SOLAR POWER GENERATION

TEAM 5 FINAL PRESENTATION
APRIL 9, 2021

Team Members

César Montilla

Connor Mattinson

Eric Liu

Peter Friedrich

Shiva Eslami

Agenda

- Business Objectives
- Context & Data Acquisition
- Data Processing & EDA
- ML Modelling
- Conclusions & Recommendations

Challenges with Solar Power

Cyclical & Intermittent Nature:

- **Peak solar potential does not always correlate with peak power demand.**
  - This lack of correlation between production outputs and consumer demands can result in wasted power surpluses or unfulfilled power demands.

- **Changing weather conditions can greatly impact outputs.**
  - The somewhat chaotic nature of weather can impact producers capacity to make firm generation commitments and therefore create uncertainty in appropriate power augmentation for distributors.

- **Equipment sensitivity can result in unexploited potential.**
  - Importance of understanding whether performance shortfalls are a result of faulty equipment, physical obstruction, etc. This issue is not unique to solar, however losses can not be recovered.

Problem Exploration

Can we generate a model to produce accurate solar potential forecasting?

Opportunistic Energy Storage: By predicting future output and comparing historical demand trends, can generators transform surplus power for additional revenue?

Powering an Electrolyzer for Hydrogen Extraction
*(*Gandikotta , GNSS Irrigation Canal)

Appropriate Maintenance Scheduling: Is it Opportunistic to Deploy Maintenance Efforts During Generating Hours?

Will Maintenance Downtime Result in Greater Losses than Reduced Overall Production Due to Failing or Faulty Equipment?

Power Commitments & Power Trading: Can the Model Accurately Predict AC Output Such that Generation Commitments can be made and Appropriate Pricing Set?

If Accurate Predictions Are Demonstrated Can Price Volatility be Reduced?

Business Evaluation

Outcome :
Improve Plant Power Management with Accurate Power Output Forecasting.

Action :
Opportunistic Energy Storage, Predictive Power Pricing, Accurate Power Scheduling, Informed Maintenance Scheduling.

Judgement :
Can the Models Accuracy be Trusted to Make Critical Business Decisions?

Cost of Prediction	
Accurate Predictions	**Inaccurate Predictions**
Reduce Potential for Wasted Power.	Potential for Revenue Loss.
Fulfill Power Commitments to Stakeholders	Over Commitment of Available Power.
Predictive Power Pricing.	Pricing of Supplied Power.
Informed Maintenance Scheduling	Production Losses Due to Untimely Maintenance.

Performance Measurements

By Accurately Predicting AC Outputs:

Power Surplus Prediction & Exploitation:

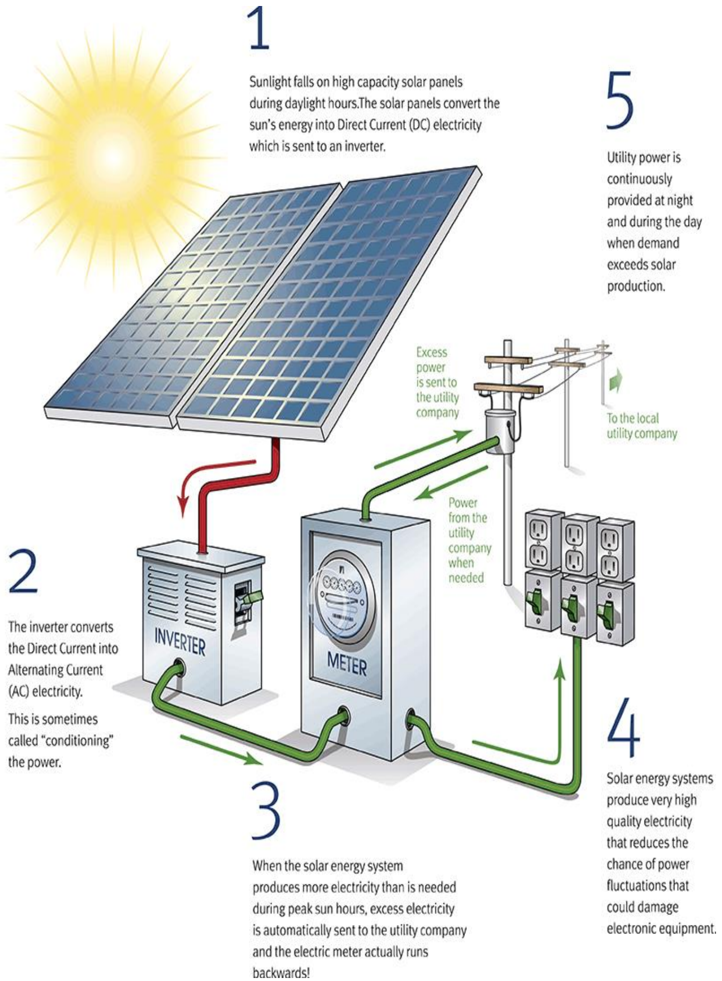
Excess power is typically wasted as the cost of terminalling to other grids is costly and in some cases logistically impossible. Understanding whether the plant is a strong candidate for energy storage infrastructure and the optimal times to engage such infrastructure would positively impact the producer and could assist in fulfilling future power demands if forecasted outputs fall short of commitments.

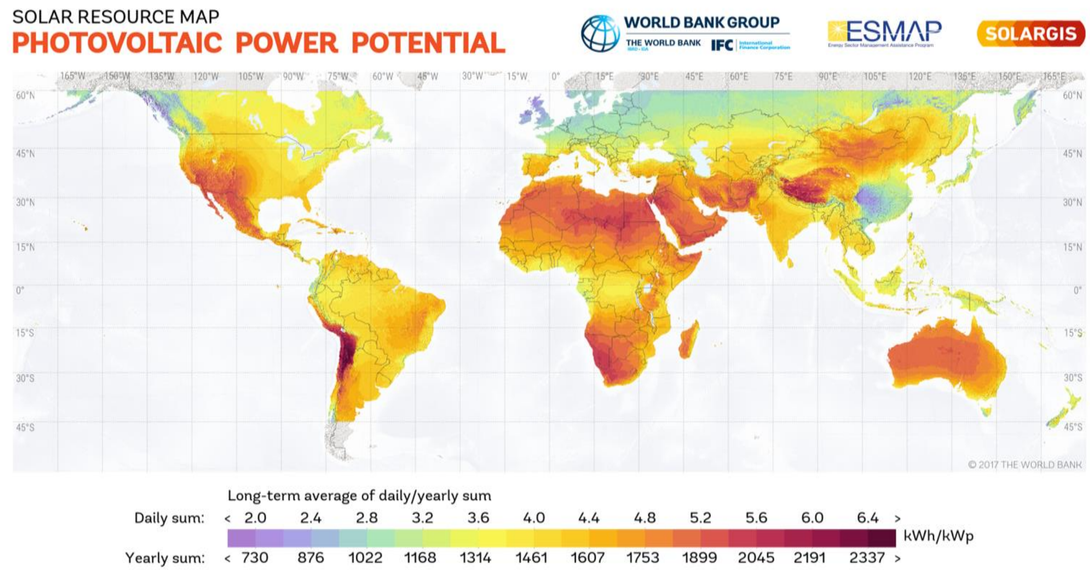
Power Commitments Forecasting and Fulfillment:

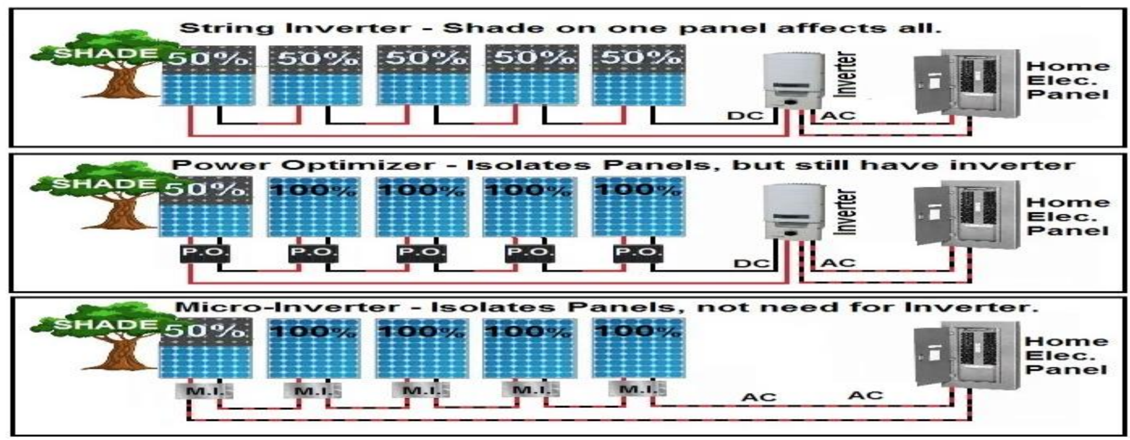
If power commitments can be forecasted to include longer periods, power suppliers can collaborate to set daily production caps and production times to mitigate instances of over producing. Pricing volatility can be avoided so long as production volatility is controlled and demand follows a standard pattern.

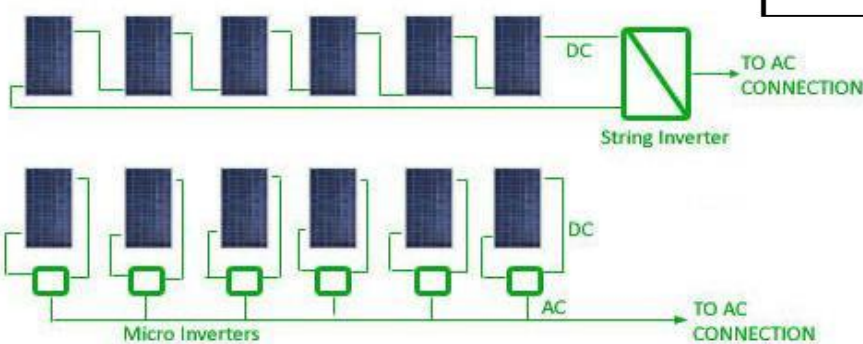
Optimizing Maintenance Scheduling & Daily Yields:

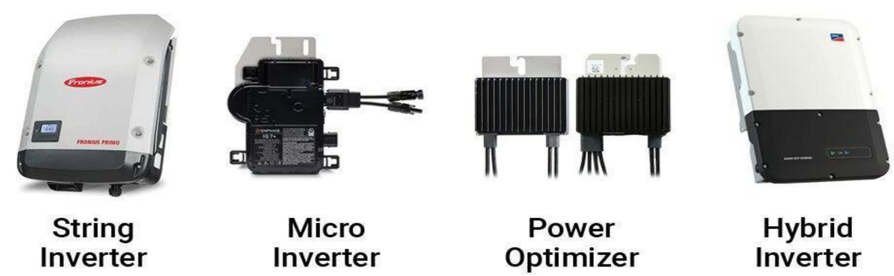
By comparing historical outputs with actual vs. predicted outputs, producers would have meaningful insights as to whether sensors are performing optimally during producing hours. Deploying maintenance efforts on an adhoc basis could significantly improve outputs.

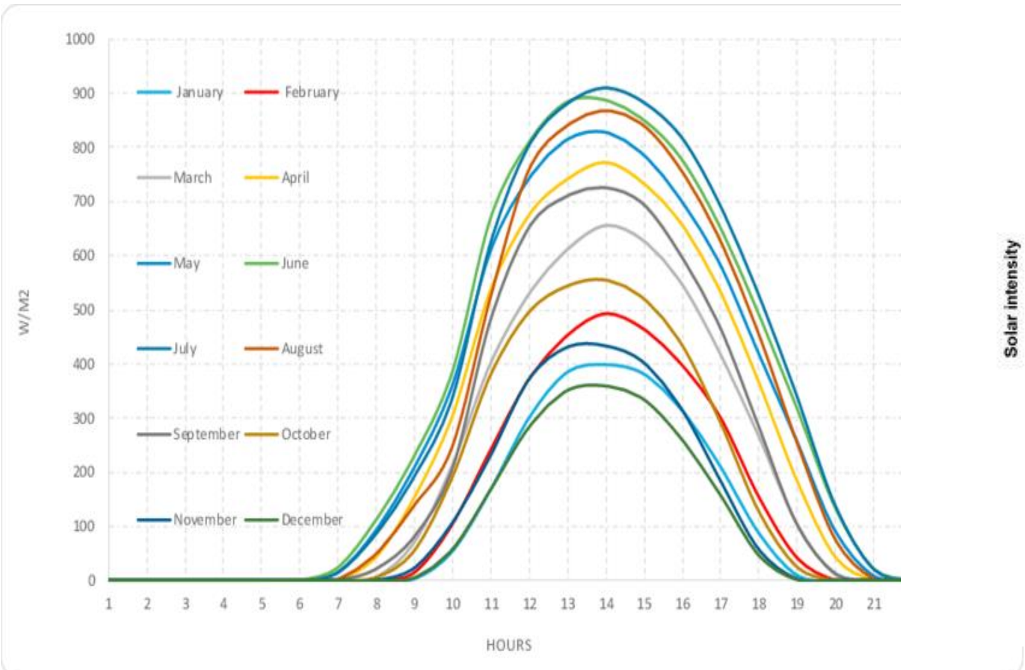

1
Sunlight falls on high capacity solar panels during daylight hours.The solar panels convert the sun's energy into Direct Current (DC) electricity which is sent to an inverter.
2
The inverter converts the Direct Current into Alternating Current (AC) electricity.
This is sometimes called "conditioning" the power.
3
When the solar energy system produces more electricity than is needed during peak sun hours, excess electricity is automatically sent to the utility company and the electric meter actually runs backwards!
4
Solar energy systems produce very high quality electricity that reduces the chance of power fluctuations that could damage electronic equipment.
5
Utility power is continuously provided at night and during the day when demand exceeds solar production.
Excess power is sent to the utility company
Power from the utility company when needed
To the local utility company
INVERTER
METER

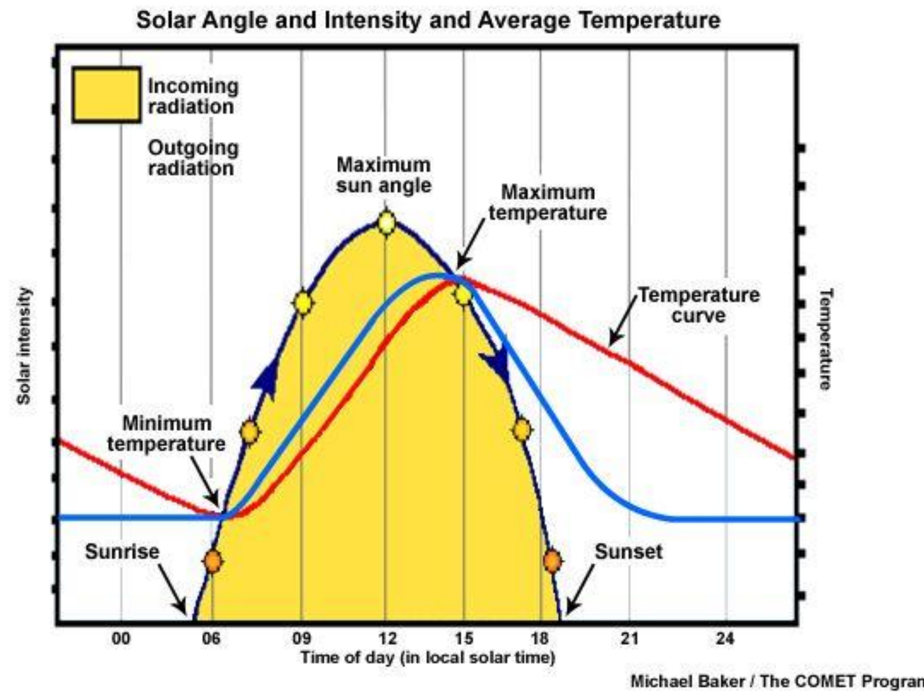

SOLAR RESOURCE MAP
PHOTOVOLTAIC POWER POTENTIAL
WORLD BANK GROUP
ESMAP
SOLARGIS
© 2017 THE WORLD BANK
Long-term average of daily/yearly sum
Daily sum: < 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8 5.2 5.6 6.0 6.4 >
kWh/kWp
Yearly sum: < 730 876 1022 1168 1314 1461 1607 1753 1899 2045 2191 2337 >


String Inverter - Shade on one panel affects all.
SHADE
50% 50% 50% 50% 50%
Inverter
DC
AC
Home Elec. Panel
Power Optimizer - Isolates Panels, but still have inverter
SHADE
50% 100% 100% 100% 100%
P.O.
Inverter
DC
AC
Home Elec. Panel
Micro-Inverter - Isolates Panels, not need for Inverter.
SHADE
50% 100% 100% 100% 100%
M.I.
AC
AC
Home Elec. Panel

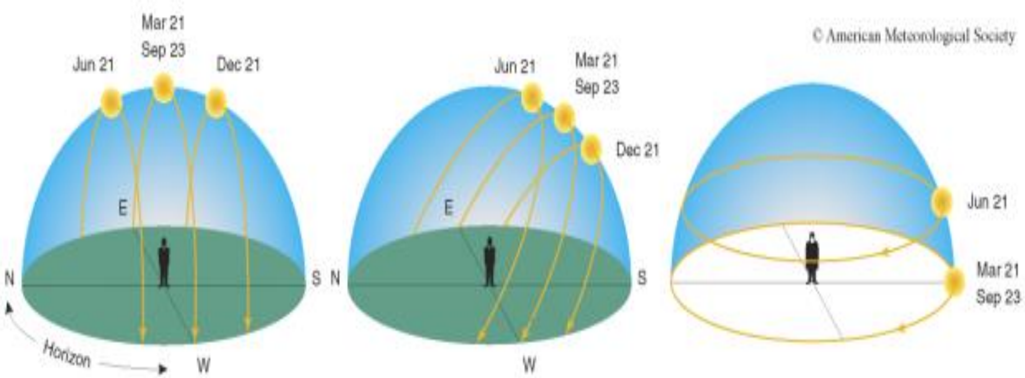

DC
TO AC CONNECTION
String Inverter
DC
AC
TO AC CONNECTION
Micro Inverters


String Inverter
Micro Inverter
Power Optimizer
Hybrid Inverter





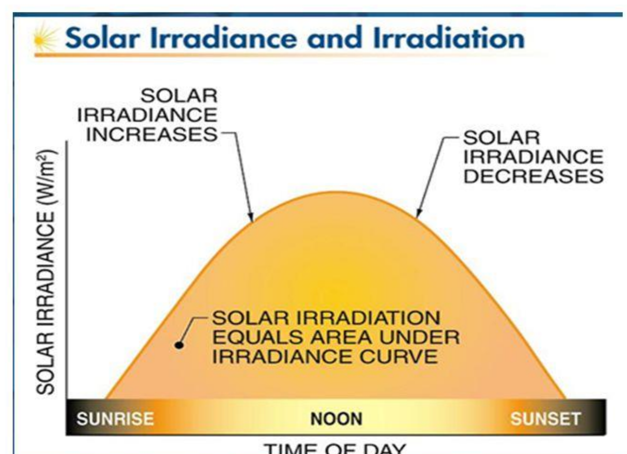
Solar Radiation



(a) At the Equator

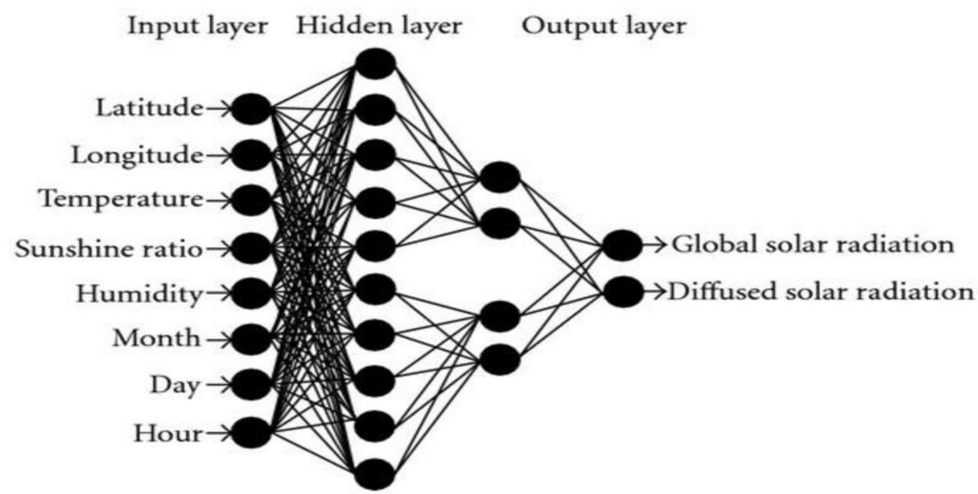
(b) At mid latitudes (Northern Hemisphere)

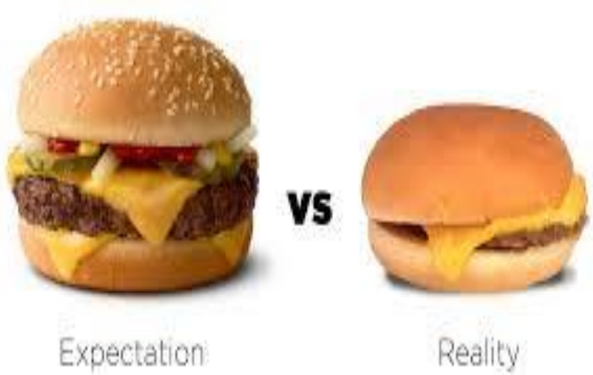
(c) At the North Pole

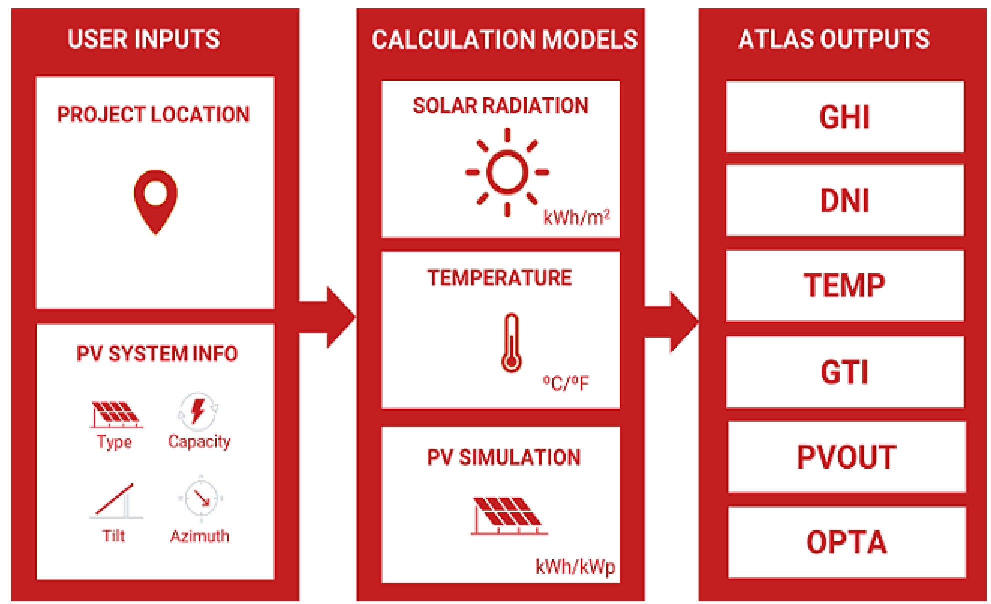


Map data (min-max range) Per day

Specific photovoltaic power output	PVOUT	3.39 –	5.24	kWh/kWp
Direct normal irradiation	DNI	2.51 –	5.81	kWh/m²
Global horizontal irradiation	GHI	3.77 –	5.64	kWh/m²
Diffuse horizontal irradiation	DIF	1.52 –	2.65	kWh/m²
Global tilted irradiation	GTI	4.13 –	6.26	kWh/m²
Optimum tilt of PV modules	OPTA	10 –	35	°
Air temperature	TEMP	-14.3 –	29.1	°C
Terrain elevation	ELE	-2 –	8586	m







Data Limitations & Assumptions

Limitations:

Data is collected at two plants between the month of May and June.

No information on physical setup and choice of solar module or inverter

Weather info collected by a single sensor. Solar irradiation info can be misleading.

Very little weather information besides temperature and irradiance.

Assumptions:

Data are accurate and timely collected

Inverters are of the same model

Series circuit wiring of solar array is identical for all inverters

The cumulative daily yield and total yield was accumulated correctly

Data Understanding - Features

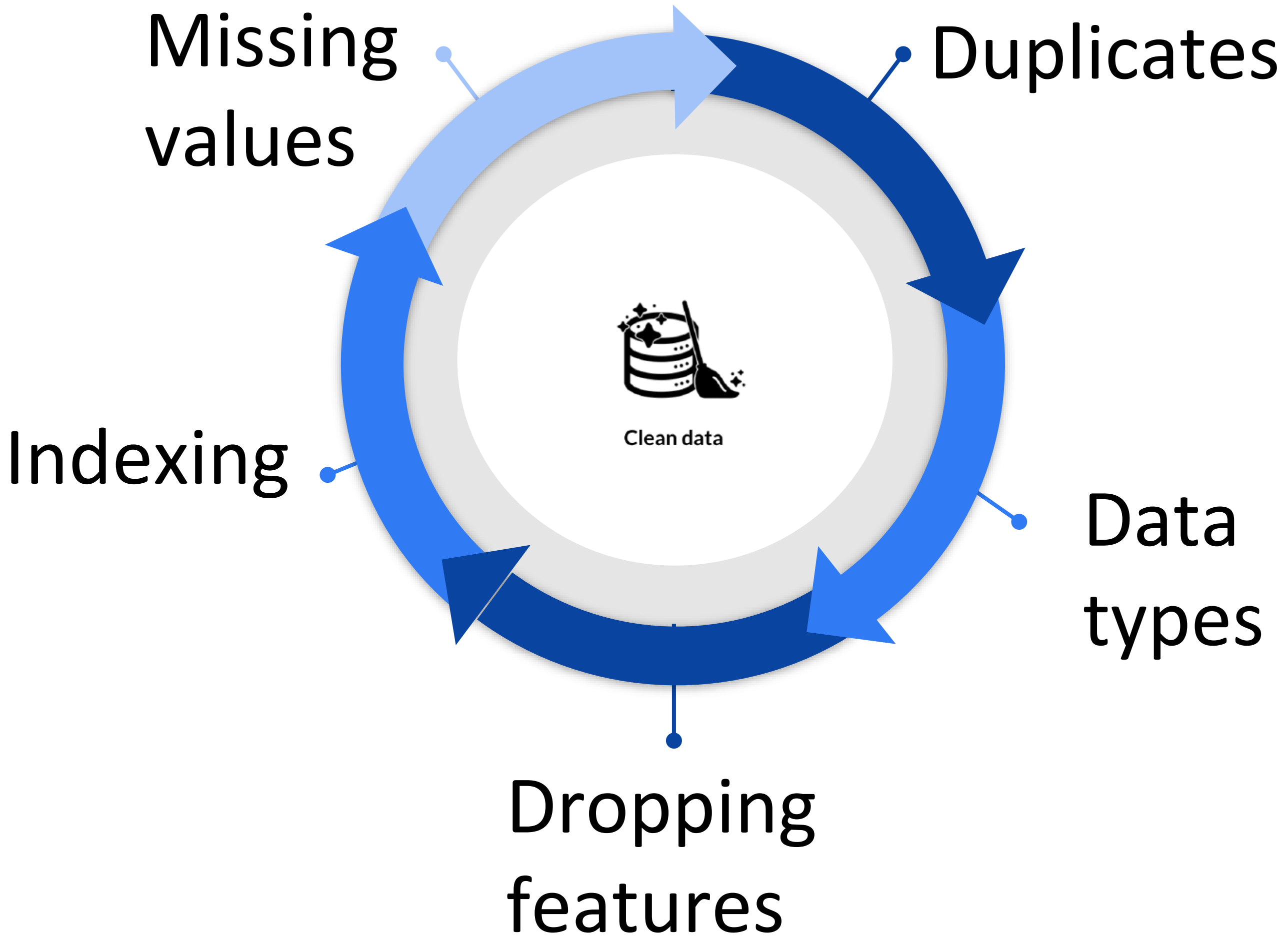
Plant data

	DATE_TIME	PLANT_ID	SOURCE_KEY	DC_POWER	AC_POWER	DAILY_YIELD	TOTAL_YIELD
0	15-05-2020 00:00	4135001	1BY6WEcLGh8j5v7	0.0	0.0	0.0	6259559.0
1	15-05-2020 00:00	4135001	1IF53ai7Xc0U56Y	0.0	0.0	0.0	6183645.0
2	15-05-2020 00:00	4135001	3PZuoBAID5Wc2HD	0.0	0.0	0.0	6987759.0
3	15-05-2020 00:00	4135001	7JYdWkrLSPkdwr4	0.0	0.0	0.0	7602960.0
4	15-05-2020 00:00	4135001	McdE0feGgRqW7Ca	0.0	0.0	0.0	7158964.0

Sensor data

	DATE_TIME	PLANT_ID	SOURCE_KEY	AMBIENT_TEMPERATURE	MODULE_TEMPERATURE	IRRADIATION
0	2020-05-15 00:00:00	4135001	HmiyD2TTLFNqkNe	25.184316	22.857507	0.0
1	2020-05-15 00:15:00	4135001	HmiyD2TTLFNqkNe	25.084589	22.761668	0.0
2	2020-05-15 00:30:00	4135001	HmiyD2TTLFNqkNe	24.935753	22.592306	0.0
3	2020-05-15 00:45:00	4135001	HmiyD2TTLFNqkNe	24.846130	22.360852	0.0
4	2020-05-15 01:00:00	4135001	HmiyD2TTLFNqkNe	24.621525	22.165423	0.0

Data cleaning & preprocessing



https://thenounproject.com/

Exploratory Data Analysis - Introduction

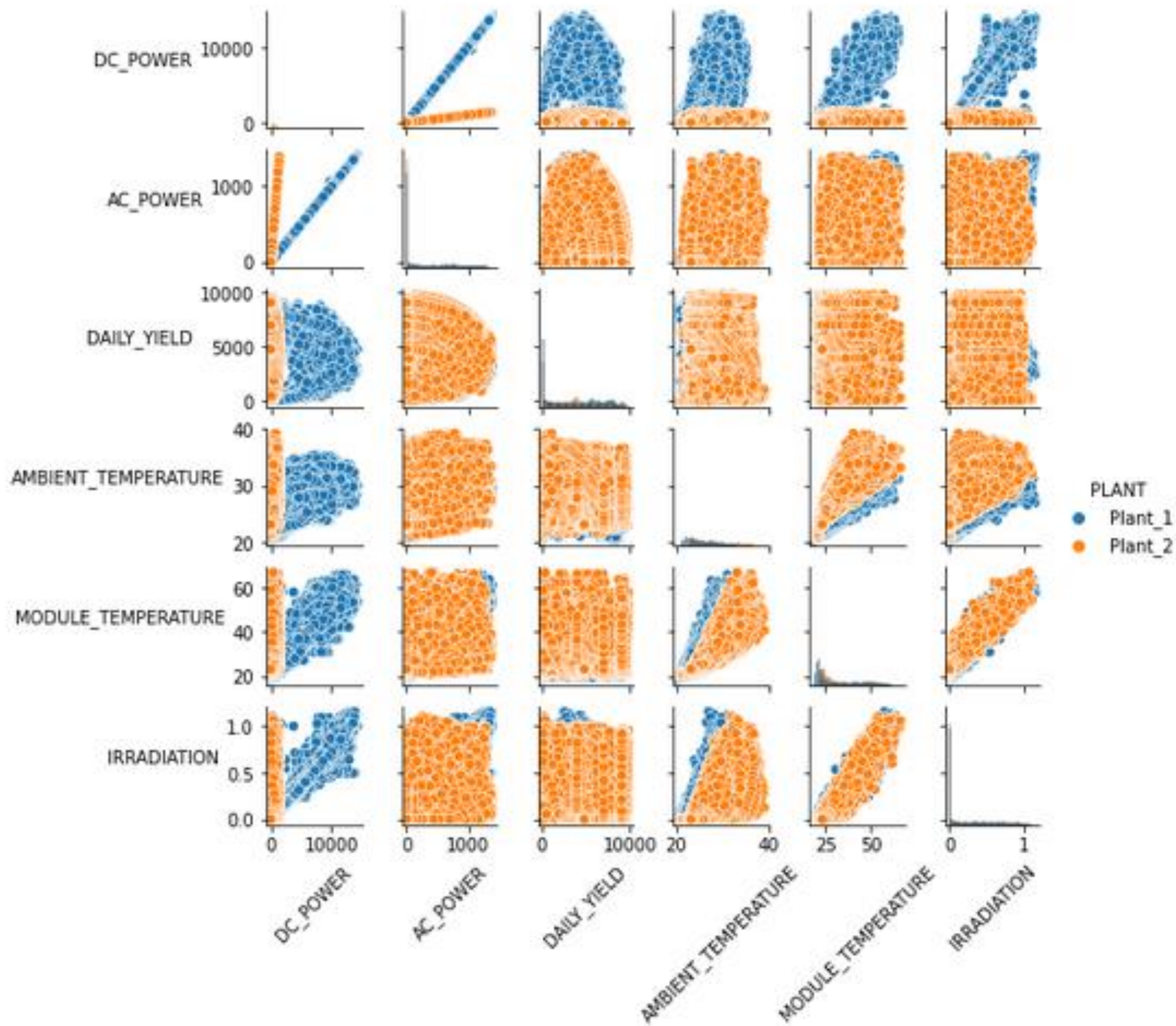
01 Problems with the dataset

02 Data ready to use

03 Answer your QuAM question

https://thenounproject.com/

EDA - Bivariate analysis



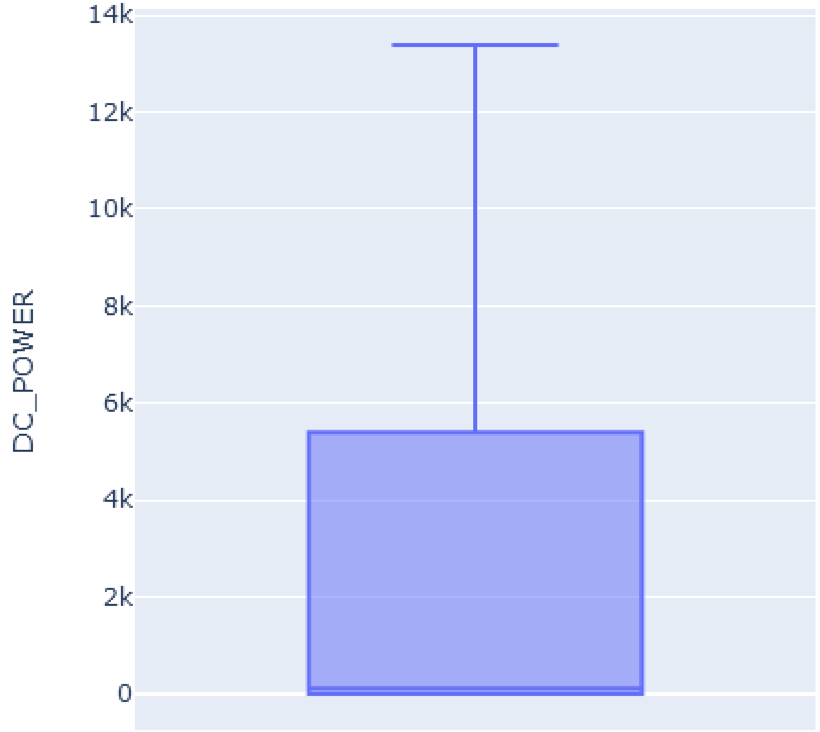
DC ∝ AC

Irradiation ∝ DC

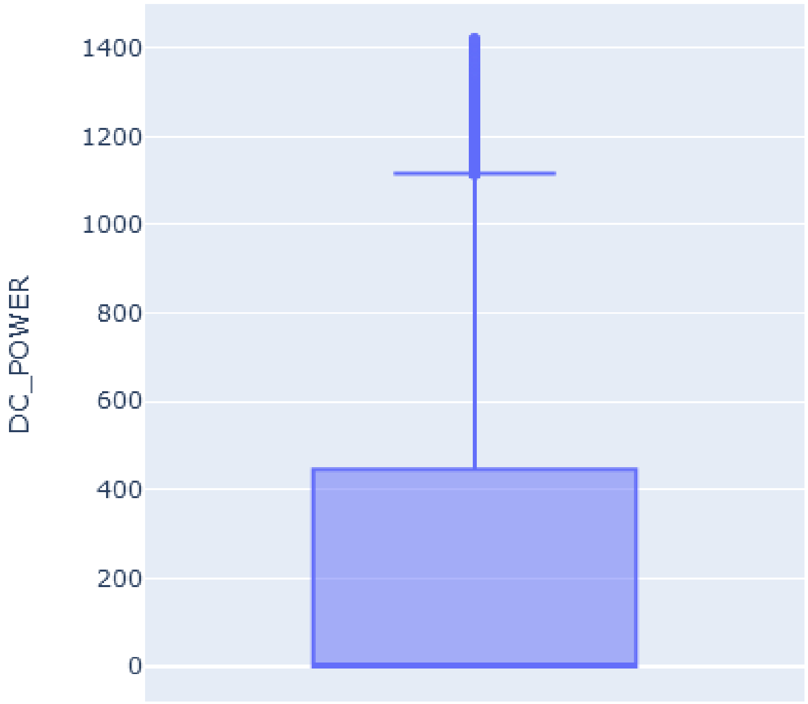
Ambient Temp ∝ Module temp

EDA - Boxplots for DC power

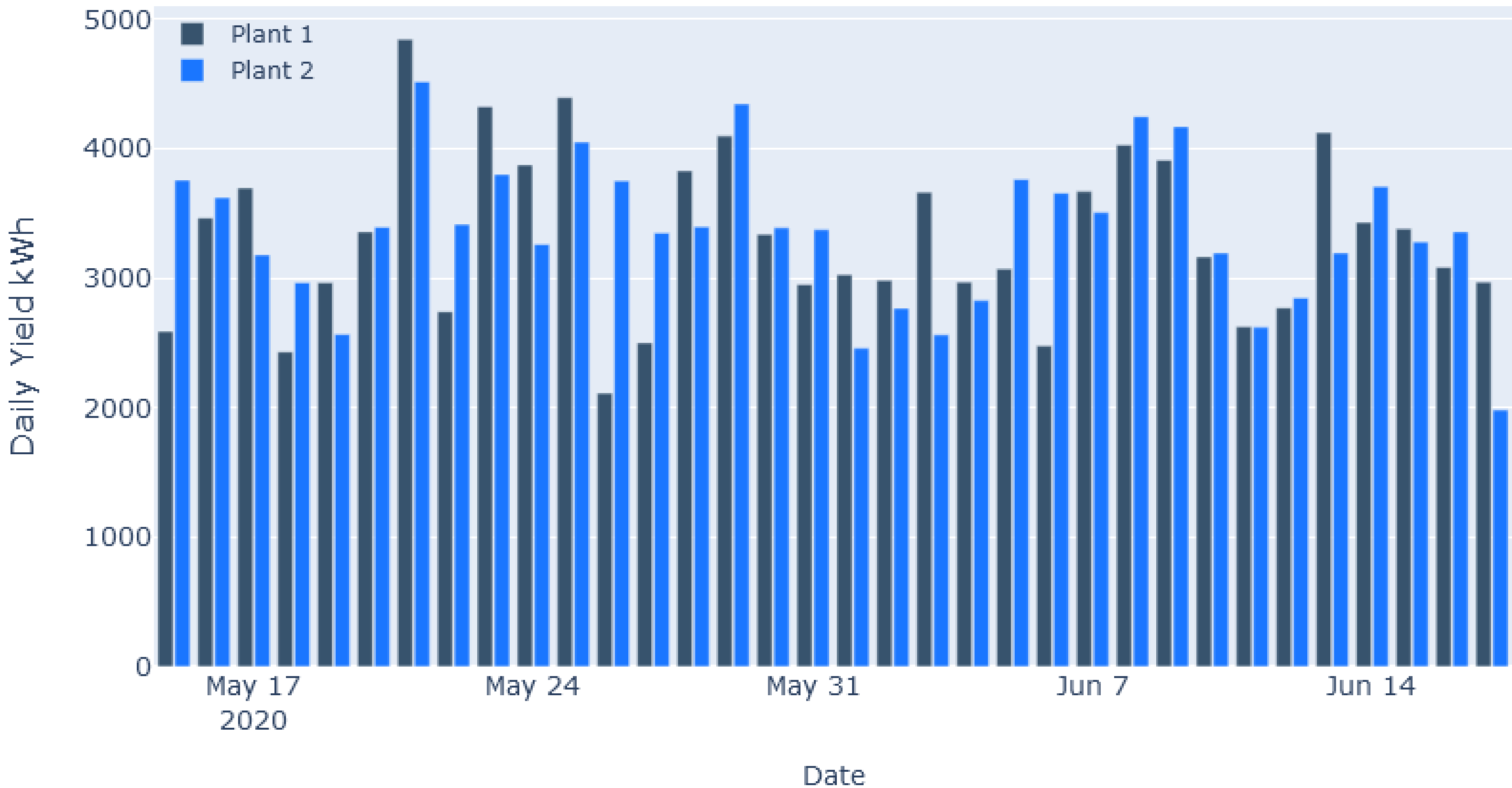
Plant 1



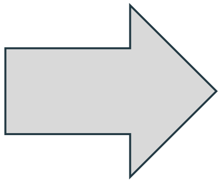
Plant 2



EDA - Daily Yield



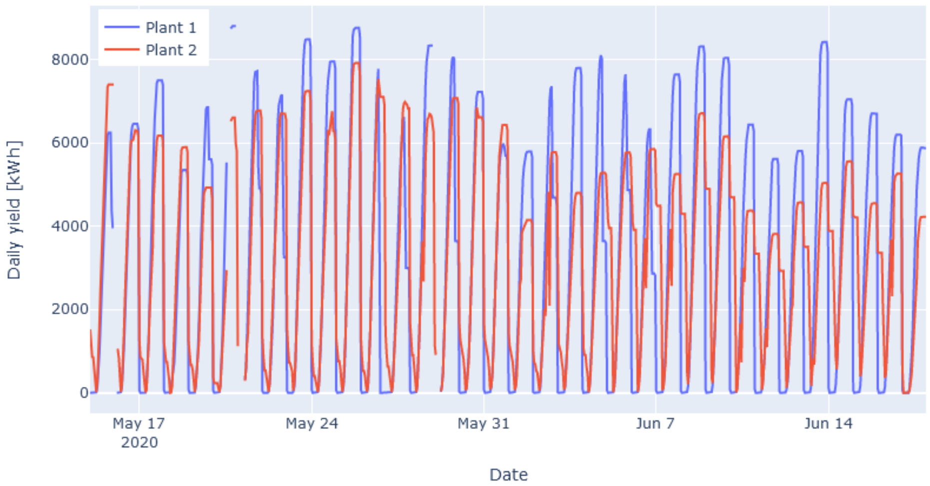
Avg. Daily yield ≈ 3300 kWh

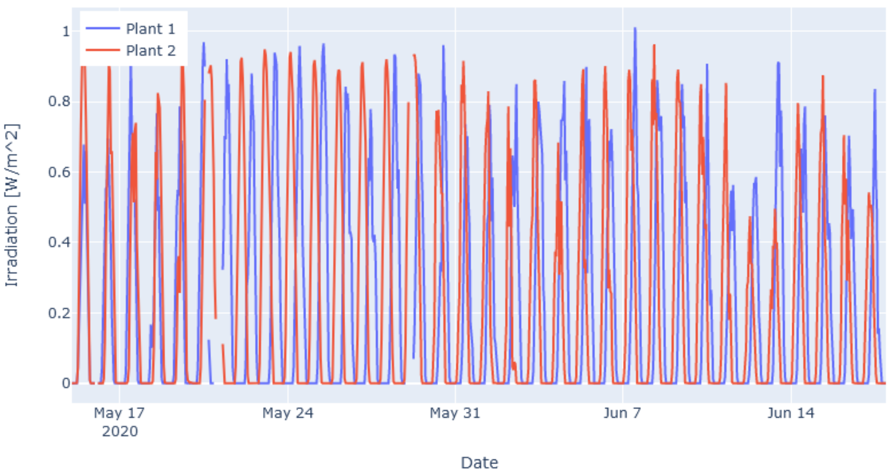


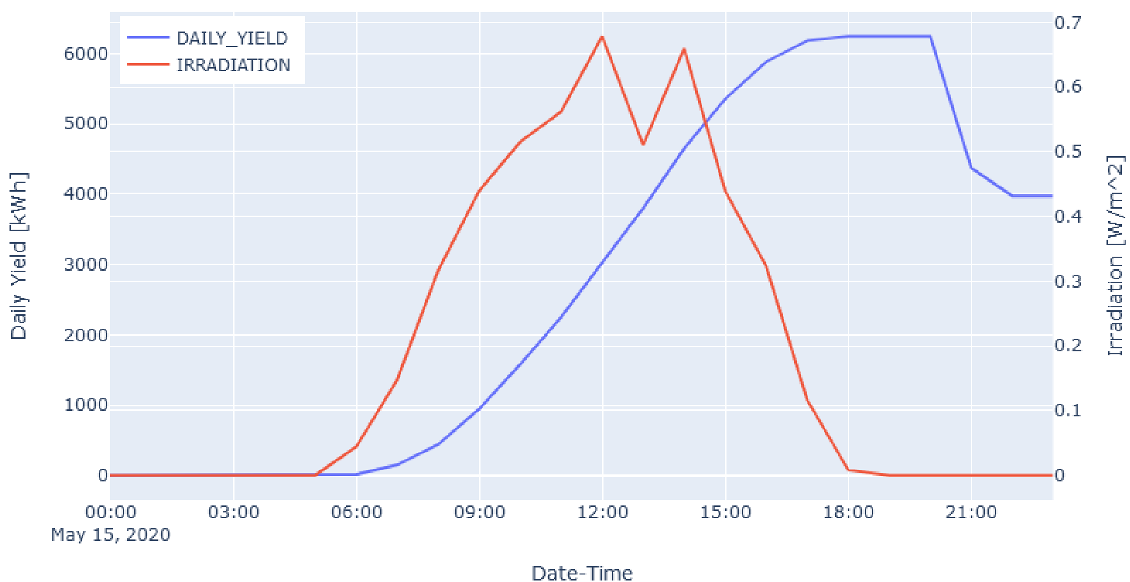
Enough to power 164 homes in Alberta*

*https://www.atco.com/en-ca/for-home/energy-101.html

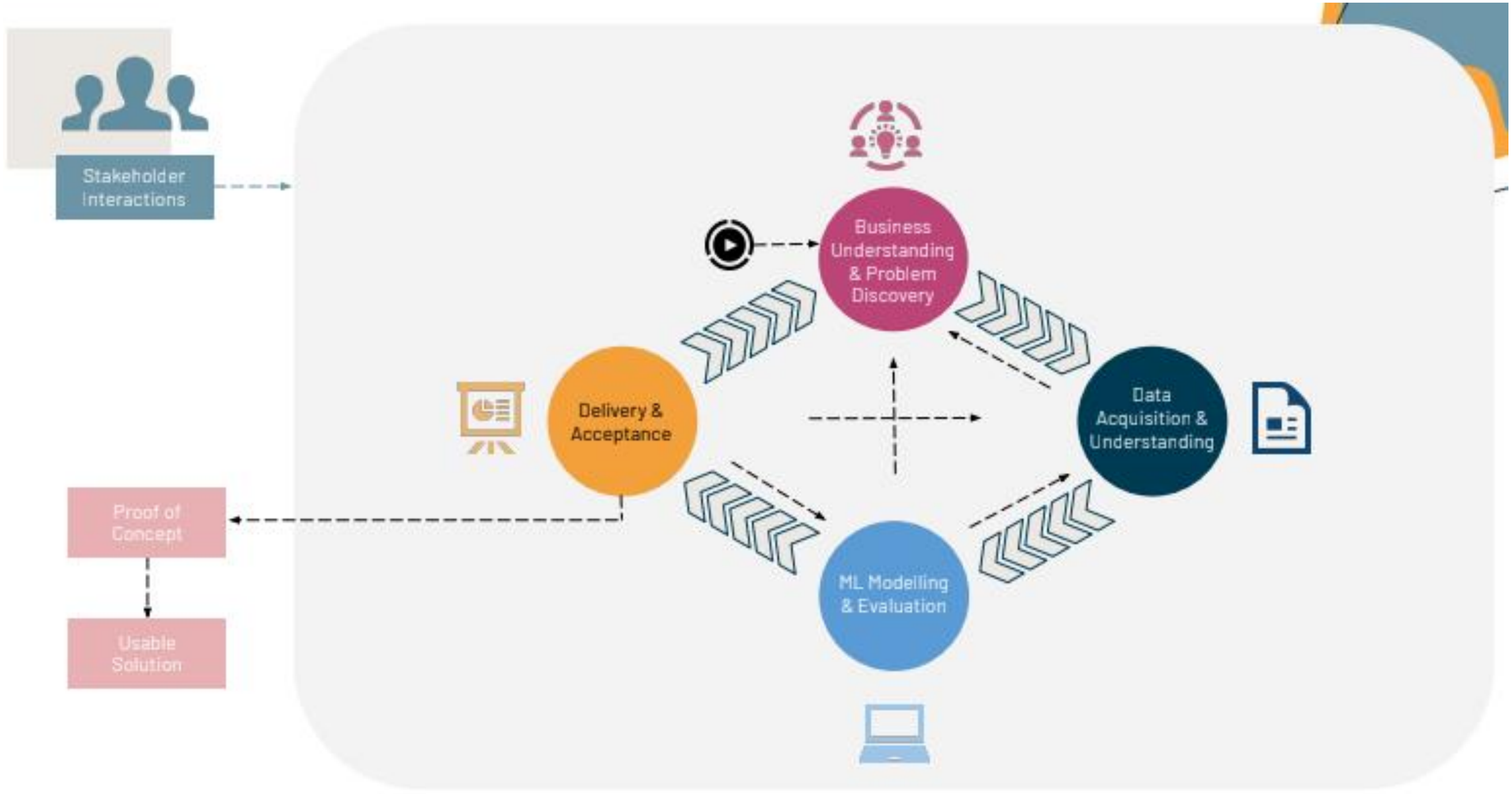
EDA - Basic time series analysis

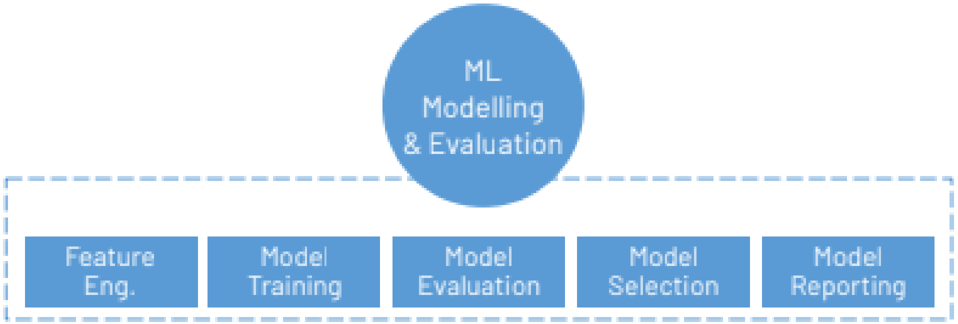






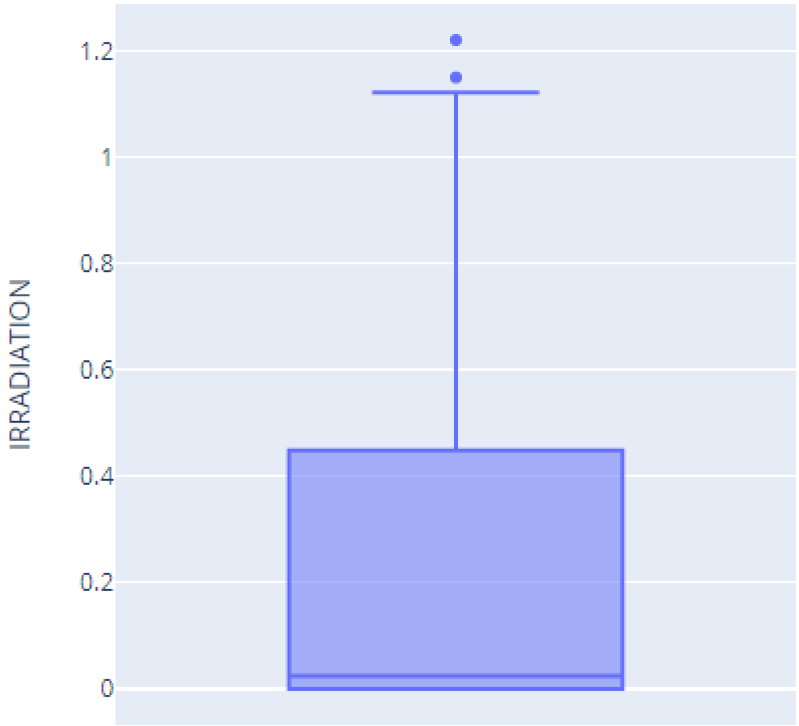
ML Modelling & Evaluation





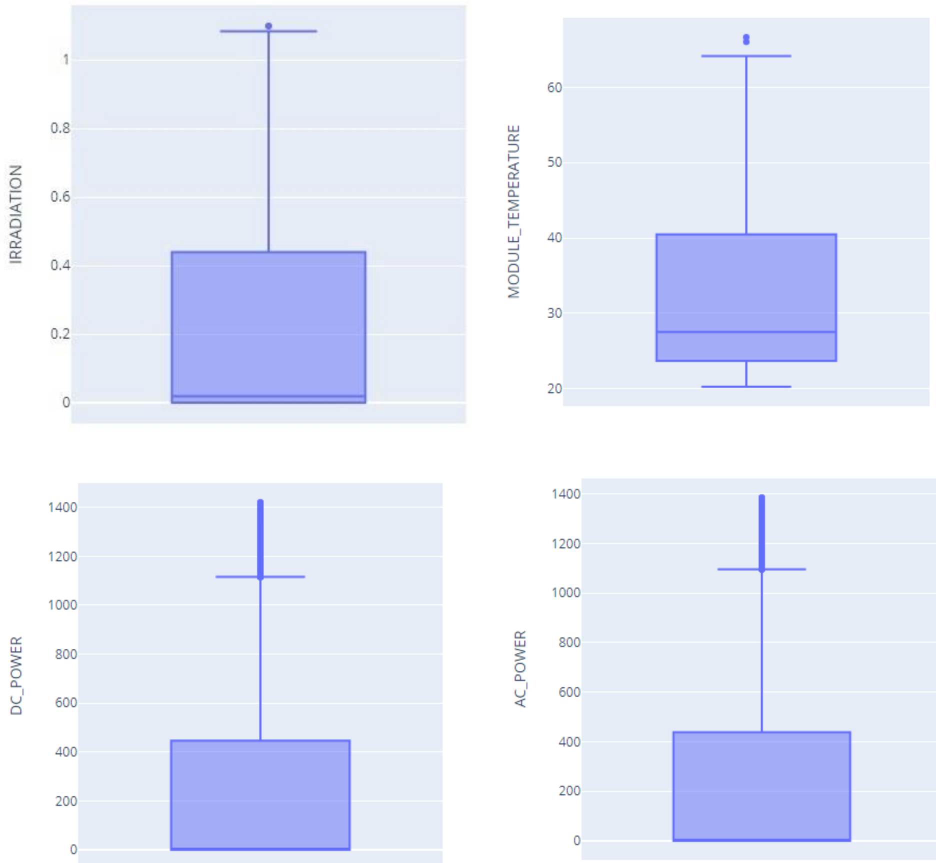
Feature Scaling/ Outliners

Plant 1



Feature Scaling/ Outliners

Plant 2

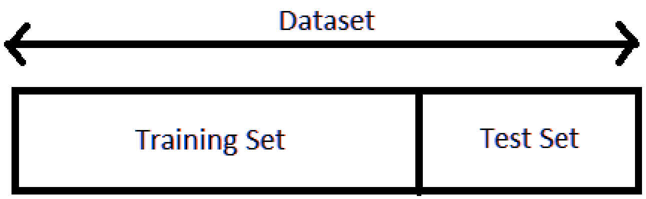


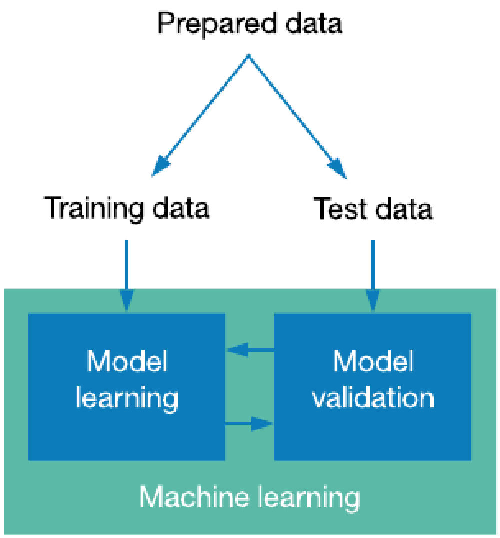
Extreme Values

```
AC_POWER              4.421105
AC_POWER_norm         4.421105
AC_POWER_std          4.421105
DAILY_YIELD           0.000000
DAILY_YIELD_norm      0.000000
DAILY_YIELD_std       0.000000
DATE_TIME             0.000000
DC_POWER              4.548140
DC_POWER_norm         4.548140
DC_POWER_std          4.548140
PLANT_ID              0.000000
SOURCE_KEY            0.000000
TOTAL_YIELD           0.000000
TOTAL_YIELD_norm      0.000000
TOTAL_YIELD_std       0.000000
dtype: float64
```

The outliers for the features AC_POWER and DC_POWER represent 4.4 and 4.5% of the data, respectively.

Train test-split

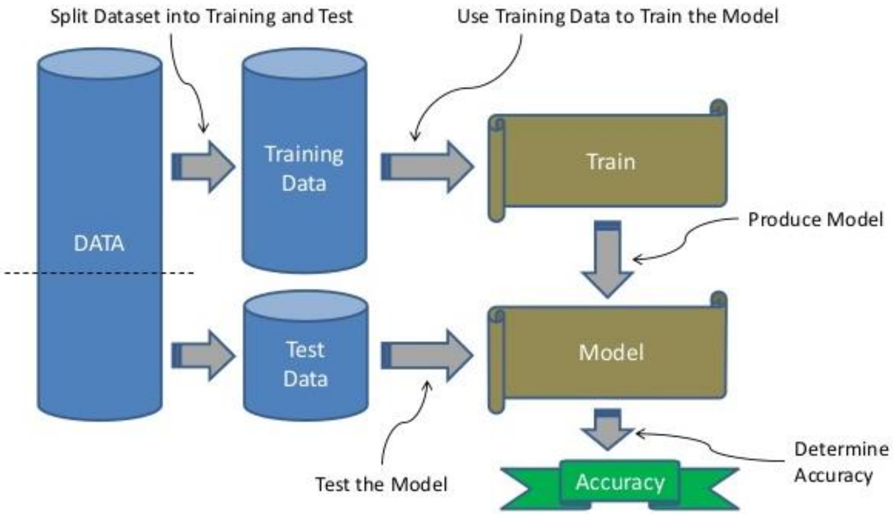




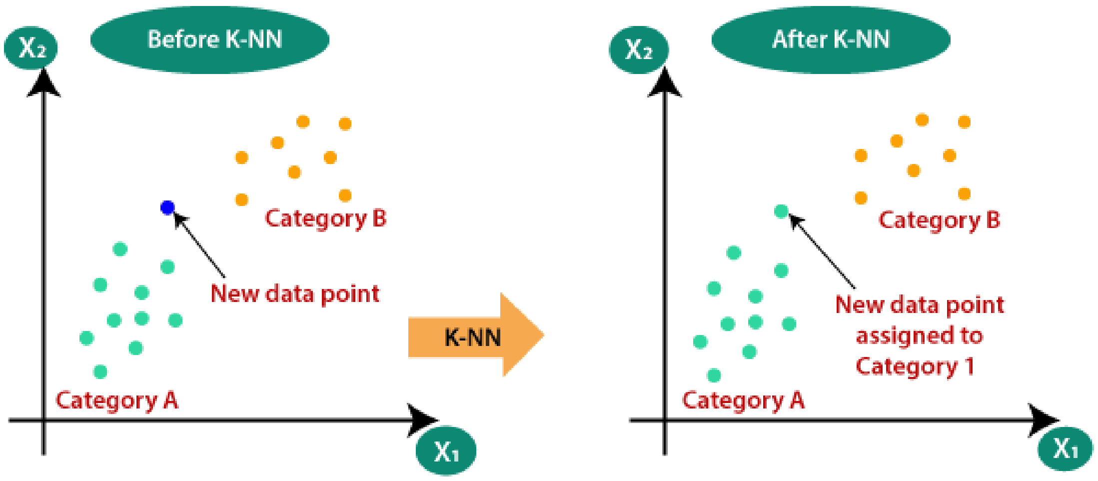
Tried two recommended types:

- Random Split, 80% train, 20% test randomly
- Time Series Split, first 80% train, last 20% test

→ Chose the second one, felt it was more robust



K-NN modelling



Target → DAILY_YIELD

```
pandas.qcut()
label_categories=["very low", "low", "high", "very_high"]
very_low = (target <= y_mean - y_std)
low = (target > y_mean - y_std) & (target < y_mean)
high = (target < y_mean + y_std) & (target > y_mean)
very_high = (target >= y_mean + y_std)
```

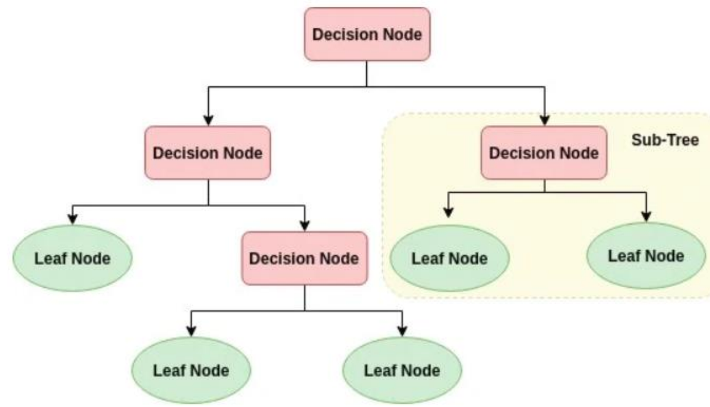
K-NN modelling

Using the 'distance' weights parameter increases the kNN model accuracy for the training dataset.

For the test dataset there is a slight decrease on accuracy.

	Uniform		Distance	
	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
kNN Neighbors number **5**	0.4808	0.3901	0.6707	0.3775
kNN Neighbors number **11**	0.4559	0.4017	0.6707	0.3833
kNN Neighbors number **15**	0.4492	0.4083	0.6707	0.3858

Decision Tree Modelling



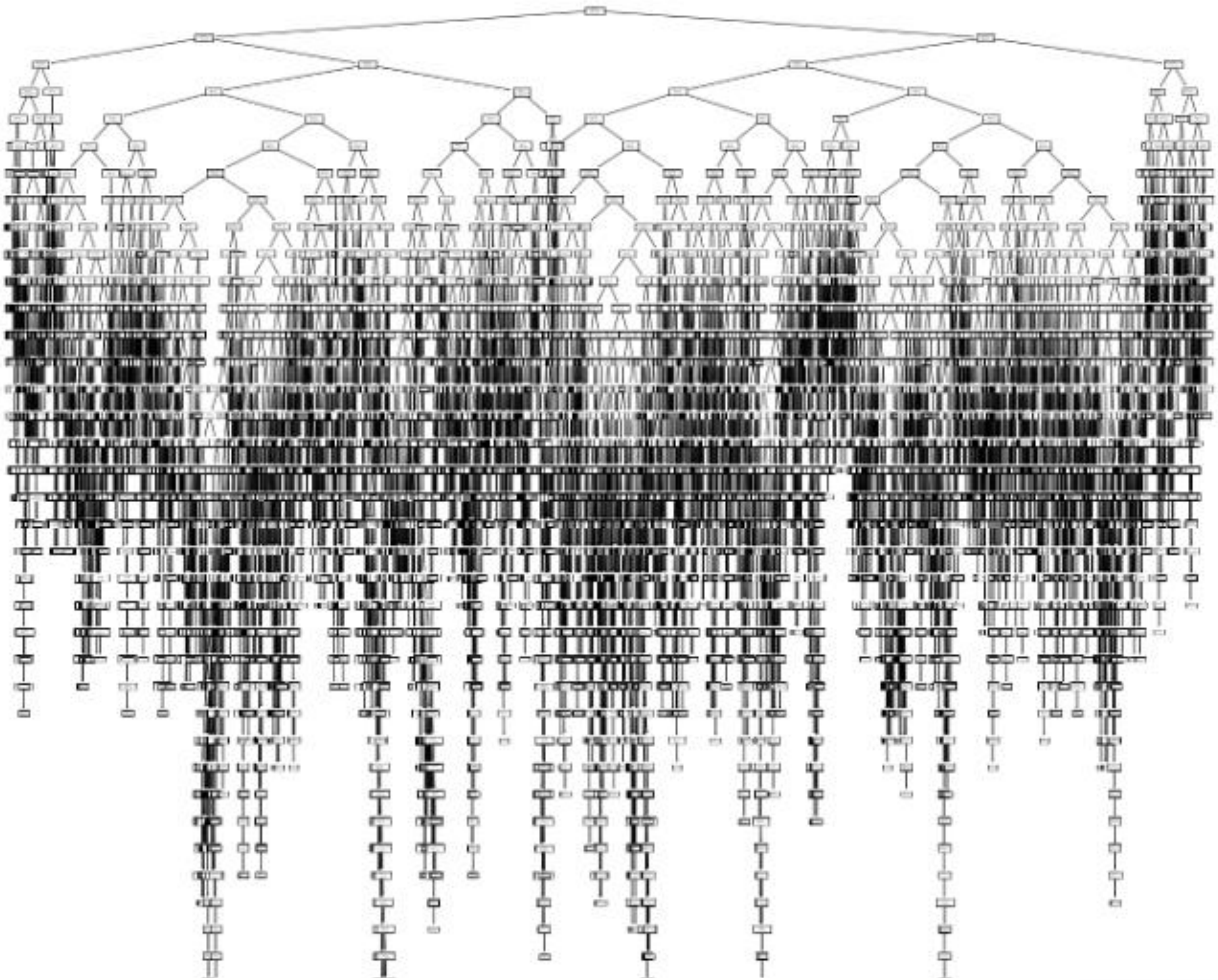
Which feature was used for the first split?

```
print(tree_rules[:250])
```

```
|--- AC_POWER <= 110.93
|   |--- AMBIENT_TEMPERATURE <= 22.83
|   |   |--- PLANT_ID <= 4135501.00
|   |   |   |--- AMBIENT_TEMPERATURE <=
|   |   |   |   |--- MODULE_TEMPERATURE
|   |   |   |   |   |--- MODULE_TEMPERAT
```

```
Results on training data:
              precision    recall  f1-score   support

        high       0.93      0.86      0.90     26967
         low       0.98      0.94      0.96     26916
   very_high       0.82      0.93      0.87     26974
    very_low       0.93      0.93      0.93     27099

    accuracy                           0.91    107956
   macro avg       0.92      0.91      0.91    107956
weighted avg       0.92      0.91      0.91    107956


Results on test data:
              precision    recall  f1-score   support

        high       0.83      0.78      0.80      6757
         low       0.92      0.88      0.90      6832
   very_high       0.75      0.84      0.79      6763
    very_low       0.92      0.91      0.92      6638

    accuracy                           0.85     26990
   macro avg       0.85      0.85      0.85     26990
weighted avg       0.85      0.85      0.85     26990
```

Decision Tree Modelling



How many leaves are in the optimal classifier/QuAM? Answer: 8609

	Train Accuracy	Test Accuracy	
kNN Neighbors number (5/ distance)	0.603	0.501	
kNN Neighbors number (11/ distance)	0.573	0.511	
kNN Neighbors number(15/ distance)	0.564	0.515	
DT / gini	0.914	0.85	
DT/ Entrophy	0.914	0.85	
DT/ splitter= best	0.914	0.85	
DT / splitter= random	0.914	0.848	
DT/ min_samples_leaf=1	0.914	0.85	
DT/ min_samples_leaf=2	0.897	0.846	

Switching to Regression

- Regression is directly predicting a continuous number instead a category i.e tomorrow's AC_POWER at noon.

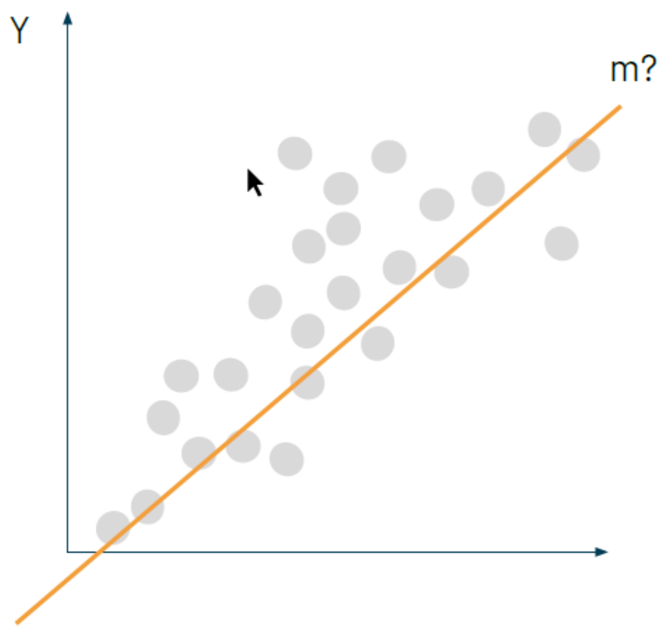
  Previous: Power = High

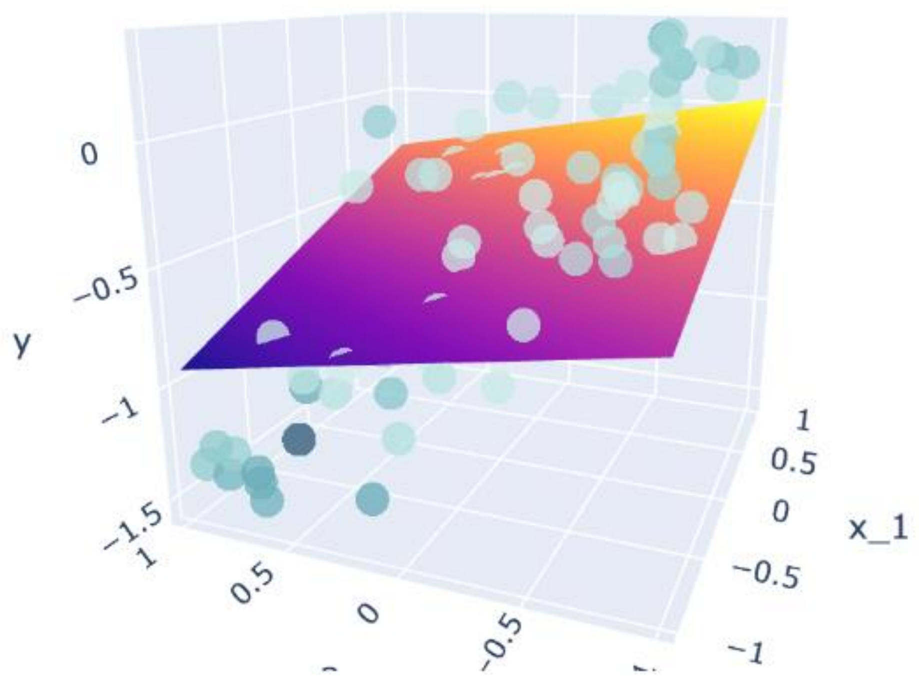
  Now: Power= 240.4 Watts

- This requires either new types of models, or reformulations of our previous models
- Split into train, test, choose baseline, then fit models and tune hyperparameters

Linear Regression Modelling

- Fit a linear function to your data that displays the overall linear trend
- A plane in 3D, a hyperplane in higher dimensions





Linear Regression Model Fitting

- Location and slope of line or plane is controlled by weights in a linear function e.g

$$\hat{y} = m * x + b$$

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + \epsilon$$

- Use gradient descent to find optimal weights that cause it to fit the data.
- we experiment with the **regularizer** to control function complexity, help the model generalize

Model evaluation - Metrics

- Baseline

```
Baseline Accuracy, Just Predict the Median
Mean Absolute Error: 243.035
Mean Squared Error: 164669.211
Root Mean Squared Error: 405.795
Coefficient of determination: %.2f -0.484
```

- Check Error against test set

	MAE	MSE	RMSE	R2_Score
Linear Regression	109.911	32994.00	181.6425	0.7026
Ridge	109.907	32995.44	181.6465	0.7026
Lasso	109.555	33523.24	183.0935	0.6979

- Even after tuning hyperparameters, determining best model is difficult

More Regressors: KNN, DT, Random forest

- KNN Regressor: take some average of neighbours
- Decision Tree Regressor: the previous algorithms reconfigured for regression
- Random Forest: Ensemble, or group of multiple DT Regressors, their average or mean prediction is taken
- Each tree is built using different parts of the dataset, so that the error in each tree is relatively uncorrelated

Model evaluation - Metrics

- KNN (Distance, n=5)
- DTree and RF (Depth=23)

	MAE	MSE	RMSE	R2_Score
KNN	281.637	186489.15	431.8439	-0.6808
DTree	129.384	56211.30	237.0892	0.4934
RandomForest	108.795	31439.26	177.3112	0.7166
Ridge	109.907	32995.44	181.6465	0.7026

- Random forest has lowest error across all

MLPL: Delivery and Acceptance

Regression is most natural formulation

- QuAm to predict plant AC_Power output 24 hours from prediction.
- Final QuAm is a Random Forest Regressor, lowest error

Could possibly choose DAILY_YIELD in 24 hours, but predicting the sum of the full distribution(I.E **nominal daily yield**)

Keep possible actions in mind:

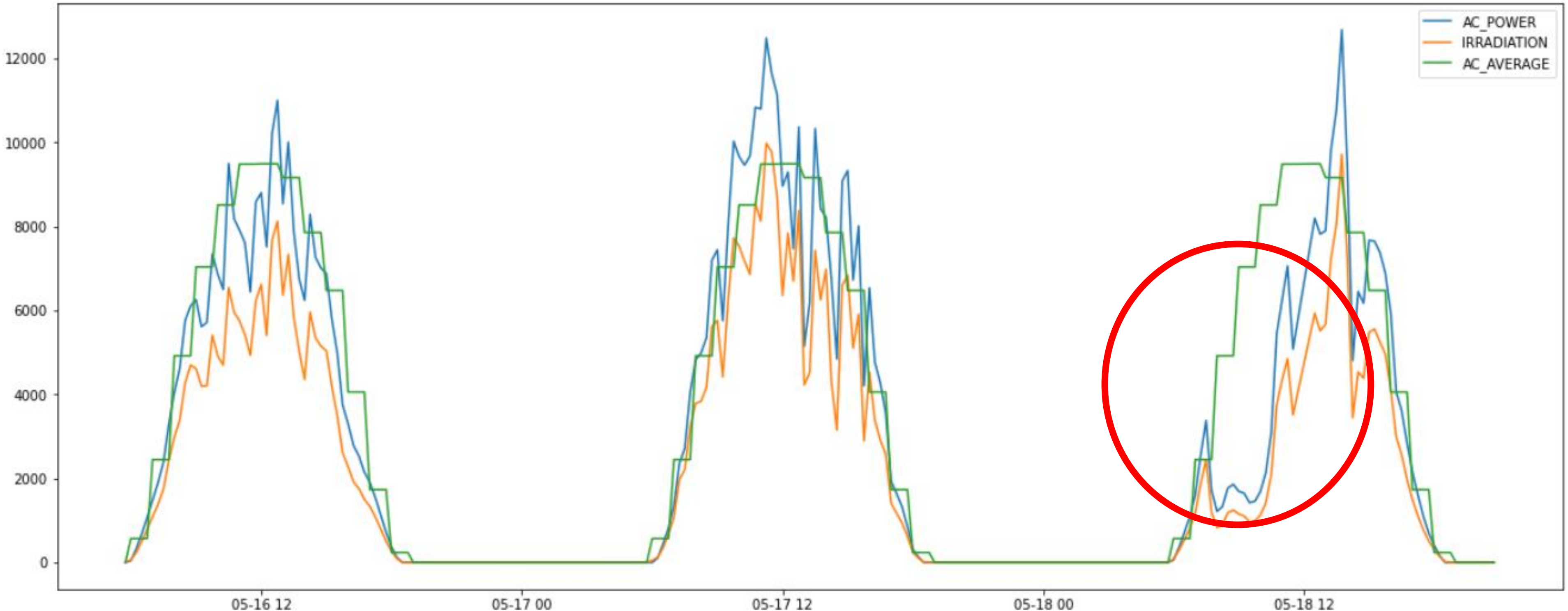
Release/Store Power, Sell/Buy Power, perform maintenance

Mean Absolute Error: 243.03454488265032

Mean Absolute Error: 243.03454488265032

MLPL: Limitations -> Future Work?

- Green is average for each hour, each day. Baseline average prediction seems unusually accurate
- Distribution is noisy, but similar most days
- Reformulating as **Anomaly prediction** may also be useful, depending on client application.



MLPL: Future Work

Additional Data Examples:

- This data is limited to 34 days in May and June, more data, to justify rest of the year, do more time series analysis
- Public Solar Angle Data, and Air pollution data found, but not implemented due to time, practical?
- Would like more site specific data gathered, more weather (humidity?)

MLPL: Future Work

Following up on MLPL:

- Would like to follow up with stakeholder power plant operators, specific actions taken to focus target (I.E energy trading, predict for storage, panel maintenance scheduling etc.)
- How workers will use the model, evaluate model drift, retraining (Climate Change?)
- Ultimately would like to work with client to go farther (Next Cycle of MLPL).

Thanks for Listening Everyone!

Acknowledgements:

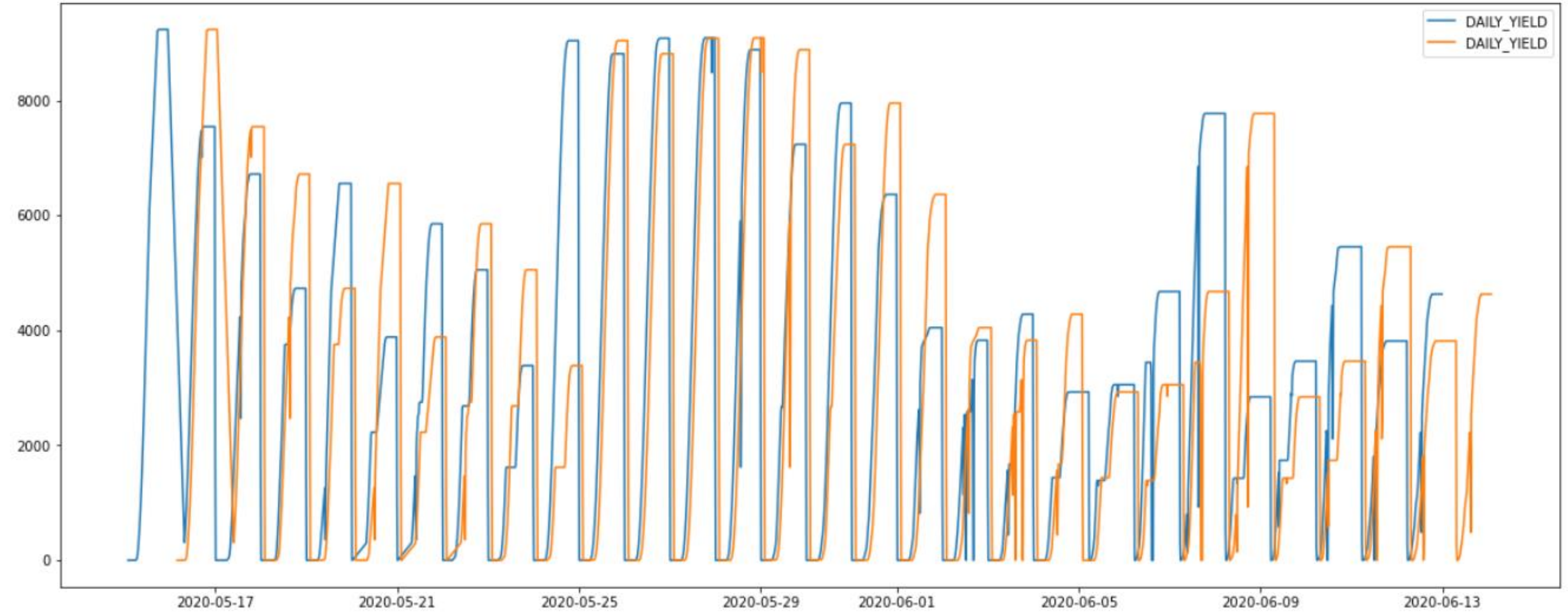
Thanks to everyone that helped us, our instructors Blanca, Mohammad, and Omid, our classmates, and Amii.

Additional Slides

Extra: Data Coherence Check

- Manually checked offset data is intact

Graph of DAILY_YIELD and DAILY_YIELD_TOMMOROW



	DATE_TIME	SOURCE_KEY	AC_POWER	PLANT_ID	AC_ONE_DAY
20	2020-05-15 07:00:00	1BY6WEcLGh8j5v7	170.014286	4135001	142.285714

Label is copy of future feature.

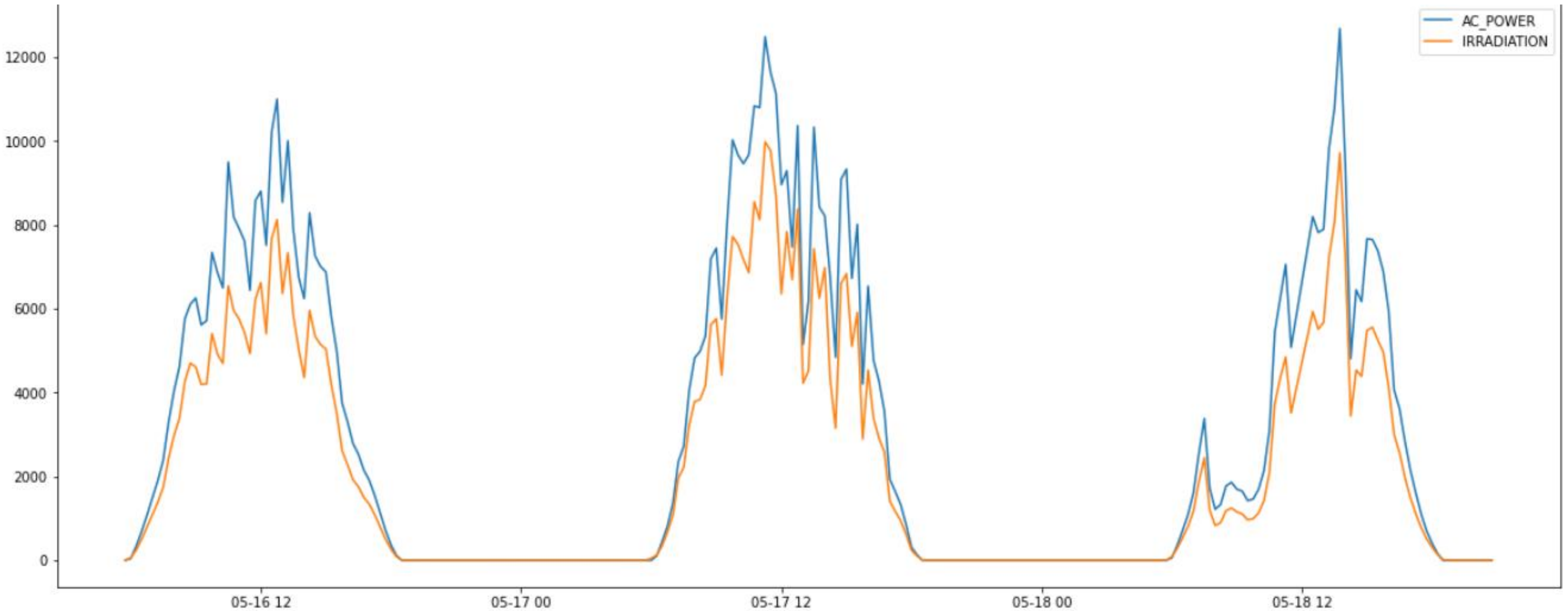
	DATE_TIME	SOURCE_KEY	AC_POWER	PLANT_ID	AC_ONE_DAY
105	2020-05-16 07:00:00	1BY6WEcLGh8j5v7	142.285714	4135001	125.071429

Validation Split: Time series, Data Leakage

- Don’t understand possible do
- Decided against a random split for final regressor.
- Afraid validation data needs to be unrelated to the training data (Experimentation seemed to confirm).
- We need Data Points that are fully separated, so that we don’t have Targets in the training data, that show up as Labels in the Validation data.

Feature Distribution Comparison

- Irradiation tracks AC_POWER closely, Is average AC_POWER day, that accurate? Baseline



Training Split options.

- randomly split, regardless
- time split across all keys, Train on first 80%, test on next 20%
- we could have done a triple split, first random validation, then test set for final

Grab the model predictions and graph them

Time series plot comparing predicted distributions

MLPL: Future Work

Go back to stakeholder with scope questions:

- Is this QuAm fine, or do they want more?
- Enough to want these new questions answered, get more data, do site specific studies, time series analysis etc.?

Extra Conclusions (Limitations)

- Target Choices
- Daily Yield is the sum of all energy generated up to that point, so basically predicts the likely distribution for the day, more difficult
- AC_POWER basically predicts what the performance would be at that moment, given possible state of the plant

Extra Limitations

- Can't make predictions after maintenance
- Predicting 5 days out will simply mean the model gives you the average solar day as an answer
- Distribution per day seems quite stable whole year, would like to study irregularities

Frameworks

- Business Framework
- Machine Learning Framework
- MLPL

Four Pillars to Move up the Spectrum

Data	Data is high quality, accessible and usable for your organization to reap long-term benefit from ML solutions.
People	Resource investment has been made in the areas of knowledge and experience.
Strategy	A cohesive ML strategy has been developed that spans across your organization's various lines of business.
Technology	Infrastructure is scalable and investment in tools and technologies allows for seamless integration of ML systems.

Extra Future Work

- Follow the MLPL Framework
- refine binning/classification formulation of prediction task
- Use/Obtain more data. Sun Angles, and Air Pollution datasets were both found, but no time to integrate
- Experiment with feature sensitivity, see how accurate weather forecasts have to be to help
- Examine feature importance for all these features together

Recommendations and Future Work

- Refine business proposal
- Work on practical use, retraining of model, and continual evaluation
- further examine dataset inconsistencies, panels going down for maintenance,

Feature and Target justification

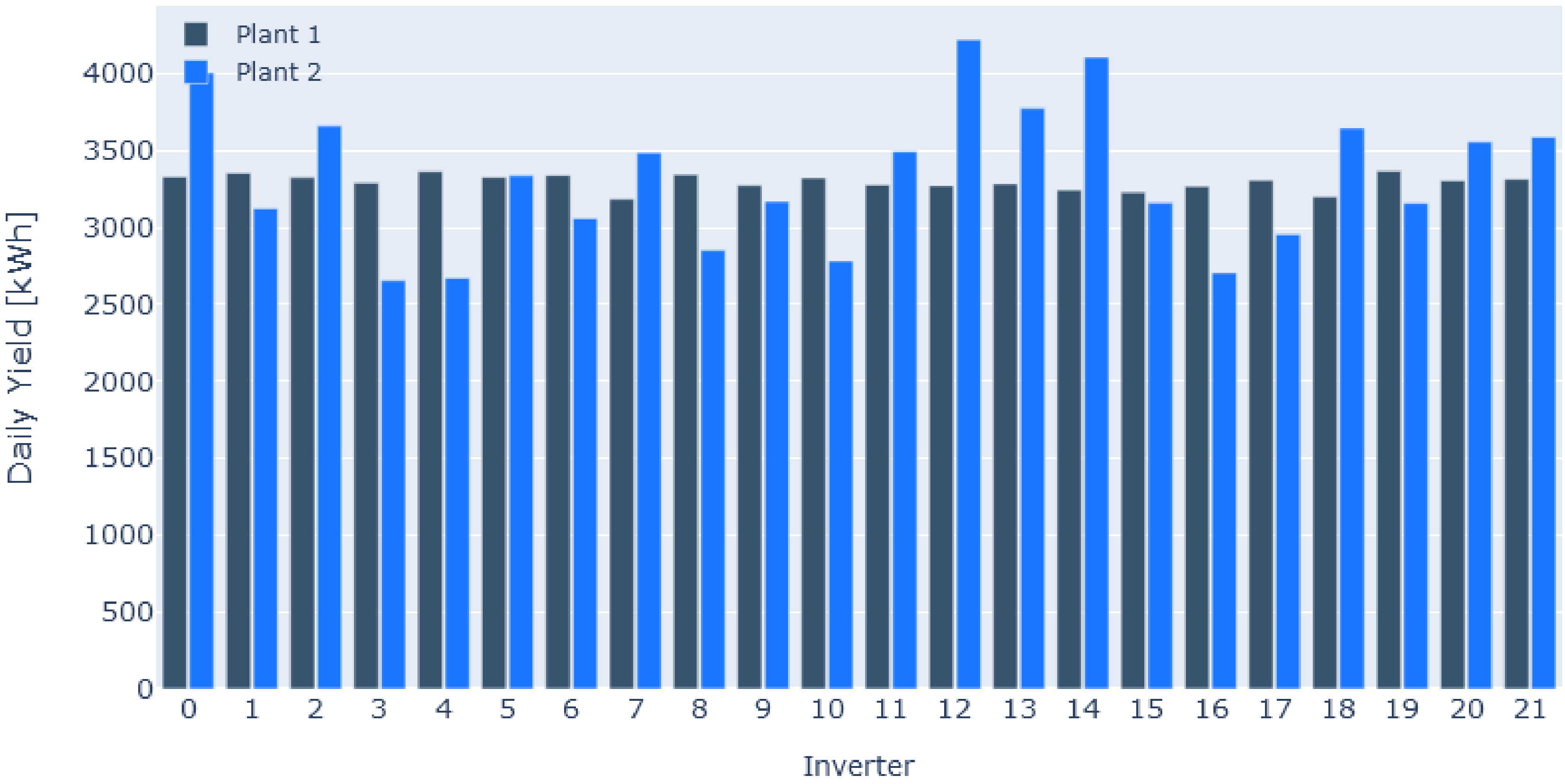
- Features: Taking in all the features from the day before could tell you something about the state of the plant.

- Target justification: We chose AC_POWER over Daily Yield
- Daily Yield tries to predict the distribution for the whole day, and so is more brittle

Data Problems

- Copied AC_POWER, TEMPERATURE, and shifted it to the next day
- Is this feature engineering, or am i causing a massive data leak, or a bit of both (or is this harmless)?
- Decided to run the models with every reasonable variation (but not messing with hyperparameters much) and compare

EDA - Inverter operation



AC_POWER VS NO AC (TIME SET SPLIT)

	MAE	MSE	RMSE	R2_Score
Linear Regression	109.911	32994.00	181.6425	0.7026
Ridge	109.907	32995.44	181.6465	0.7026
Lasso	109.555	33523.24	183.0935	0.6979

	MAE	MSE	RMSE	R2_Score
Linear Regression	142.571	41467.16	203.6349	0.6263
Ridge	142.568	41468.30	203.6377	0.6262
Lasso	142.378	41996.25	204.9299	0.6215

	MAE	MSE	RMSE	R2_Score
KNN	281.637	186489.15	431.8439	-0.6808
DTree	129.384	56211.30	237.0892	0.4934
RandomForest	108.795	31439.26	177.3112	0.7166

	MAE	MSE	RMSE	R2_Score
KNN	281.638	186489.24	431.8440	-0.6808
DTree	160.098	65863.66	256.6392	0.4064
RandomForest	138.810	40103.28	200.2580	0.6385

KNN model is quite robust, similar to baseline algorithm. It doesn’t change much after pulling data.

AC_POWER VS NO AC

(RANDOM SPLIT)

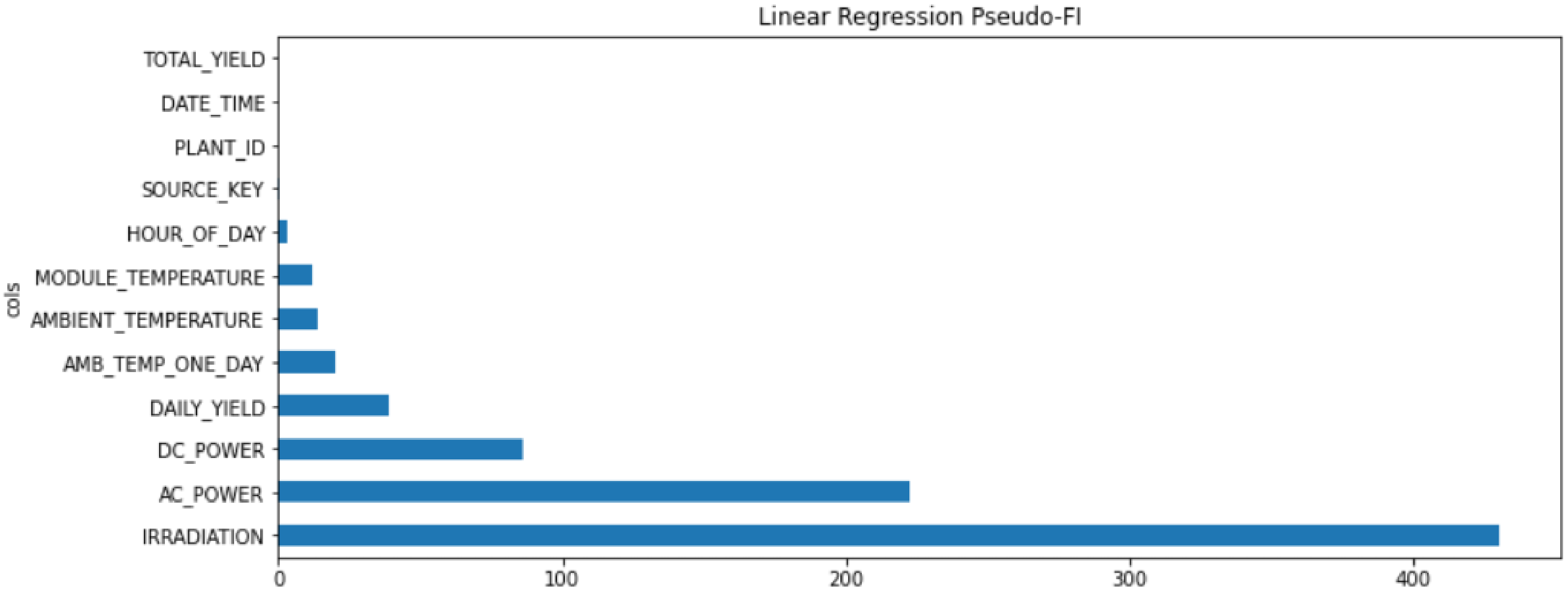
	MAE	MSE	RMSE	R2_Score
Linear Regression	132.091	47102.37	217.0308	0.6727
Ridge	132.091	47102.47	217.0310	0.6727
Lasso	132.624	47376.91	217.6624	0.6708

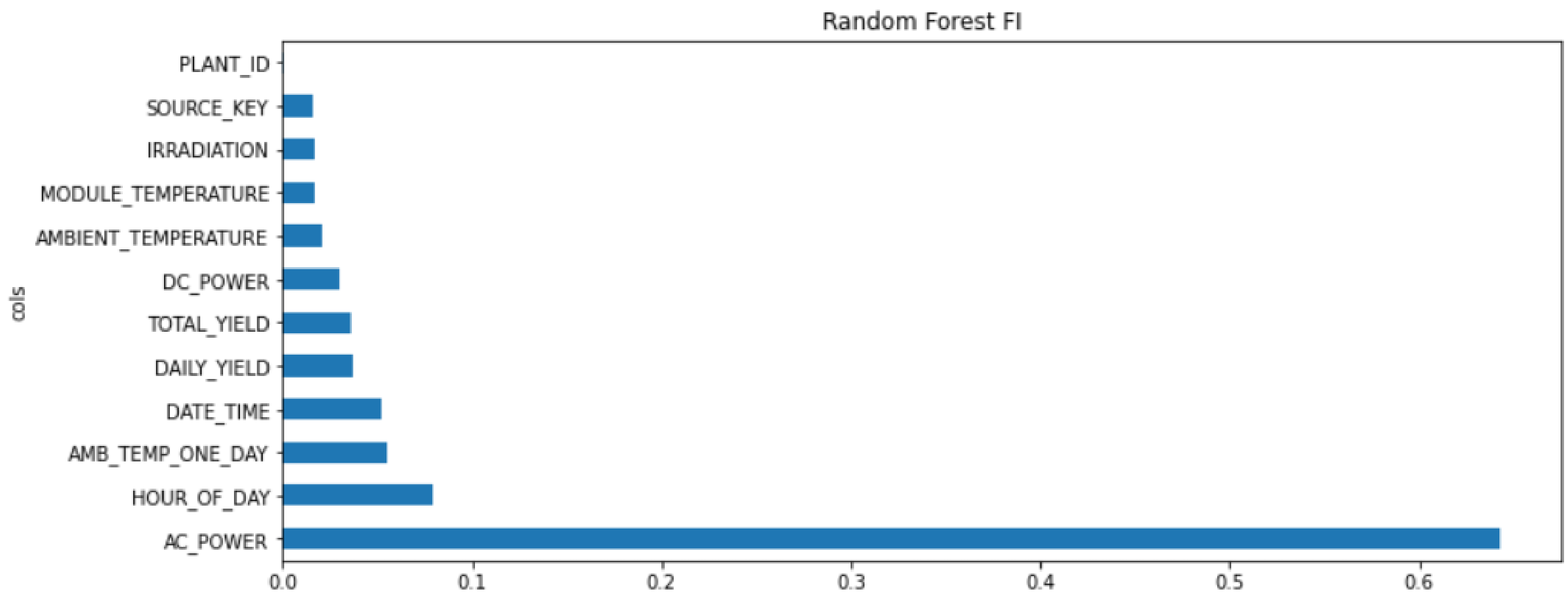
	MAE	MSE	RMSE	R2_Score
Linear Regression	170.736	61491.61	247.9750	0.5727
Ridge	170.737	61491.75	247.9753	0.5727
Lasso	171.267	61722.94	248.4410	0.5711

	MAE	MSE	RMSE	R2_Score
KNN	58.852	15244.94	123.4704	0.8941
DTree	35.699	17172.36	131.0434	0.8807
RandomForest	31.764	8873.79	94.2008	0.9383

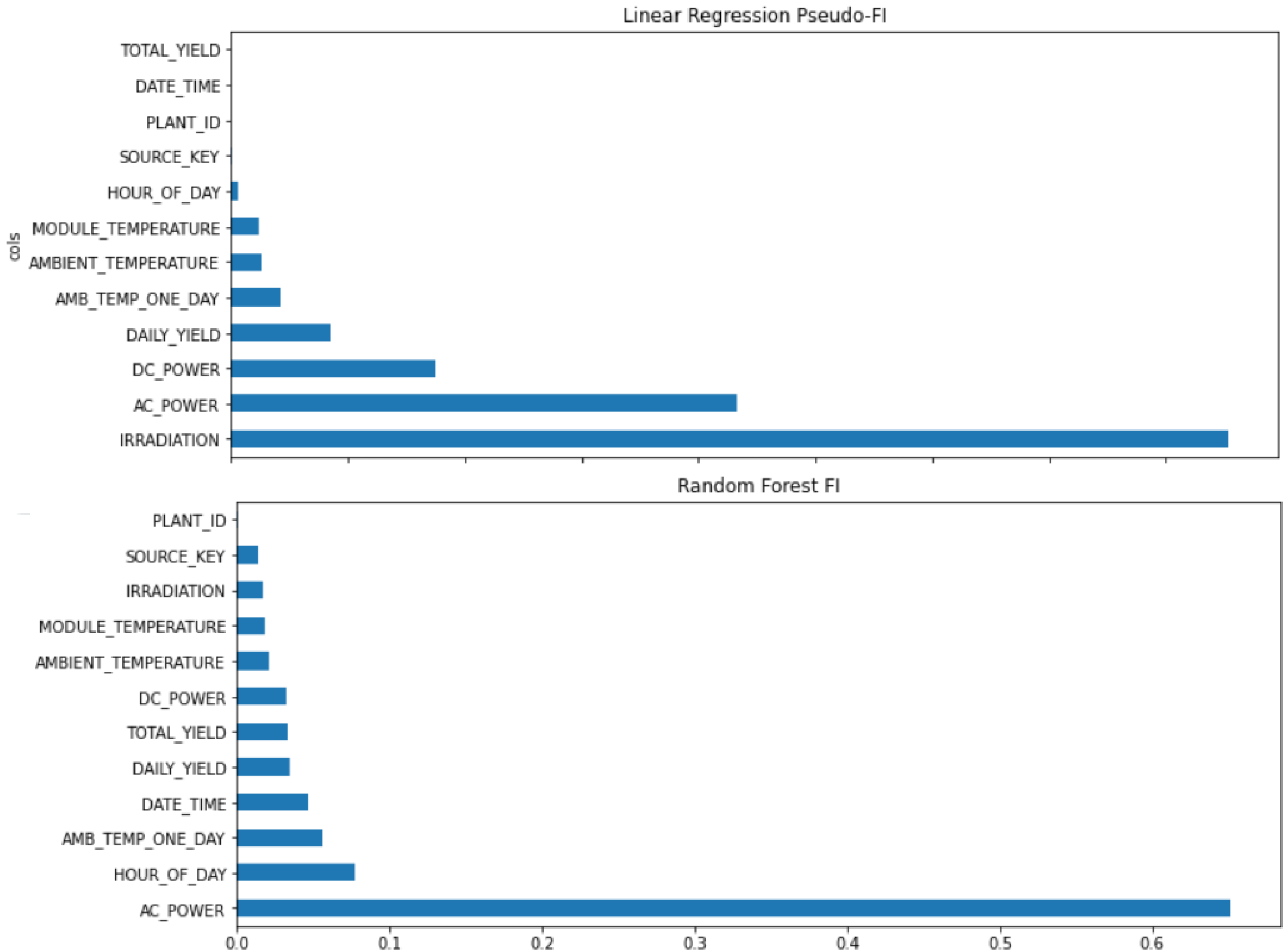
	MAE	MSE	RMSE	R2_Score
KNN	58.852	15244.94	123.4704	0.8941
DTree	34.369	15478.22	124.4115	0.8924
RandomForest	30.675	8343.10	91.3406	0.9420

Using All, Time-Split, Linear VS RF Regression FI

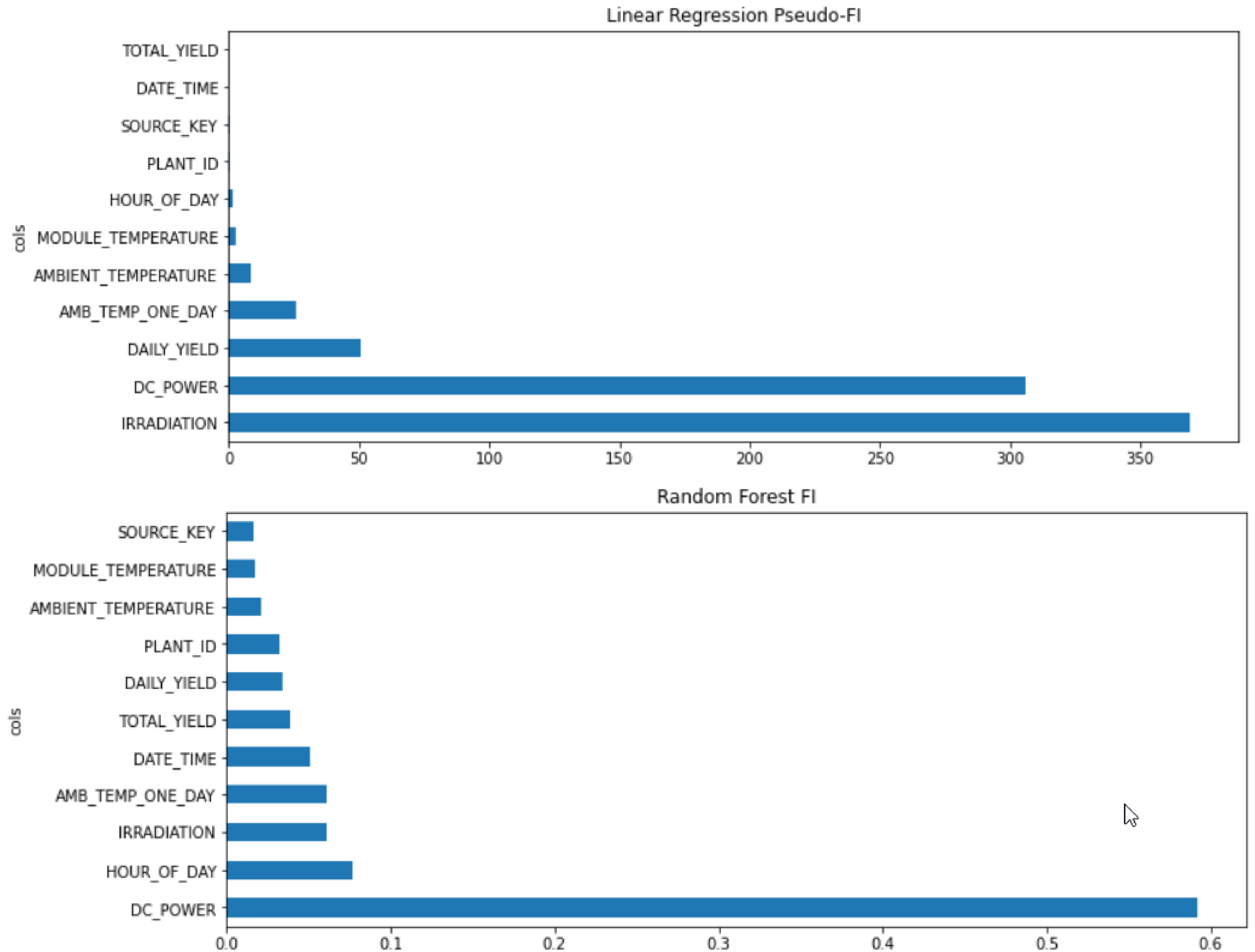




Using All, Random-Split, Linear VS RF Regression FI



Using No AC Time-Split, Linear VS RF Regression FI



Using No AC Random-Split, Linear VS RF Regression FI