In-house ML Model for Inverter Fault Prediction & Forecasting in Solar Plants

Project Review Meeting

Ishan Jain

March 17, 2021

Project statement

To develop ML model for predicting downtime (alarms where inverters are getting shut-down, i.e. active power = 0, when irradiance >50) in each inverter of the plants for the next day.

Input Data		Output
Historic Inverter data	→ **ML Model** →	Inverter-wise prediction of trips for the next day (hourly/daily)
Historic Weather data		Probabilities of inverters’ being faulty (hourly/daily)
Historic Alarm code data		Desired performance metrics

Project is divided among 4-stages

1. Data stage	2. Modeling stage	3. Validation stage	4. Iterative stage
Data Access for training (GMR, Mansa-1, Porbander)	Select one base-model (LSTM, GRU, XGBoost)	Inverter-wise prediction of trips (hourly/daily)	Increase the training data size
Pre-processing	Train the model	Probabilities of inverters' being faulty (hourly/daily)	Improvements in the data structure & alarms' selection logic
Trip's selection logic	Validate and test the model	Performance metrics	Modifications in model
1 months	2 months	1.5 months	1.5 months

Progress (Oct 2020 - March 2021)

Step 1: Fault Classification (Oct 2020)

Problem: Can we classify alarm-data into ‘trip’, ‘warning’, and ‘no-alarm’?

Classification of data into three categories (‘trip’, ‘warning’, no-alarm) for a given input data.

Outcome: Excellent accuracy with logic and decision trees.

Status: **Completed**

Step 2: Fault Detection (Nov 2020)

Problem: Can we detect ‘trip’ vs ‘no-trip’ post-classification in the given data?

A predictive modeling problem where a class label (Trip vs no-Trip) is predicted for a given input data. Given the data, **classify** if it is Trip or no-Trip.

Outcome: Excellent accuracy of fault detection using Random Forest, LightGBM, XGBoost.

Status: **Completed**

Step 3: Fault Forecasting (Nov 2020 - March 2021)

Problem: Can we forecast the next-day’s ‘trip’ vs ‘non-trip’ for the inverters?

A sequential, time-series based predictive modeling problem where we need to predict/forecast - Trip vs no-Trip for the next day.

Outcome: detailed in the report.

Status: Concluded

Two problems in the fault forecasting

Poor leading indicators for the prediction of the trips

What leads to trips? Are leading indicators present?

Our data records what happens pre- and post- the alarm/faults/trips very handsomely.

However, it seems that leading indicators based on inverter and weather data (active-power, irradiance etc.) are not predicting trips as we expected them to predict.

Class imbalance issue

where the class distribution is not uniform among the classes (‘trip’ vs. ‘no-trip’)

Many classification **learning** algorithms have low predictive accuracy for the infrequent **class**.

https://docs.google.com/document/d/1GGWRvBCqKW1rvqLWpXN4SXvIQWsjuVil/edit

Results

Is the model successful in predicting which inverter will be down tomorrow?

Yes

Is the model successful in predicting which inverter will be down tomorrow?

Date	TP	FN	FP	TN	Accuracy	Precision	Recall	F-1 Score	Actual_Trips	Predicted_Trips
Mar-09	30	5	3	7	82%	91%	86%	88%	35	33
Mar-10	29	8	5	3	71%	85%	78%	82%	37	34
Mar-11	33	10	0	2	78%	100%	77%	87%	43	33
Mar-12	36	6	1	2	84%	97%	86%	91%	42	37
Mar-13	33	7	2	3	80%	94%	83%	88%	40	35
Mar-14	29	12	1	3	71%	97%	71%	82%	41	30
Mar-15	26	17	0	2	62%	100%	60%	75%	43	26
March -16	34	9	0	2	80%	100%	79%	88%	43	34

Are there any gaps?

#	Defined questions	Results (have we solved the question?)
1	Are the models producing output in the right format?	Yes. They are producing next day's predictions in the desired format.
2	Are the models site (location) independent?	Yes. As long as the data structure and range of the features are the same. Data engineering might be additional for each site.
3	Are the models inverter-OEM's agnostic?	Yes. As long as the data structure is same. Data engineering might be additional.
4	Are the models error-codes independent?	Yes.
5	Is the current model plug-and-play?	Yes. However, it can be made plug-and-play post-data engineering at each site.
6	Is the model producing accuracy, precision, recall, and F-1 score above 75?% during testing and validation?	Partially, for some days only. Accuracy is mostly above 90% however, F-1 score takes toll.
7	Is the model giving reproducible results of performance metrics?	Yes.

https://docs.google.com/document/d/1GGWRvBCqKW1rvqLWpXN4SXvIQWsjuViI/edit

Insights and Learning

Insights & Learning

#	Topic	Time Utilized	Insights and Learning
1	Data Structure	~20 days	In the last 4 months, we have worked on 3 different types of data-structure where data shape, data volume, and initial data engineering were redone.
2	Data Predictability & Resampling	~1 to 2 months	Since raw data is not a strong predictor of the 'Trips', additional new features - time-based, categorical alarm names, difference of the previous features (delta features), and lagged target variables using lookback are used.
3	Inverter Behaviour	~10 days	Out of 45 inverters at the plant only 35 inverters had 'Trip' as per the selected logic in the year 2020.
4	'Trip' Selection Logic	~1 months	However logic based-on 'active power' = 0 is finally used and selected.
5	Class Imbalance Problem	~2 months	SMOTE (one of the oversampling techniques) performed relatively well.
6	Time Duration & Data Frequency	~1 month	Data frequency is given to be 5-min, however 1-hour to 1-day data frequency is used and finally 1-day frequency has been finalized.
7	Feature Engineering	~2 months	Total New features = 100+
8	Alarms Distribution and Analysis	~15 days	Analysis files: https://drive.google.com/file/d/1f04a7JpVAh6c95fqD08y7mwF-i7cyb6H/view?usp=sharing

https://docs.google.com/document/d/1GGWRvBCqKW1rvqLWpXN4SXvIQWsjuViI/edit

Time-travel to 2020

What would I do differently if I could go back 6-month in time?

#	Category	What will I do differently?	Time Spent
1	**Inverter OEMs database**	In Sept 2020, we started building Inverter's OEMs database to be used in our models to make it OEM agnostic, however, that turned out to be fruitless exercise.	~15 days
2	**Data Structure and API access**	From Sept to Dec 2020, we have used manual download of data from InfluxDB, however use of the current API is relatively faster and efficient. We should have facilitated API access sooner.	~ 1 to 2 months
3	**Model Selection**	LSTM vs Classification In the past (esp during the work of the freelancer), we have kept the model to LSTM. In our current approaches, we have varied to other classification techniques for forecasting.	~ 1 month
4	**Data Resampling & Predictability**	Selection of 1-hour and 24-hour frequency for model	~ 1 month
5	**Class Imbalance Problem**	Use of random undersampling techniques (because with other techniques time-to-run was too high). Now we have fixed this issue all together.	~ 2 months

Marketing Plan

Client Journey

Who can use this model as plug-and-play?

- Clients with good-quality historic (at least 1 year) data of inverter, weather, alarm for training
- Clients who are concerned about the downtime along with the trips in the inverters
- Same data structure and data fields are preferred.
- Location of plant, Inverter OEMs, number of inverters are not an issue under the current model.

How would client-journey look like?

- Data acquisition and data structure Check
- Assessment of data engineering required
- Selection of performance parameters (GreenKo vs. GMR)
    - Accuracy
    - Precision
    - Recall
    - F-1 Score
- FInal selection and deployment of the model
- Time to deploy the model:
    - with no additional data-engineering: <3 days
    - with additional data engineering: ~ 2 weeks

Recommended Commercialization Steps

In order to commercialize the current model, we have to consider following one of the following goals:

- ❑ **Better accuracy of work-order generated** - equivalent to Precision

  OR

- ❑ **Better accuracy of 'trips' forecasted** - equivalent to Recall

  OR

- ❑ Optimum accuracy of work-order generated (Precision) and accuracy of 'trips' forecasted (Recall) - equivalent to F-1 score.

#	Steps	Description
1	Daily internal validation and near-term improvements	Daily forecast and data analysis post forecasting on GMR data to build confidence internally. Immediate changes: 1. 45 days of validation on GMR 2. When will the downtime happen? 3. How long will the downtime last?
2	Pilot @ GMR client	Trial at the client site.
3.	Model fitting for the other plants	Internal model-fitting for the other locations and OEMs
4	Pilot @ other clients	Trial at the client site.

Plan for GreenKo vs. GMR

GreenKo - 800 MW, SMA

Product Stack:

1. Identification of potential down inverters
2. Identification of potential hourly window for downtime
3. Estimation of potential downtime

Process:

1. Data acquisition and quality check: **3 -5 days**
2. Assessment of data-engineering required: **3-5 days**
3. Understanding client requirements: **1-3 days**
   a. Downtime related
   b. Performance parameters related
4. Final model selection and deployment: **7-10 days**
5. Total time to deploy:
   a. with no additional data-engineering: **<2 weeks**
   b. with additional data engineering: **~3 weeks**

Potential ARR: USD 800,000
(estimated @USD 100 per MW per year)

GMR - 25 MW, SMA

Product Stack:

1. Identification of potential down inverters
2. Identification of potential hourly window for downtime
3. Estimation of potential downtime

Process:

1. Data Acquisition and Quality Check: **1-3 days**
2. Assessment of data-engineering required: **0 days**
3. Understanding client requirements: **1-3 days**
   a. Downtime related
   b. Performance parameters related
4. Final model selection and deployment: **1-3 days**
5. Total time to deploy:
   a. with no additional data-engineering: **<3 days**
   b. with additional data engineering: **~1 week**

Potential ARR: USD 25,000
(estimated @USD 100 per MW per year)

Improvements and Direction

Improvements & Direction

#	Category	Description
1	Survival analysis	Improve the model - survival analysis
2	Multi-classification models	Improve the multi-classification techniques
3	Training data size (Volume)	We have used only 2020-2021 data in the training for these results. With the increase in the size of the training data performance of the model may also improve.
4	Inverters’ OEM	Currently the model has been trained for SMA inverters. For the next steps, other OEMs should be explored and trained to build the portfolio.
5	Location data	Based-on the location of plants we can develop the model to understand and improve the geo-spatial variations.
6	Time-zone and day-time saving	Currently we have considered 7am to 5pm duration, however, this might not be accurate for a different time-zone or location. We should make a logic for considering sun-rise/sun-set or time-zone related variation.
7	Physical configuration	What changes we need to make into the model for bifacial solar panels? How will the model change, if any?
8	String and grid-side data	How to incorporate string or grid-side data? Can it be useful or improve the model's output?
9	‘Plug & Play’ model	How to make the model plug-and-play?

Bonus Project

Weather Forecasting

GMR

Project Statement

To predict the Irradiance values of GMR for tomorrow

- Status: Successful
- Daily Validation: pending
- Demo: Live