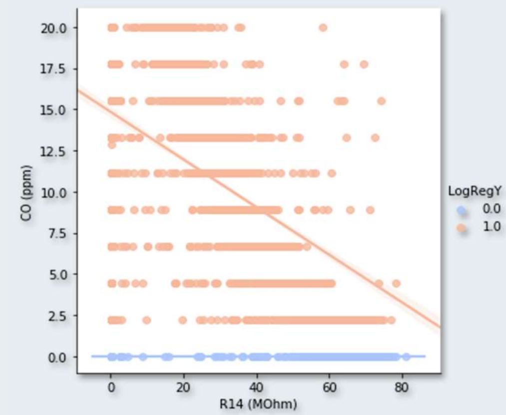

Chart 15: Hue Regression fit plot, R14 (MOhm) vs CO(ppm) vs LogRegY

- This is just explaining the same story as Chart 14 just by hue effect into one single chart, included here to create more visual understanding of this behaviour.

Modelling Process

2. Exploratory Data Analysis

2.1. Statistical Distribution Plots [e.g. Histograms, box plots etc]

2.2. Joint plots, pair plots, compare Y wrt. spesific feature

2.3. Fit lines to joint plots such like regression line plots

2.4. Heat maps, pairplots to check multi-colinearity and strong correlation features wrt. Y

2.5. Final feature selection based on steps above. This can be based on algorithm(s) too

Figure 3: Exploratory Data Analysis Flow

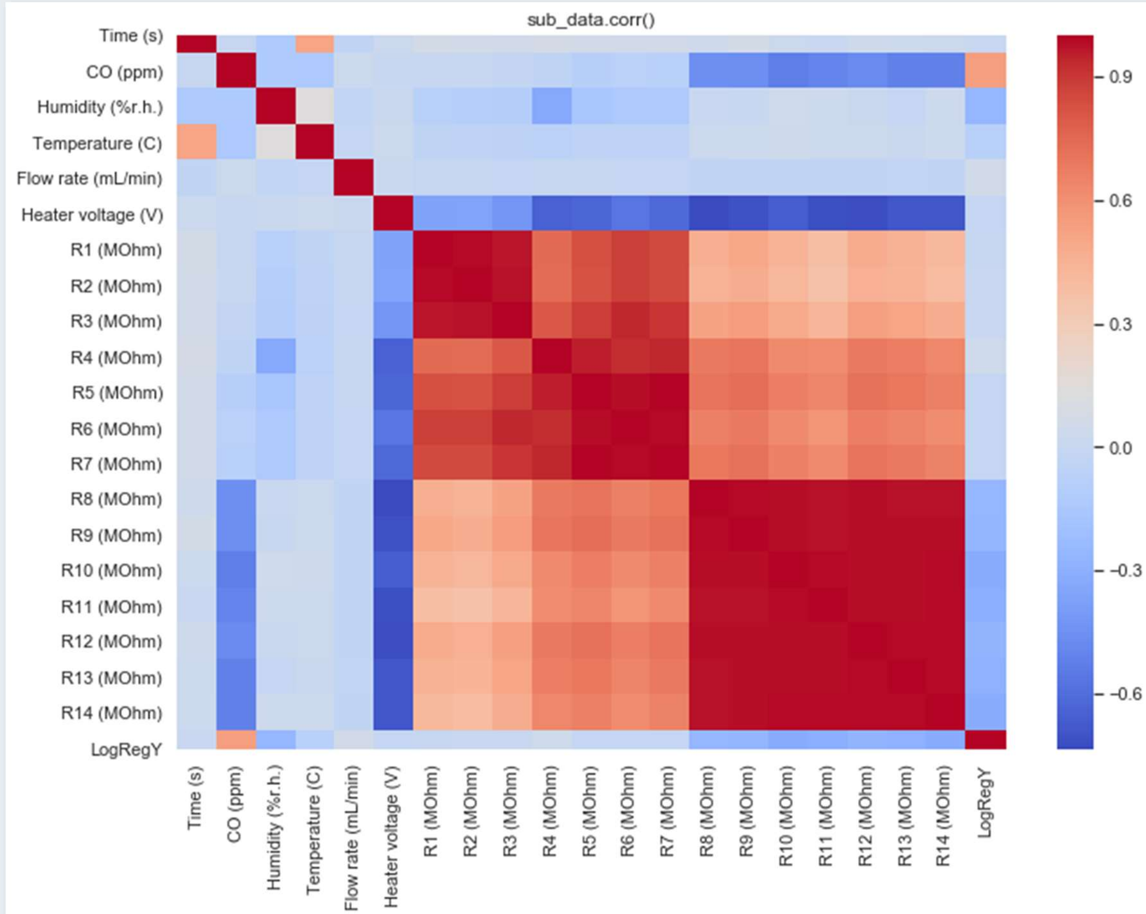

Chart 16: Heat map representing all features and their inter correlation

- In a nutshell, this plot represents the intercorrelation of all features between each other. From R8 to R14 there is a cluster of multi-collinearity [means they are behaving similar]
- R1 to R7 also have multi collinearity.
- Darker red or blue represents strong negative or positive correlation.
- We are looking at darker colours regardless of negative or positive direction, these will have information for the model.
- Diagonal line is always dark, as it is equal to 1, indicating correlation within the same feature.

Modelling Process

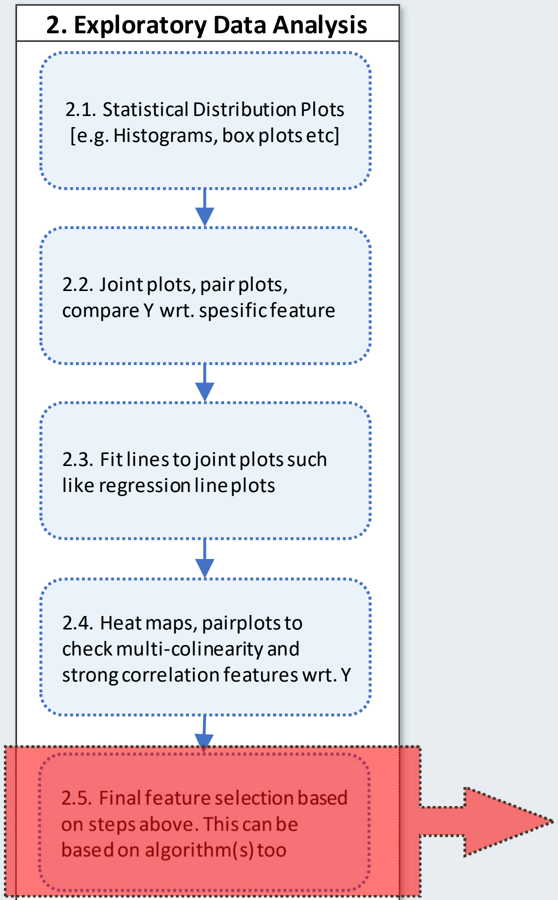

Figure 3: Exploratory Data Analysis Flow

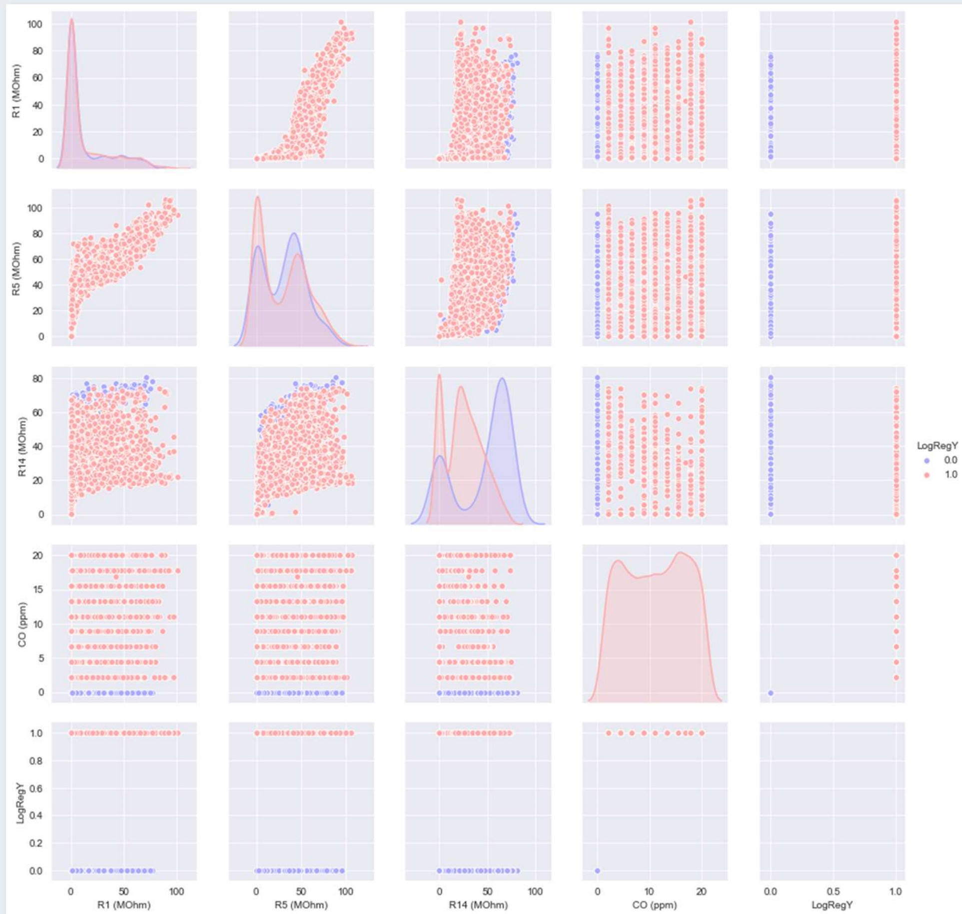

Chart 17: Hue Pair plot between final selected features

- This is a pair plot with hue based on LogRegY, where blue is scatters where LogRegY=0 and red is where LogRegY=1 [See legend].
- Diagonal plots represents the histogram lines where blue area/line represents LogRegY=0 distribution and red is for LogRegY=1.
- R1, R5 and R14 is selected to go into the model. This seemed like the most intuitive selection based on the visual EDA and steps up to now.
- From previous Chart 16, R1,R5 and R14 seemed to have a high correlation between rest of the features, so having the remaining features does not seem like adding value to the descriptive and predictive power of the model.
- Voltage is bi-model, temperature seems like forced (not dispersed enough), time is an input so should not be considered. Flow rate is also not considered as not showing enough correlation.
- Humidity has good natural distribution, has some correlation, but we will still not considered it as it is assumed to be a random input. Aim of this process is to understand and predict the behaviour of the CO based on R1-R14 sensors.

Modelling Process

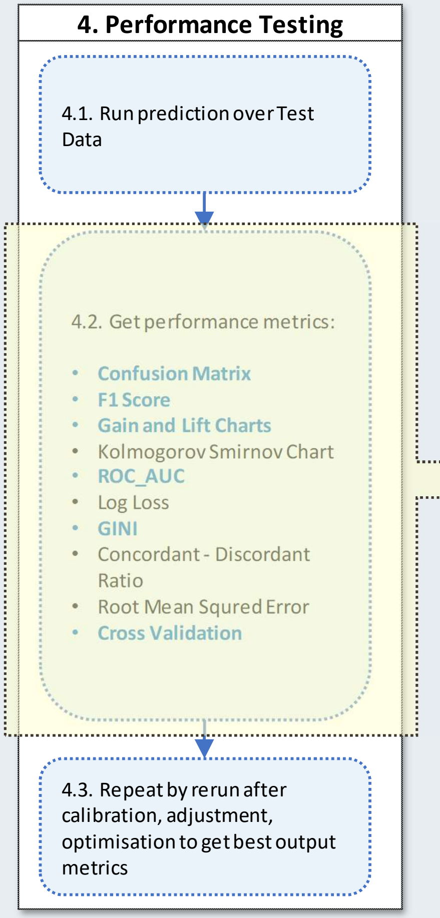

Figure 4: Performance Testing Flow

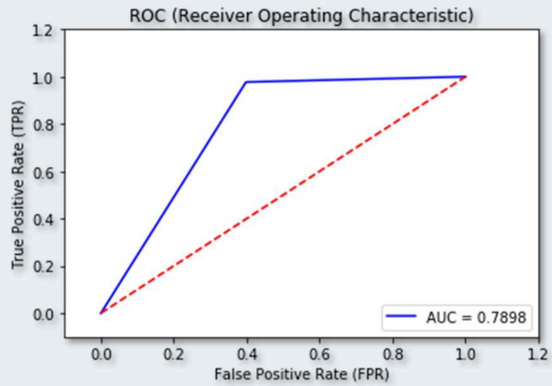

Chart 18: Log_1 Roc_AUC Plot

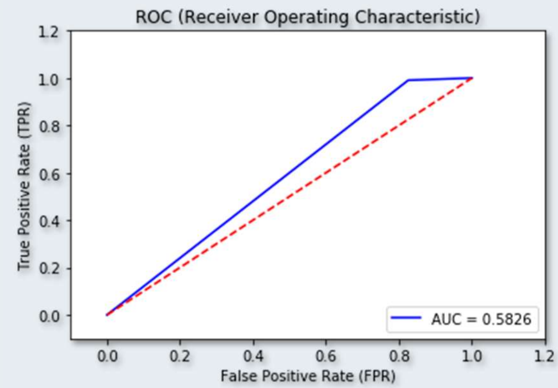

Chart 19: Log_1 Roc_AUC Plot

Model	TN	FN	TP	FP	Recall	Accuracy	GINI
Log_1	56	37	20	863	0.60	0.94	0.58
Log_2	22	104	8	842	0.17	0.89	0.17

- **Log_1** indicates the model with all features included. **Log_2** indicates model with only R1,R5 and R14 included. We can see clearly, **Log_2** is performing poorly here. GINI, Accuracy and Recall are all lower.
- TN is True Positive, FN False Negative, TP True Positive, FP False Positive [Refer **Appendix** for more info on this]. In short, we are trying to increase TN and TP and decrease FN and FP.
- Our LogRegY rate is around 90% [See previous slides showing histograms], so even a random guess would provide near 90% accuracy, so anything around 0.9 accuracy is a poorly performing model. Our **Log_2** is a very poor model in this case.

Modelling Process

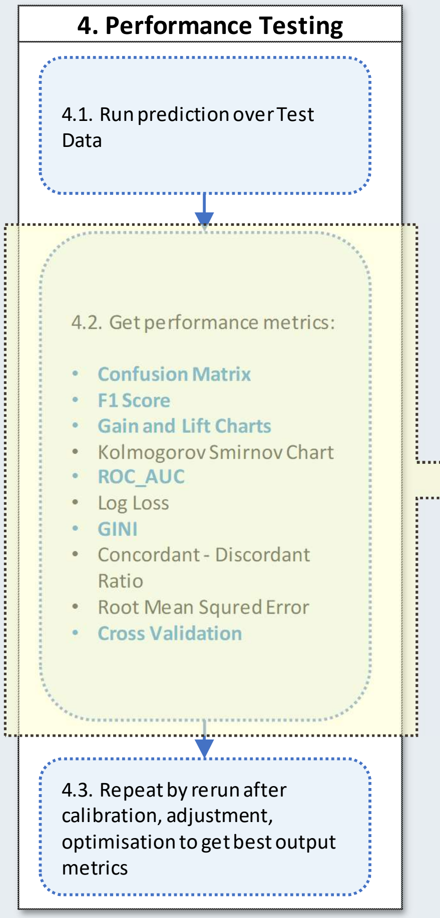

Figure 4: Performance Testing Flow

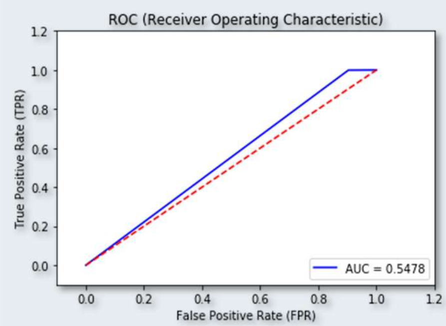

Chart 20: SVC_1 Roc_AUC Plot

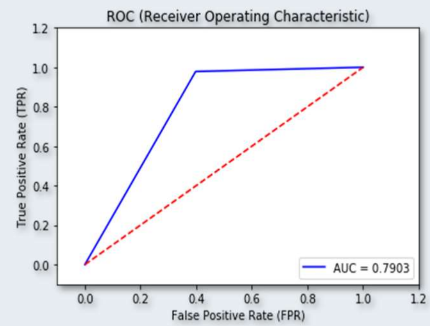

Chart 21: SVC_Grid_1 Roc_AUC Plot

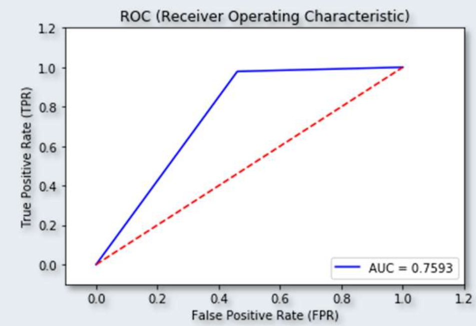

Chart 22: SVC_Grid_2 Roc_AUC Plot

Model	TN	FN	TP	FP	Recall	Accuracy	GINI
SVC_2	9	84	1	882	0.10	0.91	0.10
SVC_Grid_1	56	37	19	864	0.60	0.94	0.58
SVC_Grid_2	68	58	18	832	0.54	0.92	0.52

- **SVC_2** Support Vector Model with R1,R5 and R14 included with no GridSearch, **SVC_Grid_1** means all features included with GridSearch optimisation. We didn't tabulate the run results for **SVC_1** because it gave zero GINI.
- GINI, Accuracy and Recall are ok in **SVC_Grid_1** and **SVC_Grid_2**. And incredible result by adding GridSearch especially to **SVC_1** where GINI increase from nothing to 0.52. **SVC_Grid_2** is significant improvement as only with 3 features, we were able to describe almost all of the prediction power of **Log_1** with all features included [see previous slide].

Model Results

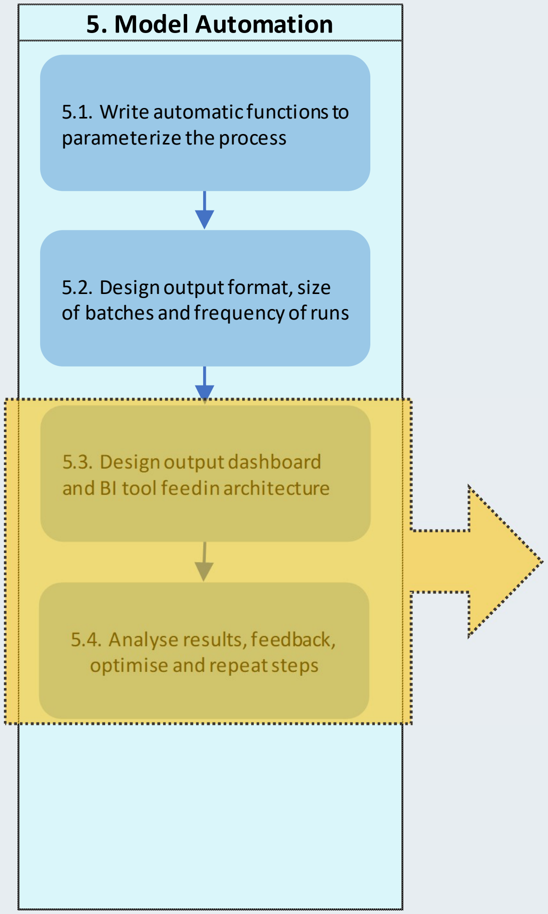

Figure 5: Model Automation Flow

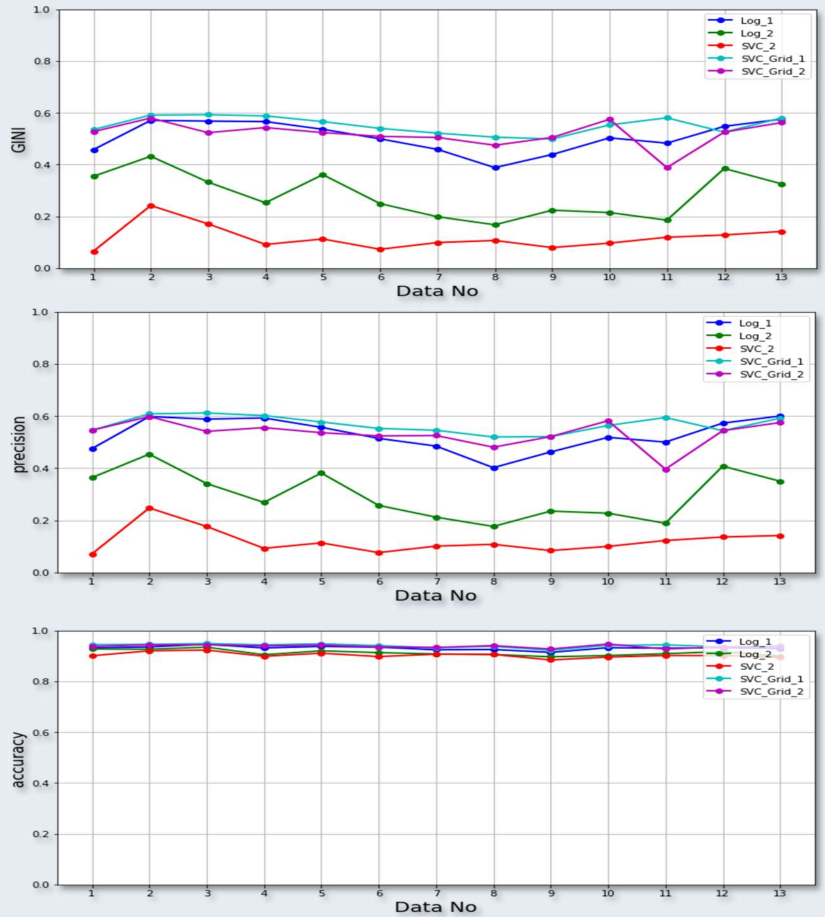

Chart 23: GINI, Recall, Accuracy - Combined Plot

- We have designed a dashboard as such, where can be compared 5 model selection performance based on time/date dependent data.
- We have utilized all of the available 13 data sets, so Data number is indicated in the horizontal axis [See **Appendix** for list of data corresponding date].
- We have come to a come to a conclusion that either full feature **Log_1** or **SVC_Grid_2** is the best option at this stage [This is for demonstration only, there are more to consider into this – will be explained in the meeting].
- The reason for our conclusion is although all models shows stability on results [which is a reflection of data consistency], **Log_1**, **SVC_Grid** models showed superiority over other options. **SVC_Grid_2** performed good considering with only 3 features which reduces complexity and eases implementation & monitoring.
- **Next steps** going forward could be possible to check other supervised algorithms e.g. [Though listed below tend to work better for nonlinear features, still worth checking]
- These could be XGBoost, Random Forrest.
- We can try un-supervised classification(s) to start with and model different classifications separately e.g. If we have two voltage input(s) - two waves or bi-modal input, this may suggest we may have some classification(s), where we can use algo's like:
- K means, PCA: Principle Component Analysis, to idenfity these classification(s), cluster(s).
- If we have resource time, we can check Deep Learning Algorithms like Neural networks.
- These could be, perceptron models with Tensorflow & Keras, Backpropagation and activation function optimisation and so on.

Process Architecture

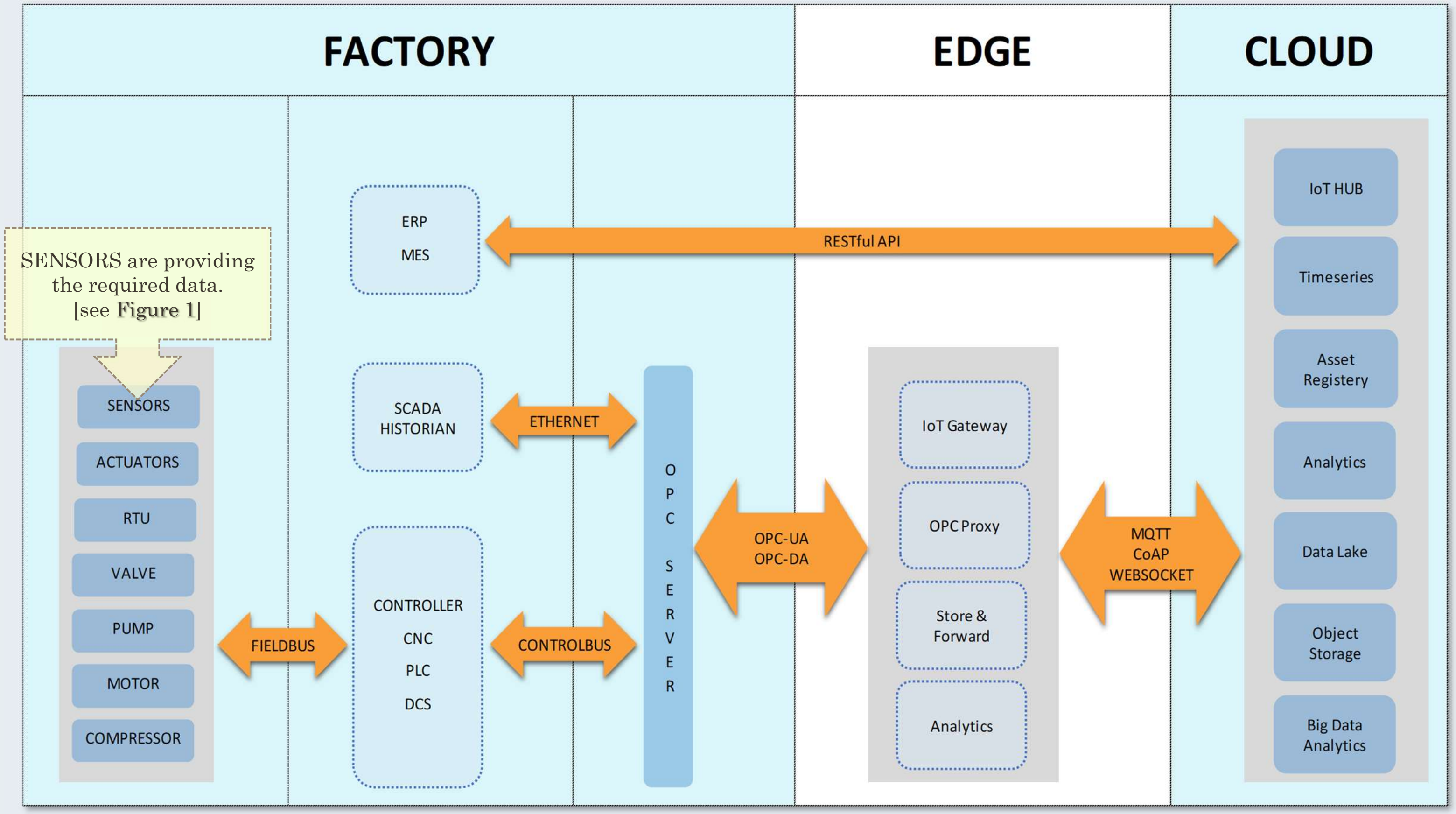

Figure 6: Process Architecture - Flow Diagrams

- Cloud and General System Considerations include:
- Frequency of data refresh, lag/real time transfer, speed, security, ML analytics type e.g. cloud vs server.
- This is high level Industrial IoT Process Architecture. For our solution, we receive data from **SENSORS** and follow the process.
- ML Model is part of Cloud Analytics.

- **ERP**: Enterprise Requirements Planning
- **MES**: Manufacturing Execution System
- **CNC:** Computer Numerical Control
- **PLC**: Programmable Logic Controller
- **DCS**: Distributed Control System
- **REST**: Representations State Transfer
- **RESTful**: Web services that conform to the REST architectural style
- **OPC**: Open Platform Communication
- **OPC-UA**: OPC Unified Architecture
- **OPC-DA**: OPC Data Access
- **MQTT**: Message Queuing Telemetry Transport
- **CoAP**: Constrained Application Protocol
- **IoT**: Internet of Things

Process Architecture

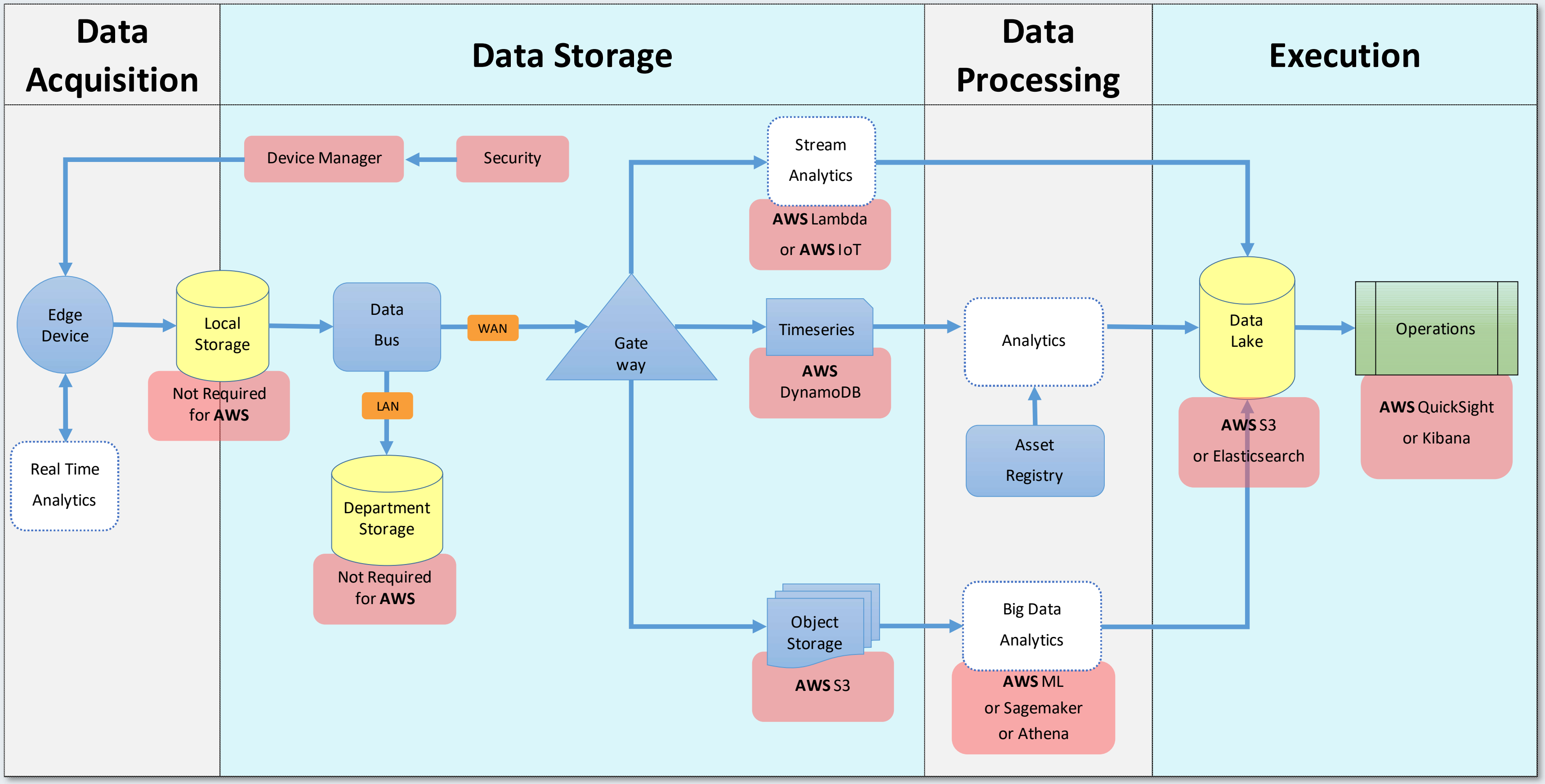

Figure 7: Process Architecture – End to End Data Flow with/without AWS