Solar Power Generation Dataset

Team P26:
Kristin Petersel, William Mukose, Martti Praks, Kaarel Tark

1st year Data Science MSc students

https://www.kaggle.com/anikannal/solar-power-generation-data
https://github.com/marttipraks/ML2020-p26

Project Goals

1. Can we predict the power generation for the next day?

Propose a way to address questions:

2. Can we identify the need for panel cleaning/maintenance?
3. Can we identify faulty or suboptimally performing equipment?

Note: No project owner and low domain knowledge in team

- Data from 2 solar plants
- Data from Solar inverters every 15 minutes
- 34 days
- 22 + 22 inverters

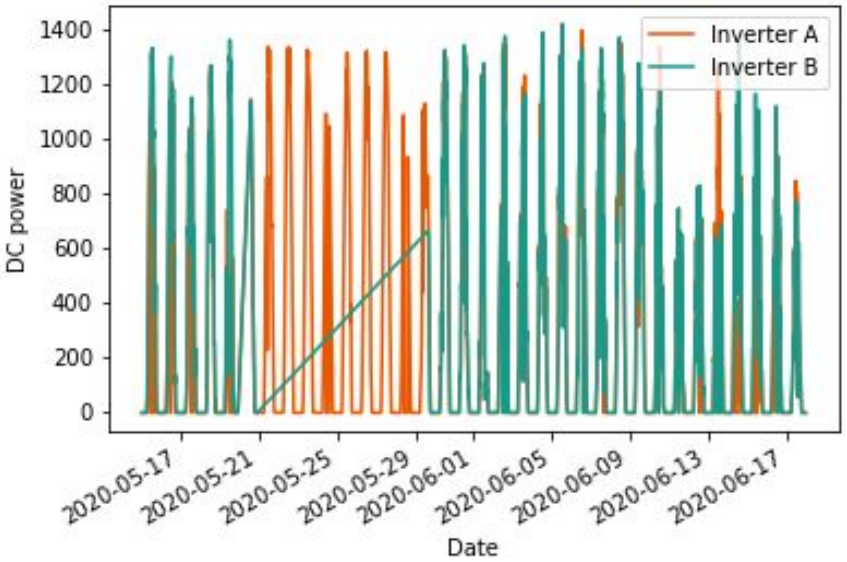
*Picture used from: https://www.cleanenergyreviews.info/blog/2014/5/4/how-solar-works

Methodology and Process Steps (Goal 1)

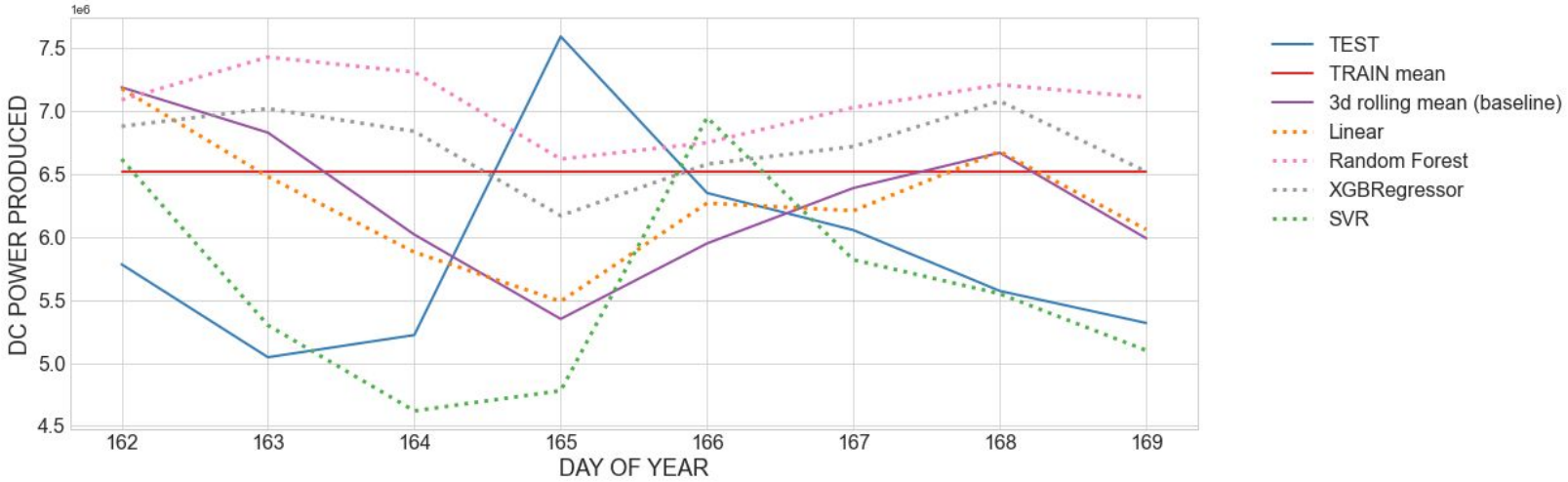
- Data split: 26 days training , 8 days test
- Data cleaning: temperature info added to the ones missing, missing records added with no electricity generation
- Features Engineering: previous 3 days' solar irradiation and power generation could be used to predict the next day's power generation
- Predictions on inverter level, but target on sum of inverters generated DC POWER for the day
- Last 3 days rolling mean as baseline
- Linear Regression, XGB Regressor, Support Vector Regression and Random Forest
    - All models use different selection of features identified during CV

Missing data

- Percentage of missing values (during daytime hours)
  - Plant 1 - 2.5 % DC - 2.7 %
  - Plant 2 - 11.7 % DC - 16.8 %
- Missing DC power values in some inverters



Plant 1

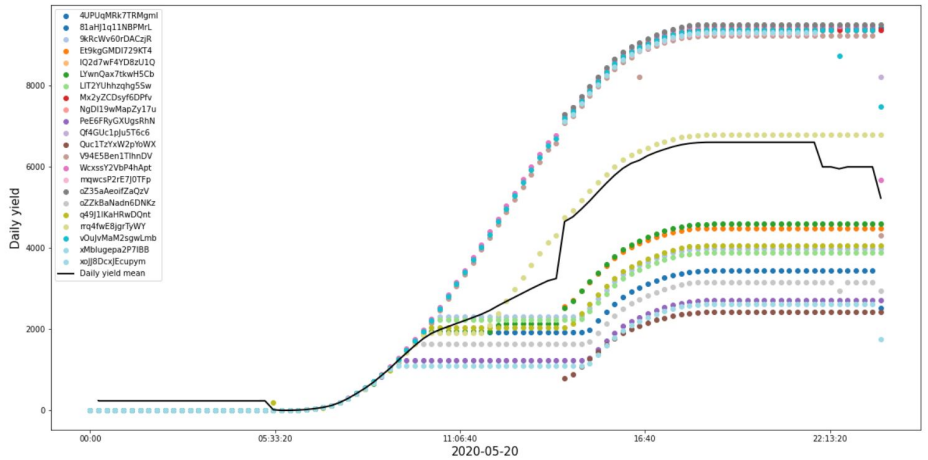


Method	**1 day pred. RMSE**	**1 day pred. MAE**	**15m inverter RMSE**
3d rolling mean (baseline)	3 079 888	1 823 703	2572
Linear Regression	1 262 887	1 091 779	1682
Random Forest	1 568 941	1 444 176	1808
XGBRegressor	1 320 368	1 214 055	1609
Support Vector Regression	**1 089 902**	**698 708**	1768

Results Other initial goals

- **Can we identify the need for panel cleaning/maintenance?**
- **Can we identify faulty or suboptimally performing equipment?**

➔ Could be detected by single inverter AC and DC differences compared to other inverters



We concluded that these questions could be answered with data analysis and do not need machine learning models.

Conclusions & lessons learned

- Possible to predict 1 day better than 3 day rolling mean (based on RMSE and MAE)
- To get better results
  - Need to know why missing data or sometimes generation info drops to 0
  - Need to get some info from the future - for example predicted ambient temperature/irradiation based on weather forecast (introduce new data)
  - Longer period of time data
- Lessons learned:
  - Spend more time on data preparation and features engineering
  - Check data quality (0 and also missing records!)
  - You cannot see into the future
  - Features engineering for time-series
  - A lot about regression in machine learning, as in class we mainly looked at classification problems
  - How to divide work in a team for a data science project

Thank you!

Link to the repository

https://github.com/marttipraks/ML2020-p26