Predicting Solar Power Production

W207: Applied Machine Learning

Summer 2022

Julia Bobrovskiy, Nic Brathwaite, Greg Chi, Denny Lehman, Jacob Petrisko

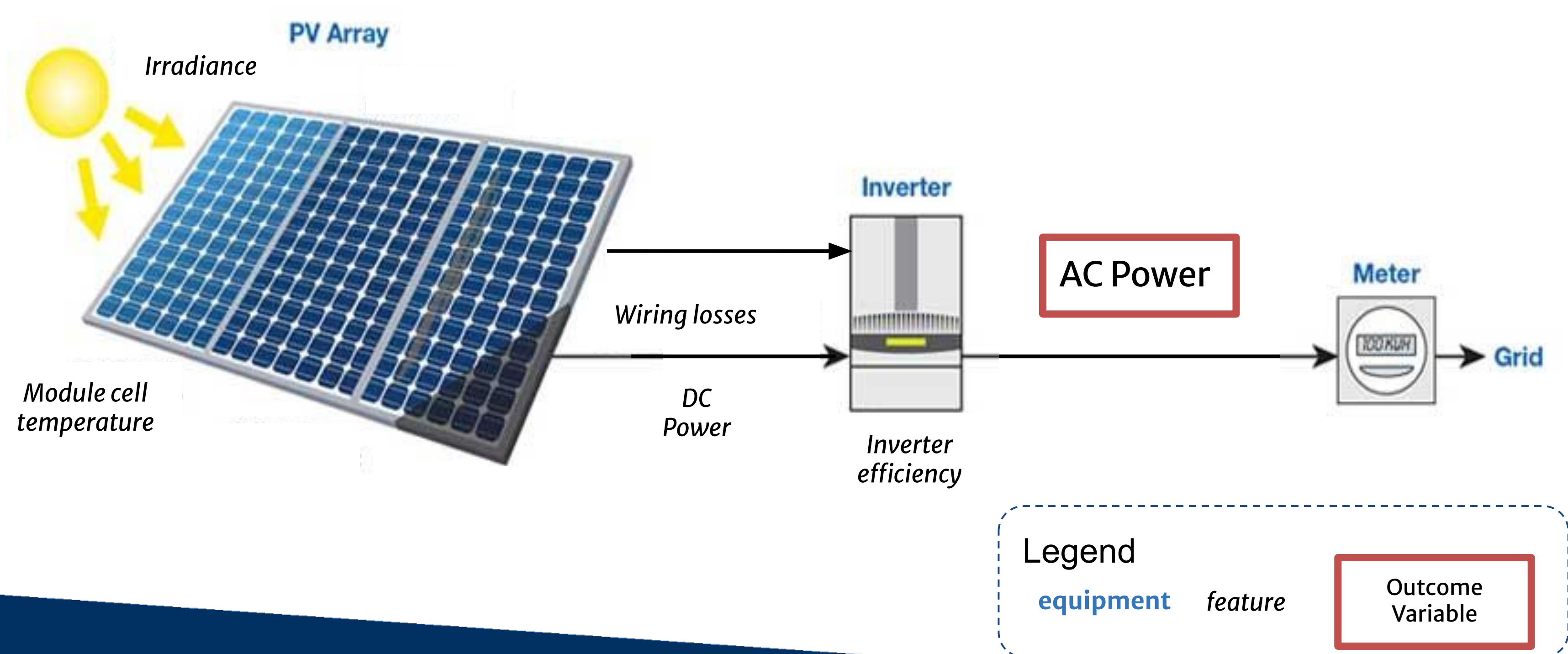
Berkeley
UNIVERSITY OF CALIFORNIA

Motivation

How does solar power work?

Key Takeaway
Irradiance is industry term for sunlight and inverters capture AC power, our outcome variable

1. Solar panel converts irradiance to electricity
2. DC power transfer
3. Inverter converter DC to AC



Berkeley
UNIVERSITY OF CALIFORNIA

img src: https://www.saurenergy.com/solar-energy-blog/here-is-how-you-can-calculate-the-annual-solar-energy-output-of-a-photovoltaic-system

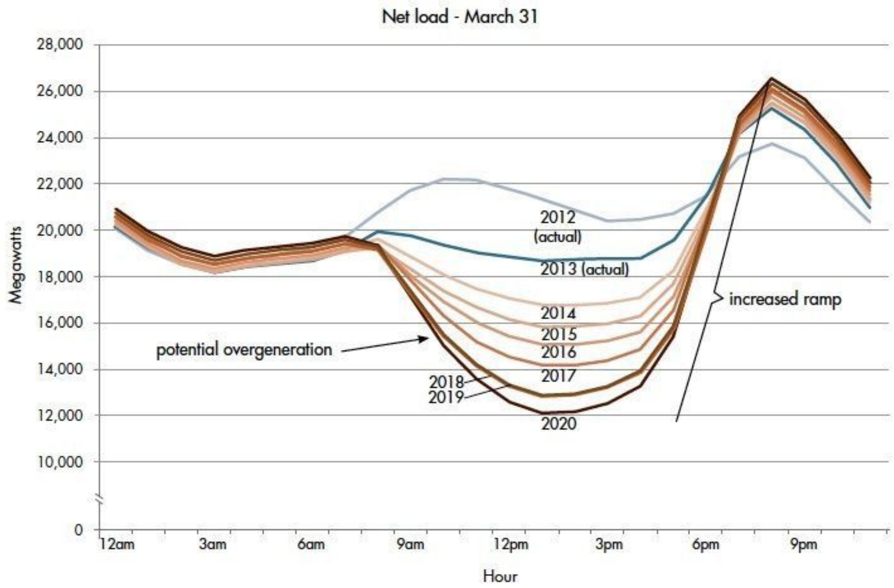
Question - The Duck Curve

Grid energy is a mix of sources, including solar and fossil fuel power plants. As solar energy becomes more prevalent, fossil fuel plants must

- Ramp down when sun rises
- Ramp up when the sun sets

Causes huge voltage variability on grid, reliability and maintenance issues.

To better address the supply and demand of California's $1.3 Trillion energy grid, we ask



Can we predict AC electricity production at a 15 min level using equipment and weather data?

Berkeley
UNIVERSITY OF CALIFORNIA

The grid reliability problem

Key Takeaway
Solar energy sloshes onto the power grid intermittently. With prediction, we can improve grid reliability

The good: Solar is growing!

The bad: intermittent power causes huge voltage variability on grid, reliability and maintenance issues.

To better address the supply and demand of California's $1.3 Trillion energy grid, we ask

Can we predict AC electricity production at a 15 min level using equipment and weather data?

Berkeley
UNIVERSITY OF CALIFORNIA

What has been done in this space

Key Takeaway
Big data and Machine Learning are revolutionizing the space, best models have ~7 normalized MAE

Big data is driving operations research thanks to

- IOT devices in solar plants
- Weather station data

Key findings

- Irradiance is best predictor, highly correlated with AC power (Feng, 2018)
- Temperature is a good predictor, inversely correlated with AC power (Feng, 2019)
- Seasonality is critical to high performing models (Boland, 2020)
- ML/AI is promising with better performance than traditional time series models (Wood, 2022)

State of the art uses multi-model machine learning to predict 1 hour ahead solar production with a normalized MAE of 6.5 (Feng, 2018)

Berkeley
UNIVERSITY OF CALIFORNIA

Our Plan

For 15 minute intervals:

Plant ID, Inverter ID, Ambient Temperature, Module Temperature, Irradiation, Time

→

AC Power

Models

- Linear Regression
- Decision Tree
- Random Forest
- Gradient Boosting Trees
- Time Series
- Neural Network

Metrics

- Mean Absolute Error
- Root Mean Squared Error

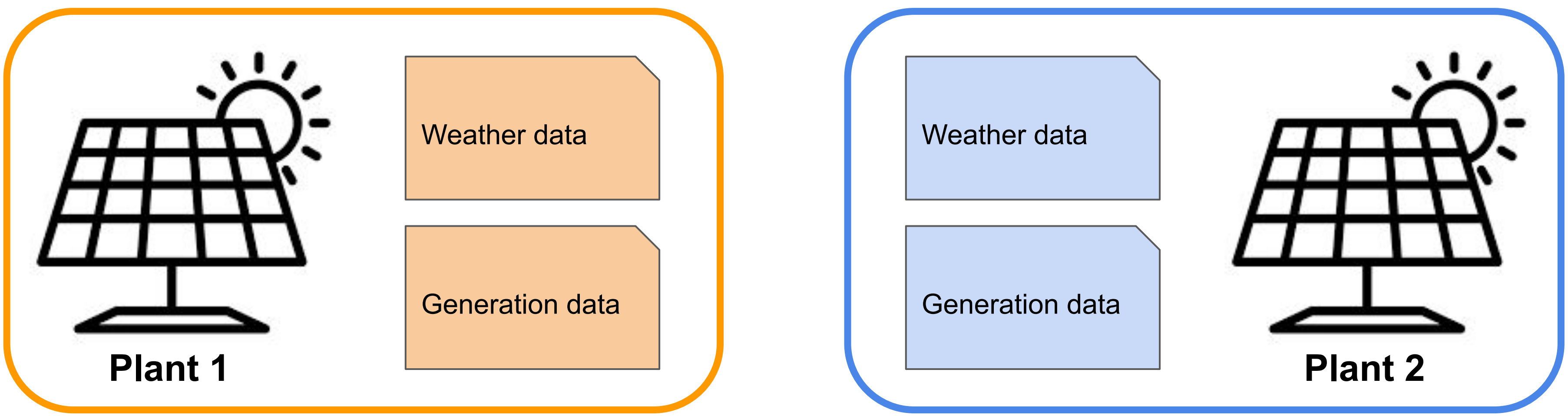
Results Summary

At 15 minute intervals, we can predict the AC power output of an inverter with a 9.5% error rate using a feed forward neural network.

Data

Data

Our data is the solar generation dataset from Kaggle. The data consists of 2 photovoltaic solar power plants in India over a 34 day period. Each plant has its own weather and electricity production data. In total, there are 4 files in the dataset, listed here:

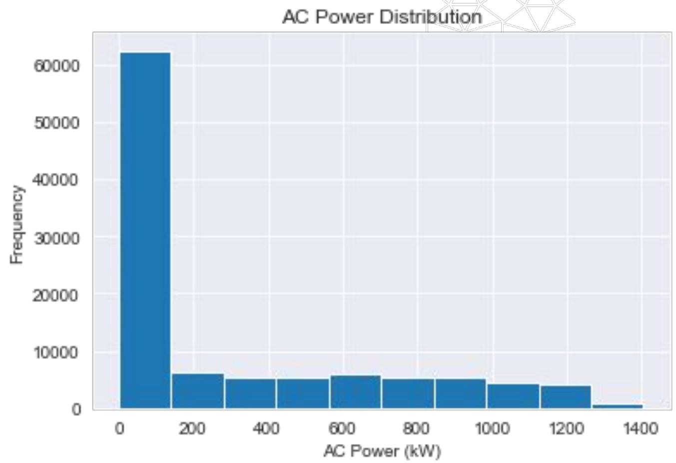


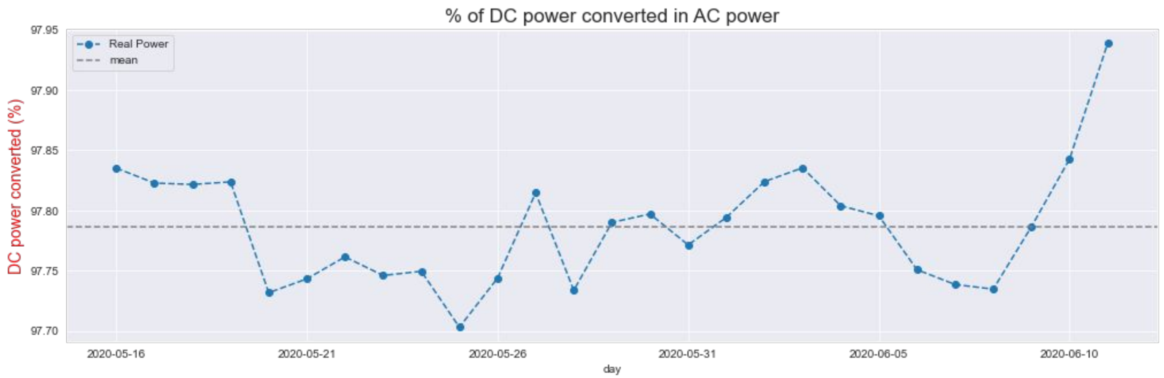
Berkeley
UNIVERSITY OF CALIFORNIA

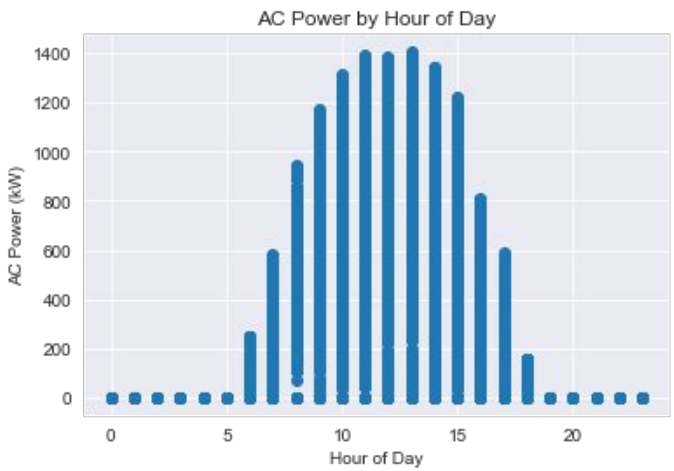
Image source: https://www.vecteezy.com/free-vector/solar-energy-icon

EDA

- **Total records:** 135k generation data from 44 inverters, 6.4k weather data observations
- **Observation Frequency:** 15 minute intervals
- **Features:** Ambient Temperature, Module Temperature, Irradiation, Weather Capture Date/Time, Inverter ID
- **Time range:** May 15, 2020 through June 11, 2020
- **Location:** India across two plants





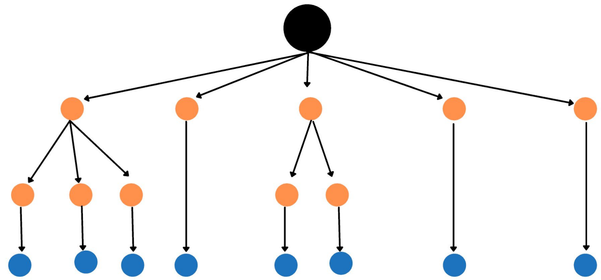


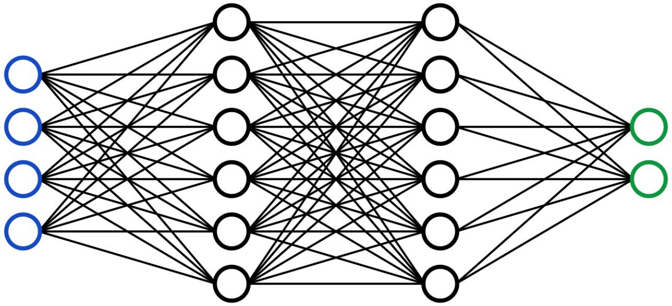
Berkeley
UNIVERSITY OF CALIFORNIA

Approach

Our Approach

- **Baseline:** Linear Regression
  - Linear regression to establish a baseline
- **Model 1:** Decision Tree
  - Improve on linear regression with a decision tree
- **Model 2:** Random Forest
  - Apply random forest to beat a simple decision tree
- **Model 3:** Gradient Boosting Decision Trees
  - Beat random forest with gradient boosting decision trees
- **Model 4:** Neural Network
  - Build a FFNN and RNN to improve upon decision trees





Berkeley
UNIVERSITY OF CALIFORNIA

Our Approach

Modeling Approach

- **Baseline:** Linear Regression
  - Linear regression to establish a baseline
- **Model 1:** Decision Tree
  - Improve on linear regression with a decision tree
- **Model 2:** Random Forest
  - Apply random forest to beat a simple decision tree
- **Model 3:** Gradient Boosting Decision Trees
  - Beat random forest with gradient boosting decision trees
- **Model 4:** Neural Network
  - Build a FFNN and an LSTM to improve upon decision trees

Evaluation Approach

Each model was evaluated against the baseline based on the following metrics:

- **Mean Absolute Error** (MAE) in kW: to establish the average absolute error between predicted AC output and actual AC output

$$MAE = \frac{1}{n}\sum |e_t|$$

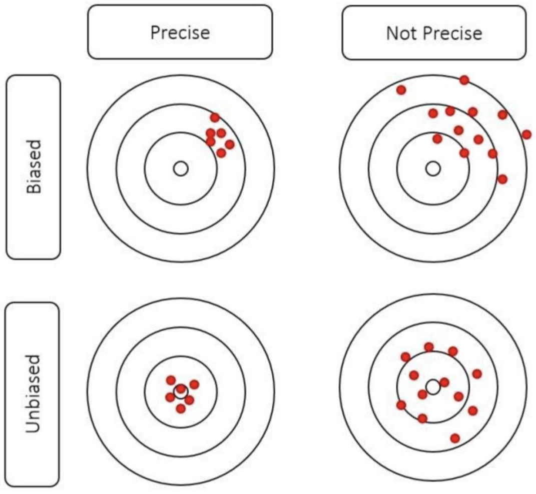
- **Normalized Mean Absolute Error (NMAE %):** the MAE divided by the average AC output

Berkeley
UNIVERSITY OF CALIFORNIA

Evaluation

- **Bias:** average error (across history)
- **Precision:** spread between forecast and actuals

e = (predicted power generation) - (actual power generation)



- **MAE:** Mean Absolute Error
  - evenly distributed, good with outliers
- **RMSE**: Root Mean Squared Error
  - correct on average, minimizing bias

$$MAE = \frac{1}{n}\sum |e_t|$$

$$RMSE = \sqrt{\frac{1}{n}\sum e_t^2}$$

Berkeley
UNIVERSITY OF CALIFORNIA

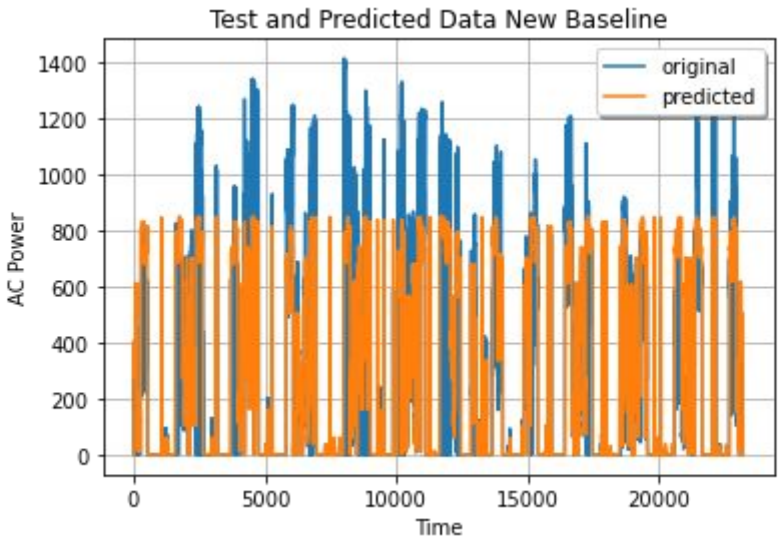
Experiments

Baseline Model

Key Takeaway
Baseline model established baseline for error rates in future model explorations

Logic for Baseline Model

The mean of the AC power generated in a given 15-minute timeframe is calculated across all inverters

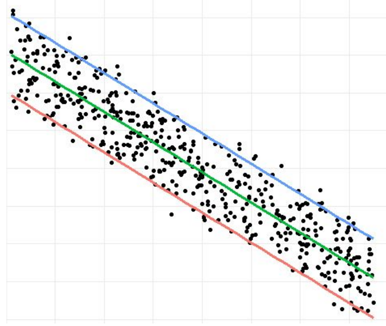


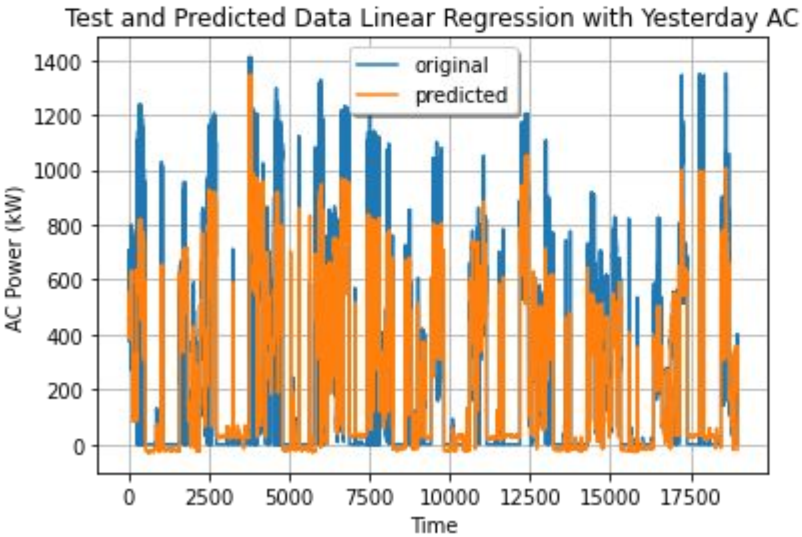
MAE = 163 kW
(35% error rate)

Berkeley
UNIVERSITY OF CALIFORNIA

Linear Regression

Key Takeaway
Multivariate linear regression saw significant improvements over our baseline.





Why use linear regression?

Intuitive explanation of model drivers and their impact on total AC power production.

Features in Lowest MAE Specification:

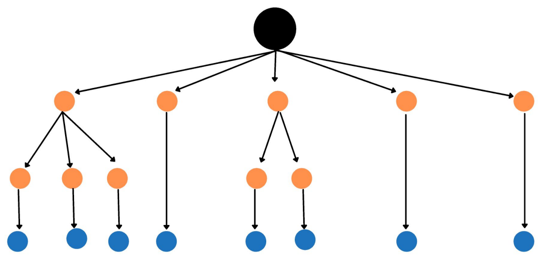
Irradiation, Ambient Temp, Module Temp, Time, AC Power 24h Ago, Time, Plant ID

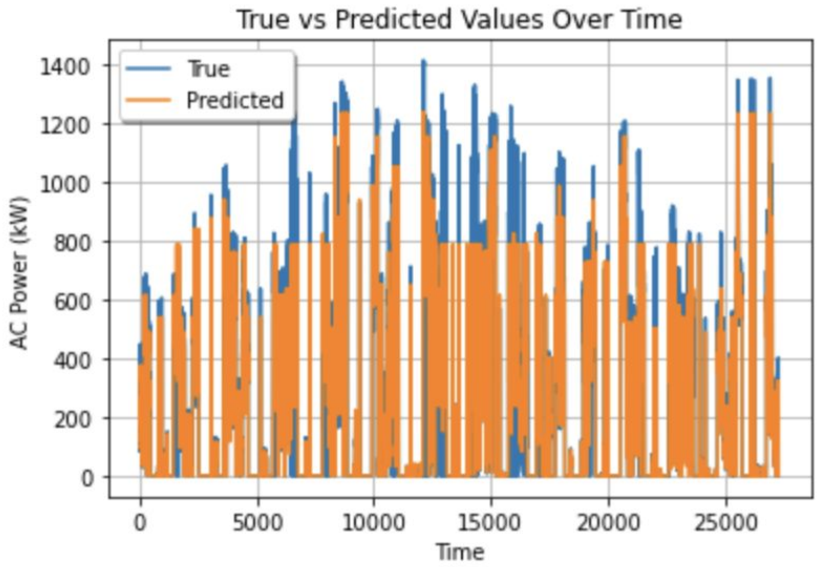
MAE = 50.4 kW
(10.9% error rate)
3.2x better than baseline

Berkeley
UNIVERSITY OF CALIFORNIA

Decision Tree

Key Takeaway
Decision tree outperforms linear regression.





Why use a decision tree?

Our data is rule-based: night or day, warm or cold, cloudy or sunny.

Optimal hyperparameter:

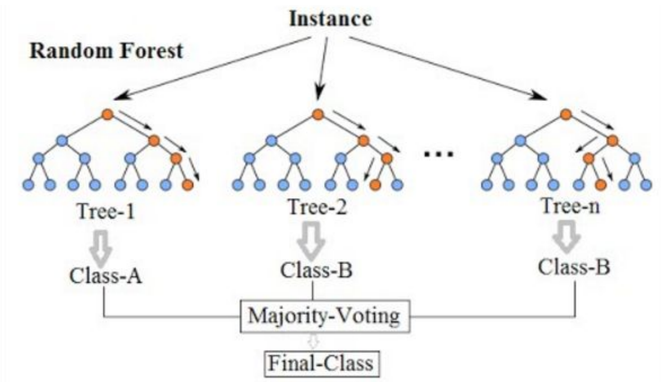
Max Depth: 7

MAE = 45.7 kW
(9.92% error rate)
3.6x better than baseline

Berkeley
UNIVERSITY OF CALIFORNIA

Random Forest

Key Takeaway
Random forest slightly underperforms a simple decision tree.



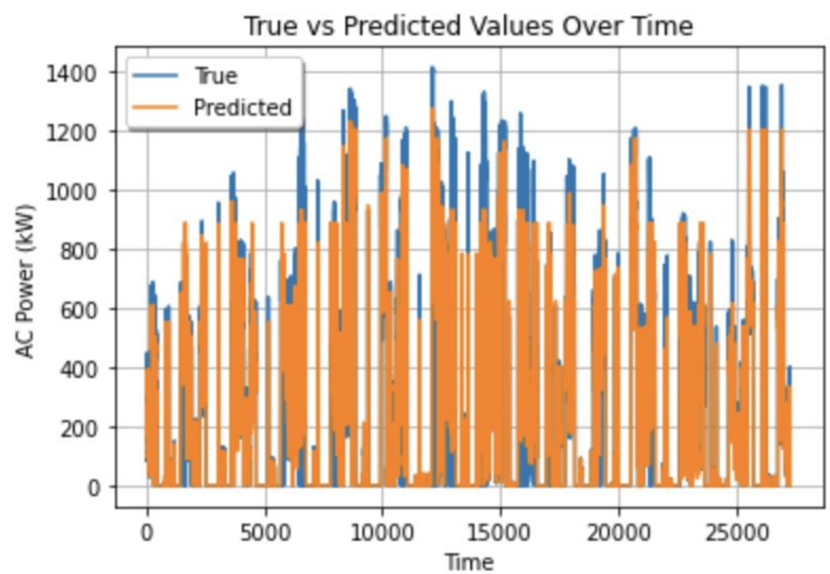
Why use a random forest?

Improve upon the simple decision tree using an ensemble.

Optimal hyperparameters:

Max Depth: 6

Num Estimators: 3



MAE = 46.2 kW

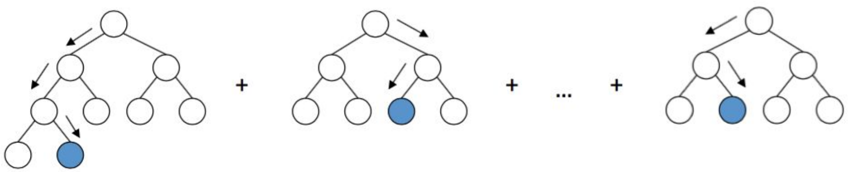
(10% error rate)

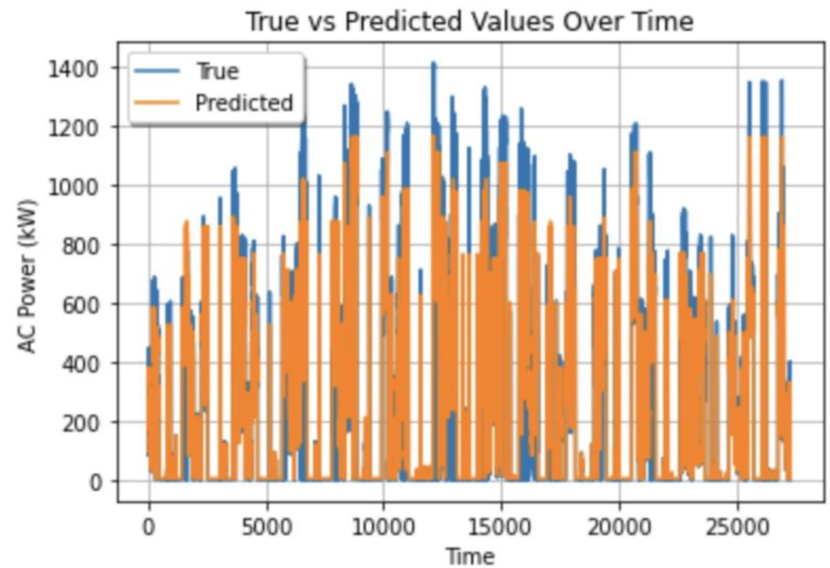
3.5x better than baseline

Berkeley
UNIVERSITY OF CALIFORNIA

Gradient Boosting Trees

Key Takeaway
Gradient boosting trees result in the same error rate as random forest.





Why use gradient boosting trees?
Apply a more advanced ensemble method to improve upon random forest.

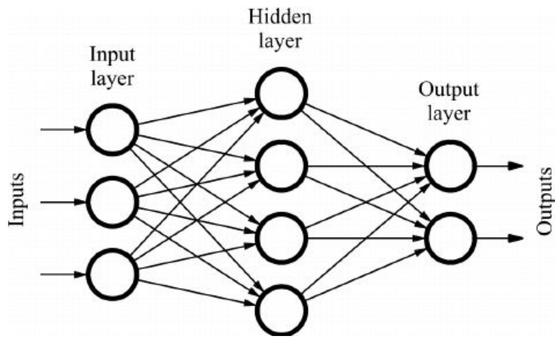
Optimal hyperparameters:
Max Depth: 5

Num Estimators: 5

MAE = 46.2 kW
(10% error rate)
3.5x better than baseline

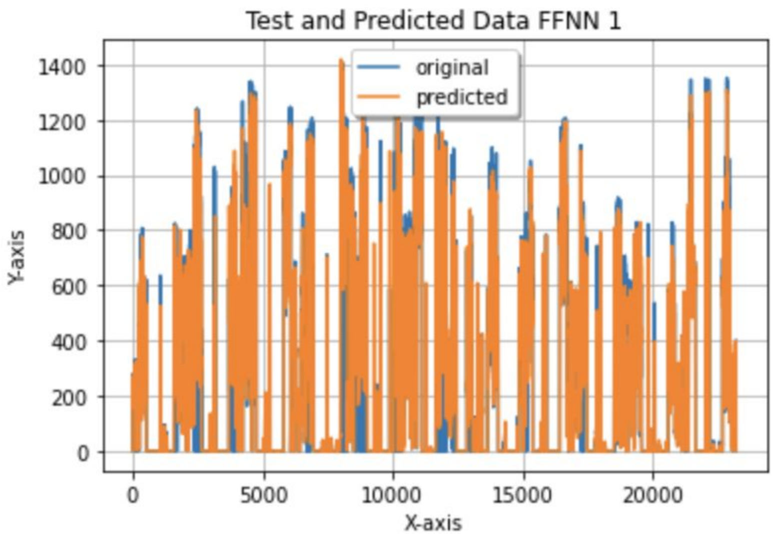
FFNN

Key Takeaway
Feed Forward Neural Networks outperform all decision tree variations.



Why use Feed Forward Neural Networks?
Apply the use of hidden layers and optimized activation functions to improve our prediction.

Optimal Hyperparameters:
Hidden Layers: 4 (ReLU)
Epochs: 20
Batch Size: 64

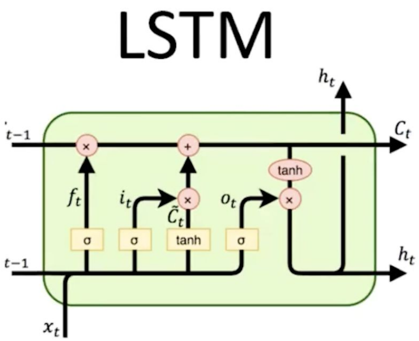


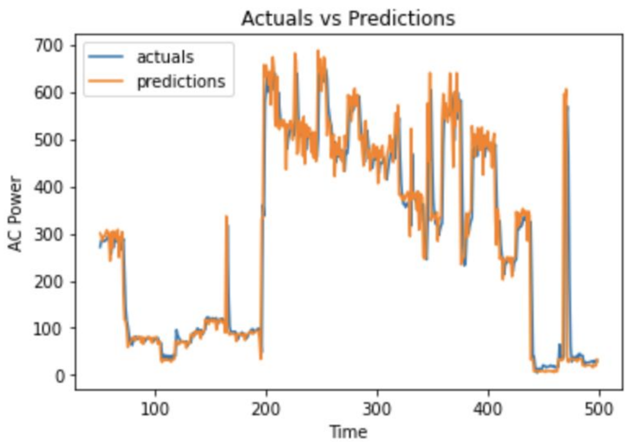
MAE = 43.5 kW
(9.5% error rate)
3.7x better than baseline

Berkeley
UNIVERSITY OF CALIFORNIA

LSTM

Key Takeaway
Time alone is more predictive than we thought for AC Power





Why use Long Short Term Memory?
Uses time series data in RNN structure, including previous outputs as inputs to the next node.

Optimal Hyperparameters:
Window Size: 8
Epochs: 200
LR: 0.0001

MAE = 52 kW

(11% error rate)

3.1x better than baseline

Berkeley
UNIVERSITY OF CALIFORNIA

Conclusion

Results

	MAE	RMSE	NMAE
FFNN	42.255606	91.632888	9.17
Decision Tree	45.699147	95.919128	9.92
Gradient Boosting	46.183677	91.837501	10.02
Random Forest	46.224350	94.543866	10.03
Linear Regression Ridge	50.394521	96.295946	10.93
Linear Regression	50.404929	96.293097	10.94
Linear Regression Lasso	50.576399	96.423121	10.97
LSTM	52.886768	109.996044	11.47
Advanced Baseline	163.035980	227.684901	35.37
Simple Baseline	286.158529	339.658231	62.09

Lessons Learned

- Decision trees can perform similar to neural networks
- Times series data may not always require a time series model
- Limitations:
  - Not applicable outside of 34 days in summer, India, and 2 plants
  - Likely overfit to summer months

Berkeley
UNIVERSITY OF CALIFORNIA

Future Work

- **Geographic expansion** – Expand to US plants
- **Seasonality** – get more data to show seasonal effects
- **Application** – Use to predict day ahead solar power supply in CA
- **Other Methods** – Try time series models, like ARIMA
- **Other Outcomes** – Consider variance predictions (e.g. GARCH) for scenario analysis (“What’s the worst that could happen?”)
- **Analytical pipeline** – Chain prediction model to an optimization model to optimally meet power demand
- Collect several years worth of data rather than several months

Questions?

Berkeley
UNIVERSITY OF CALIFORNIA

Code

https://github.com/denny-lehman/power_production_w207_final_project_2022

Berkeley
UNIVERSITY OF CALIFORNIA

Contributions

Julia: pre-processing, EDA, LSTM model 1, LSTM model 2, slides

Nic: EDA, domain research, feature engineering, hyperparameter tuning, slides

Greg: EDA, feature engineering, linear regression iterations, meeting mgmt, slides

Denny: domain research, baseline model, preprocessing, eda, project mgmt, slides

Jacob: EDA, pre-processing, linear regressions, decision trees, random forests, gradient boosting trees, FFNNs, slides

Berkeley
UNIVERSITY OF CALIFORNIA