Indian Solar Panel Modelling

Paul Metcalf, Principal Engineer

Cloudforest Technologies Ltd

V. 1.0

Doc. Number WJ-20-0001

Contents

Introduction: Why Model Solar Panels?

Identifying underperforming Solar Panels is a key capability for maximising plant performance & ROI

- Solar panel arrays have a high initial capital cost, repaid by generating predictable quantities of electricity from the sun
  - Investment cases are predicated on being able to generate a certain amount of power over a given time period
  - ROI is maximised by panel up-time - Faulty panels reduce ROI by failing to generate saleable power
- Solar panels should follow predictable patterns of behaviour, once weather, solar load and degradation rates are known
- Expected solar load can be generated from time of day, season, and weather (sunny, cloudy, rain etc.)
- Can panels in need of cleaning or maintenance be identified from plant data? Can faulty panels be identified?
- Can future output be forecast, based on historical data?

The Objectives of this Modelling Work

Project Objectives

- To develop methods for **identifying failing panels / panels in need of maintenance**
- To develop methods for **identifying when panels collectively need cleaning**
- Develop **Output Power** predictions based on time, weather and historic performance

Scope of the Project

- A solar power station in India with two arrays is analysed
- Local plant weather is captured
- Panel performance at the inverter level is captured
- This data is used to identify panels that are underperforming
- This data is also used to predict current generation output from conditions
- Service and maintenance records are not available, so assumptions will have to be made in this area

Exploring & Working with the Dataset

Generation & Weather data for two Solar Plants are captured in separate files, linked by timestamps

Generation Data

Parameter	Notes
Datetime	• Date and time stamp initially stored as string in 15 minute increments
Plant ID	• String labelling source plant
Source Key	• String labelling the inverter
DC Power	• Rate of DC Power generation in preceding period (pre-inverter) in kW (assumed average)
AC Power	• Rate of AC Power generation in preceding period (post-inverter) in kW (assumed average)
Daily Yield	• Cumulative sum of generated power for that day in kWh
Total Yield	• Cumulative sum of generated power since last reset of records in kWh

Weather/Condition Data

Parameter	Notes
Datetime	• Date and time stamp initially stored as string in 15 minute increments
Plant ID	• String labelling source plant
Source Key	• String labelling the sensor (one per plant)
Ambient Temperature	• Ambient temperature at the plant in °C
Module Temperature	• Temperature of one panel directly attached to sensor in °C
Irradiation	• Average solar irradiation in preceding 15 minutes, in kW/m2

Issues with missing/incorrect values have required thoughtful management to create a modelling dataset

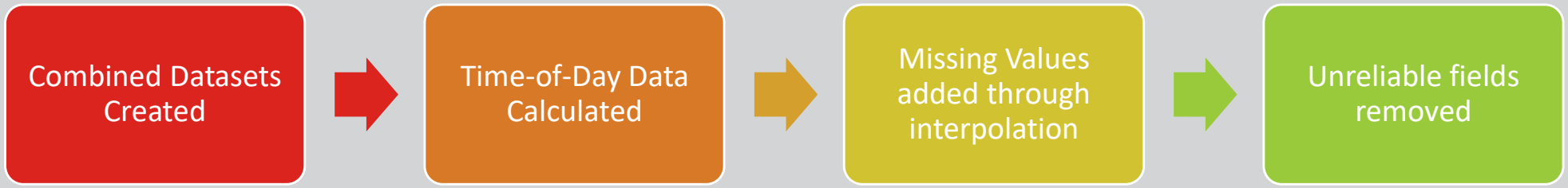
Issues Encountered

- Generation & weather data for each asset are in independent files [Segregation]
- Not all cells have data for all time stamps [Missing Values]
- Some data is clearly incorrect [Erroneous Values]
- Some data fields (i.e. Total Yield) appear unreliable [Values not as expected]

Solutions Implemented

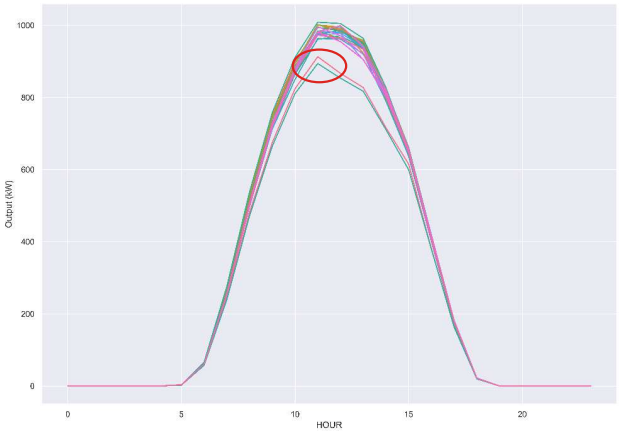
- Data has been combined into plant-specific (1 & 2) datasets
- Missing values are initially handled as NaN
- Time-of-day data added
- Generation during night-time hours set to zero

Data Munging Pipeline



Engineered Features deriving hourly and daily metrics provide insights into panel performance

Plant 1 Mean Daily Yield

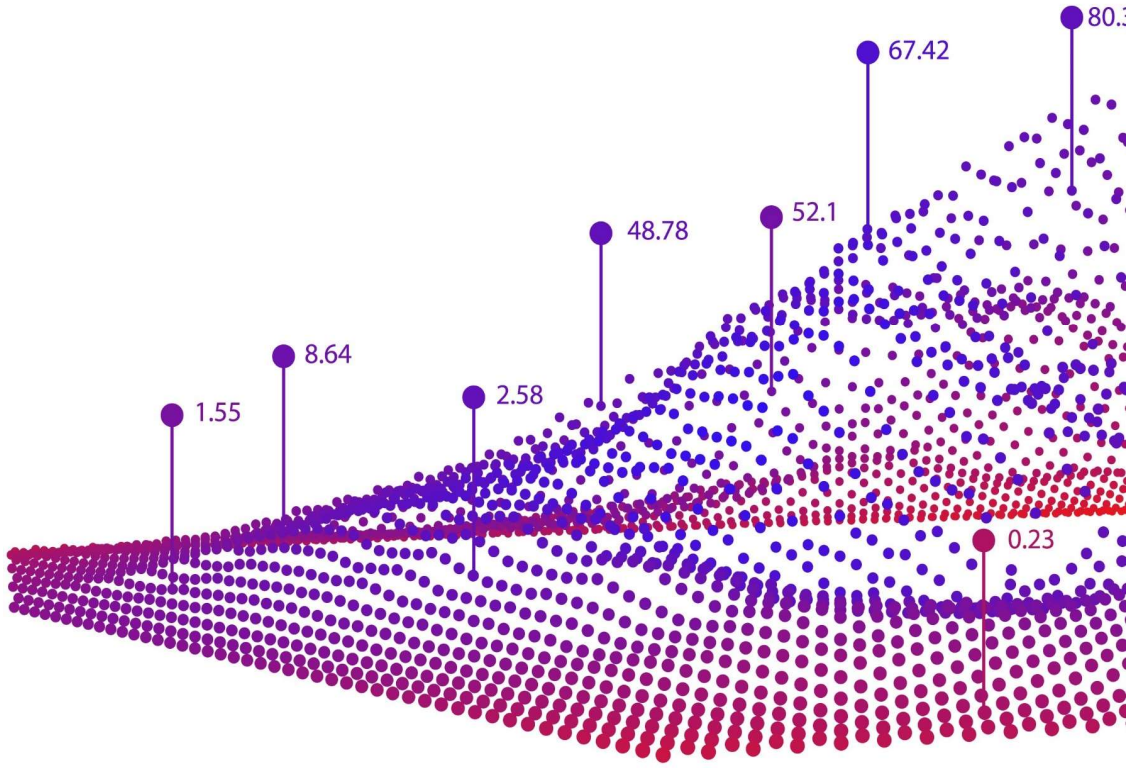


Plant 1 Mean Hourly DC

Plant 1 Mean DC-Irradiation Ratio

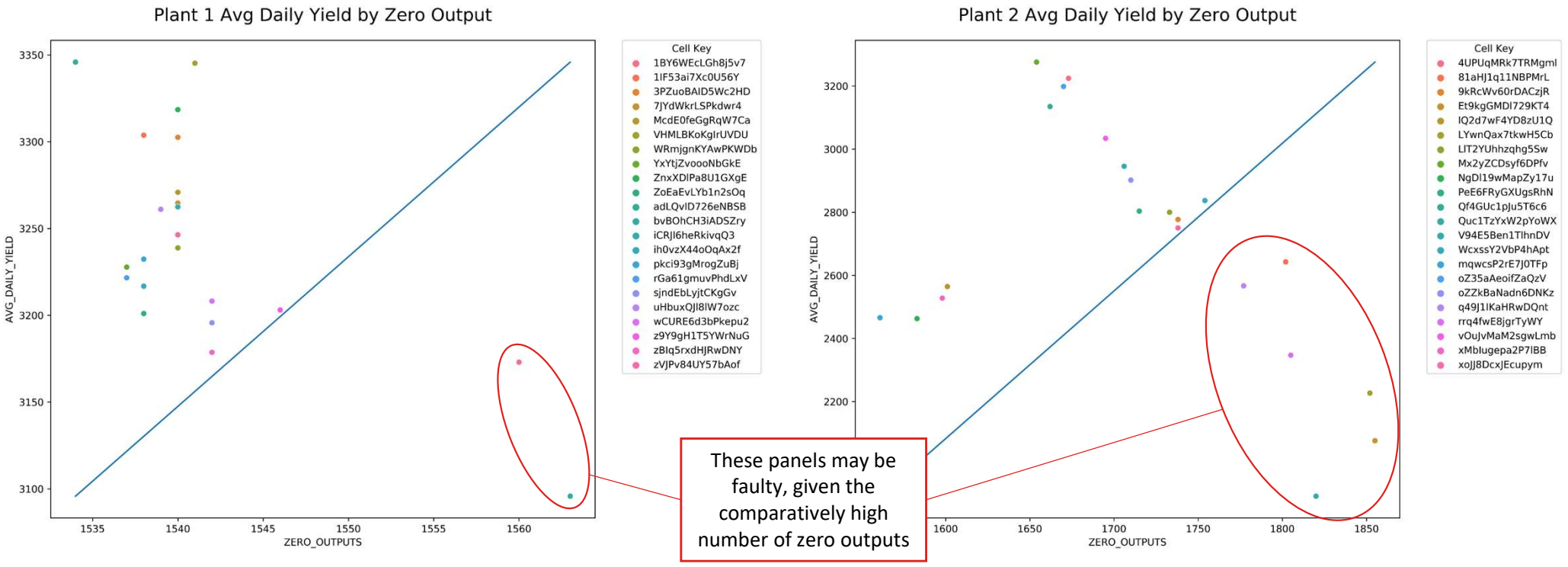
- Features Added:
  - Hourly & Daily DC Output, AC Output, and Yield Means
  - The ratio of DC Output to Irradiation
- These features help visualise where performance and output may be less than expected; An early indicator of panels in need of cleaning or maintenance
- They also allow the data to be clustered into higher performing/lower performing sets with machine learning
- AC output features were eventually removed following exploration, as these were found to be duplicates of DC outputs

Identifying Failing Panels from Data



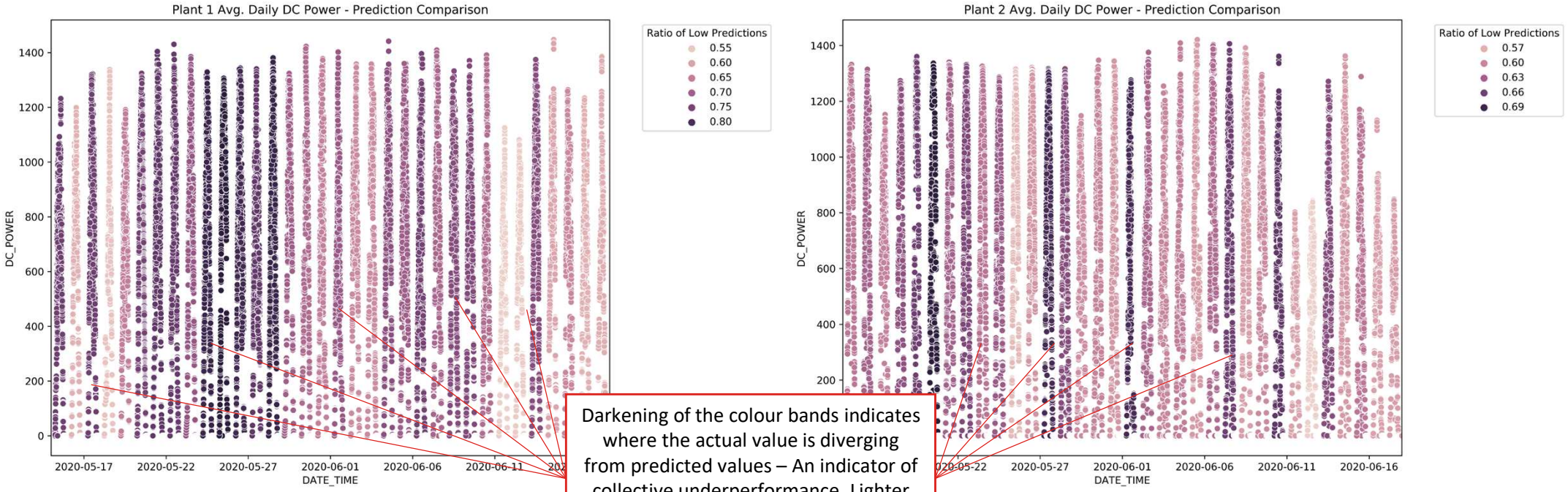
How can we identify failing panels from the data? Lower performance for given conditions may provide clues

- A good inverter should be consistently productive, so features related to consistency and productivity could create insights into failing panels:
  - One option could include average daily yield divided by standard deviation in daily yield
  - Another could be a count of the number of zero output instances
- An assumption can be made that panels will have zero output if they are faulty, rather than a lower output
  - Therefore, panels with large numbers of zero outputs compared to peers can be investigated for faults:

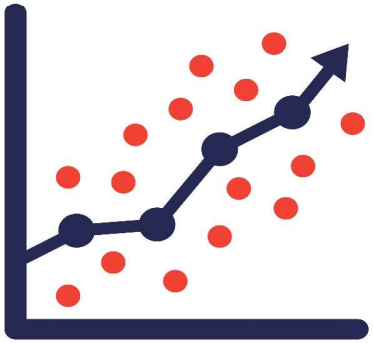


Panel expected output can also be used to identify when panels need cleaning

- Areas where many panels are producing less power output than predicted, for a given level of irradiation (i.e. at the same time), indicate that the panels, as a collective, may need cleaning
- A simple linear regression model can be trained to forecast output for a given set of conditions, and the number of below-prediction real values can be measured
- This effect seems to be born out in both plant datasets, where mean power output drops, and then recovers, possibly as a result of (undocumented) cleaning:



Developing Predictive Models for Realtime Generation

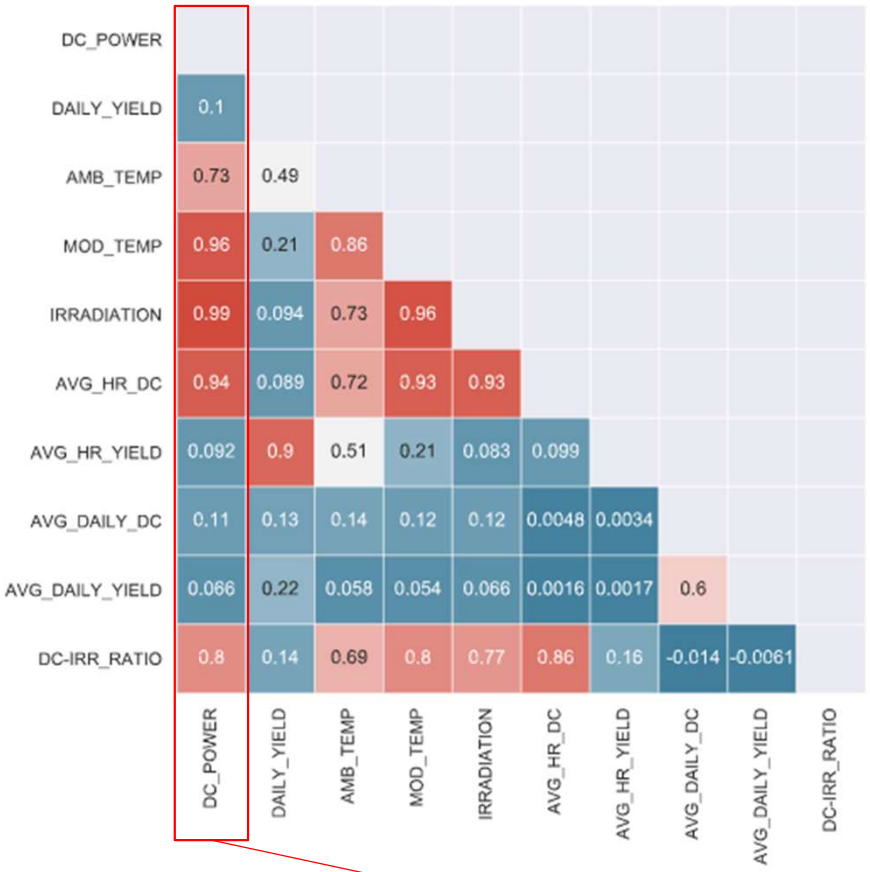


Developing Predictive Models to help support plant operations

- Predictive models seek to estimate the solar panel output (in the form of DC output) for a given plant at prescribed conditions
- These predictions can be used to support plant operation and energy grid integration, and can be used to identify when panel performance is not as expected (panels may need cleaning or maintenance)
- Baseline predictive models have been built based on linear and polynomial regression of DC output to plant-specific operating conditions
- Equivalent models using Deep Neural Networks (DNNs) have also been trained
- Although desirable, it isn't possible to develop forecast models for future plant output without some form of weather forecast data

DC Power serves as the target variable for predictive modelling, with decent linear correlation to other variables

Plant 1 Correlation Matrix



Plant 2 Correlation Matrix

Target Variable

Model performance is comparable between baseline and DNN models

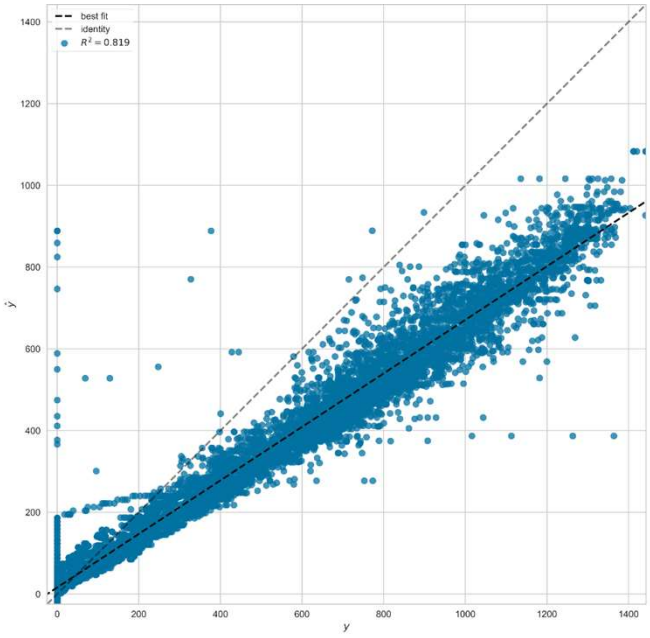
- Linear Regression, Polynomial Regression and Multi-Layer Perceptron (MLP) models (Neural Networks) all perform similarly
- DNN performance was improved by input data normalisation; LR/PR not so
- DNN models underwent a parameter optimisation process to refine performance
- All models run into a natural limit created by inclusion of underperforming panels in the dataset
- Models have quite different coefficients between plants 1 and 2, with considerably more (unexplained) variance in the base dataset for plant 2
- Forward-forecasting isn't really possible, due to the absence of near-future weather data, which is a key input to any forecast model

Model	RMSE	MAE	R^2
Linear Regression Plant 1	57.7	26.01	0.979
Polynomial Regression Plant 1	54.4	23.69	0.981
DNN-MLP Plant 1 (Optimised)	55.07	22.33	0.981
Linear Regression Plant 2	232.2	131.33	0.609
Polynomial Regression Plant 2	217.43	101.63	0.658
DNN-MLP Plant 2 (Optimised)	217.56	97.39	0.657

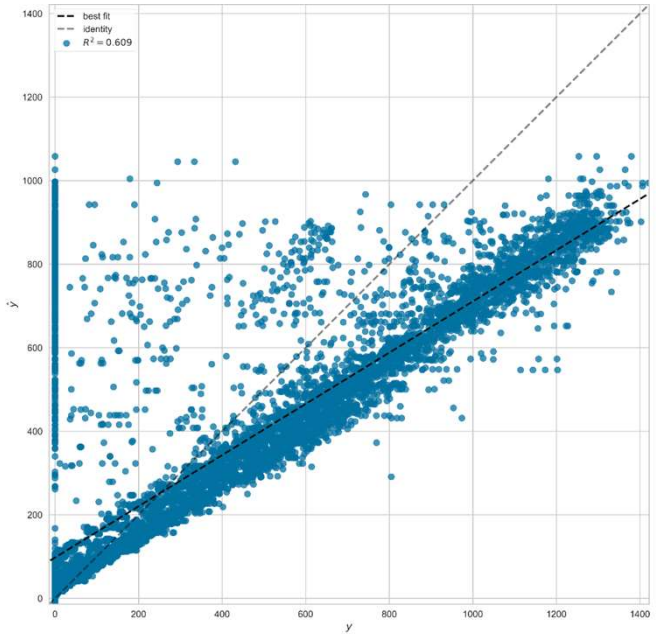
Linear Regression Models – Trends are modelled; Underprediction present at higher outputs

- Prediction and Residual Error plots for both plants show biases created during training, with underpredictions in output at higher absolute power levels.
  - This is probably due to the presence of faulty panels in the training datasets themselves
- Plant 2 contains considerably more variance in the training data, which generates a wider spread of errors in the Prediction and Residual Error plots.
  - The same broad trend of underprediction at higher power output levels can also be observed in plant 2 – Again likely due to the presence of already failing panels in the training dataset.

Plant 1 LR Model Prediction Error



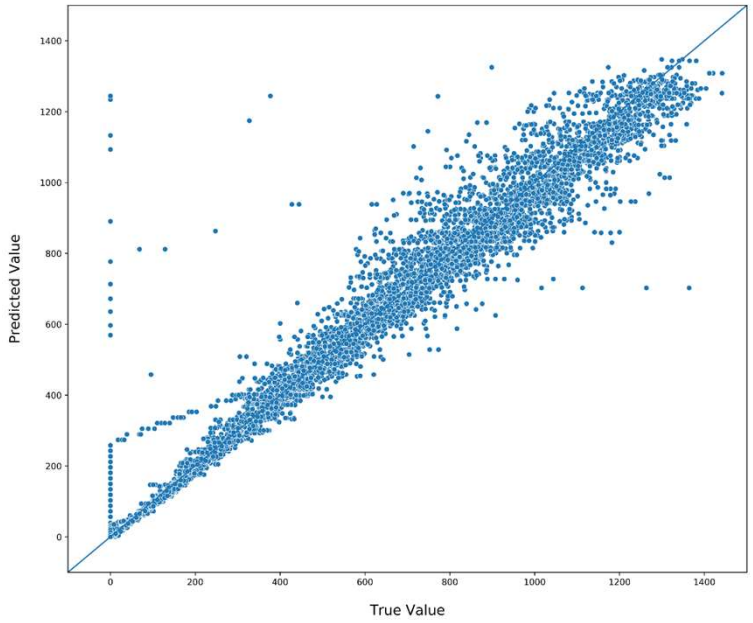
Plant 2 LR Model Prediction Error



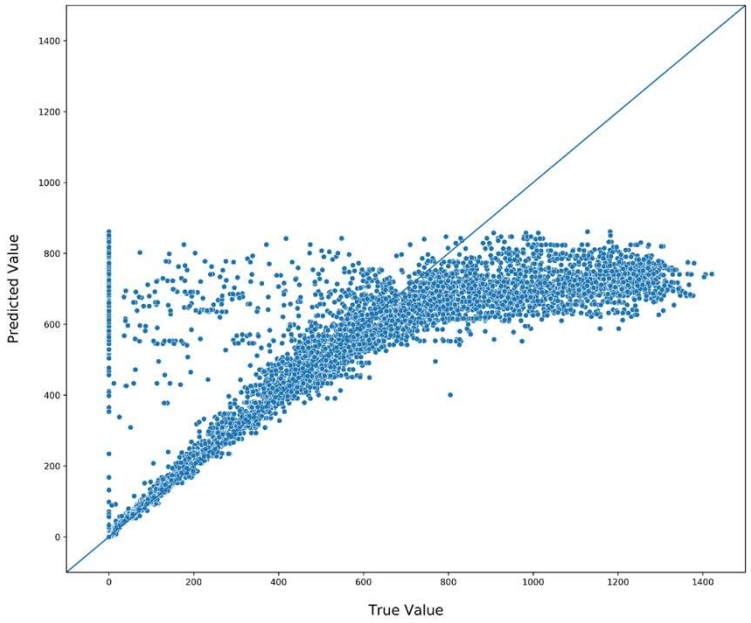
Neural Network Regression Models follow similar trends, with some differences

- Multi-Layer Perceptron (MLP) Neural Network Regression Models show a couple of interesting trends:
  - Firstly, the predictor for plant 1 doesn't appear to exhibit an underprediction at higher power outputs, although RMSE/ MAE measures for the model are comparable to models that do
  - Secondly, the model for plant 2 does underpredict at higher power levels – The reasons for this are not immediately apparent, but do suggest underprediction may be linked to 'unexplained' variance in the datasets

Plant 1 MLP NN Model Prediction Error



Plant 2 MLP NN Model Prediction Error

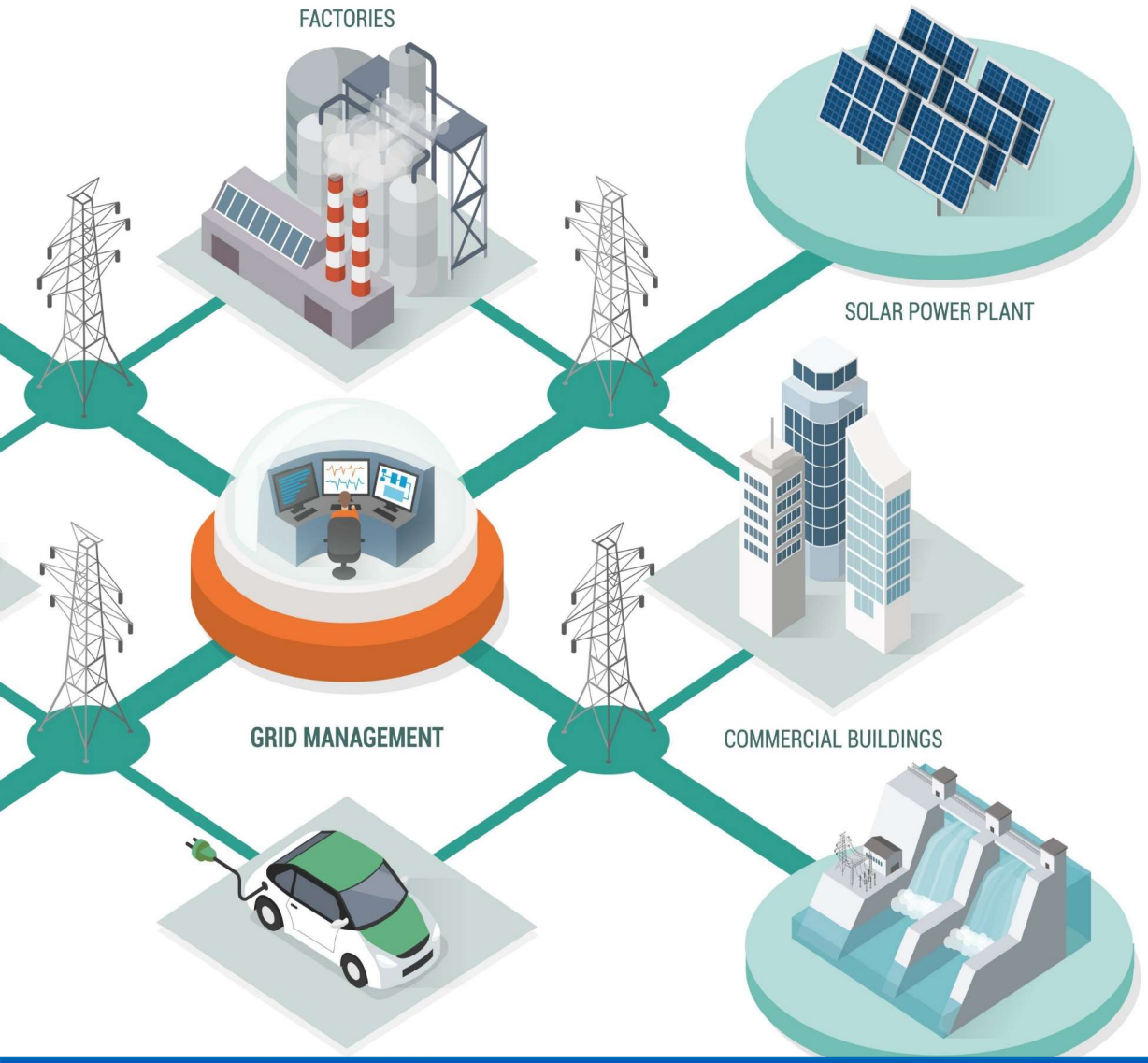


What about forecasting output? Lack of near-future weather data makes this difficult …

- It would be a useful capability to forward-forecast plant output into the future (days and weeks), as this can help with integrating the plant effectively into the wider energy grid
- However, modelling of current output shows that predicted power output is very dependent on environmental conditions – Particularly irradiation; Less so temperature
- In order to develop an accurate forecasting model, some form of weather forecast to the prediction horizon is needed as an input to the model
- This isn't readily available in the dataset, but could be made available from a weather API
- Another option would be to use LSTM cells in the NN models, as this would capture near-history weather patterns that may extrapolate into the near-future

Summarising the Project

It has been demonstrated that it is possible to identify underperforming panels through analysis & ML



- Although the dataset isn't perfect, data munging and feature engineering can be applied to generate useful data
- Identifying panels with lower outputs than their peers is a good way of identifying the need for a maintenance intervention
- Visualising when panels collectively underperform, by comparing actual with predicted output, can help identify when cleaning is required
- Forward forecasting isn't possible because of the absence of confounding weather forecast data, but ML models can be trained to make predictions on actual output for given conditions
- Developing a proper forecasting capability would be useful for wider integration into a smart grid

Recommendations for Future Investigations & Work

Weather Data

- Collecting weather data (humidity, cloud cover, windspeed, air pollution) to go alongside the existing weather and performance datasets would be invaluable
- Historic weather data when coupled with forward weather forecasting would enable forecast modelling for future generation, to the horizon of the weather forecast accuracy

Maintenance Records

- Maintenance data should be collected alongside performance data, to generate event labels that are impactful on panel performance (such as cleaning)

Digital Twins

- Predictive output models can be used to feedback to plant operation, advising when cleaning or maintenance interventions might be required in the future