Solar Intensity Estimation using Time-Series weather data

Machine Learning Project

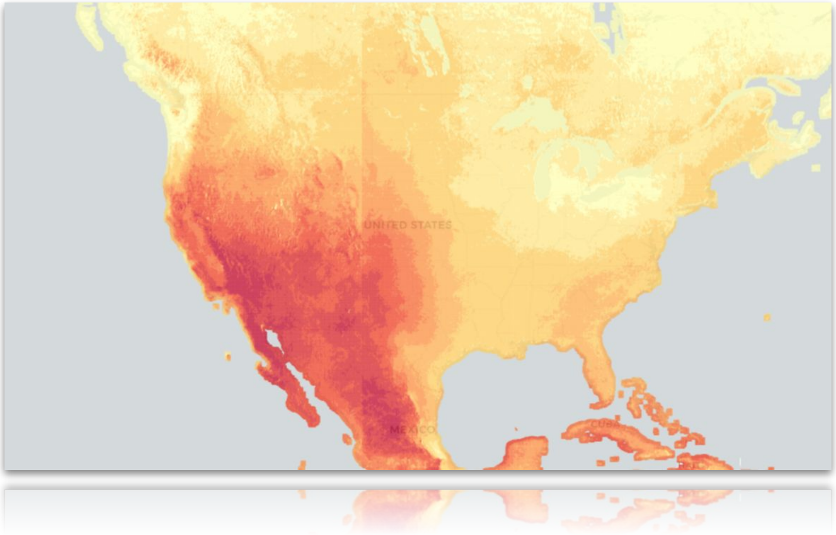
Rishi Singhal
Kushal Junejal
Naval Kumar Shukla
Udit Narang

Team Number - T3

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Motivation

- <u>Goal</u> : Provide high-confidence forecasts of solar power generation using readily available weather data.



Solar Power Intensity over North America [NSRDB]

Motivation

- Solar energy is inexhaustible energy resource.
- Solar panels only capture small amounts of energy from the sun.
- Highly variational nature of solar energy puts strain on conventional fossil fuel based energy sources.
- We aim to predict solar intensity 48 hours into the future using time-series weather data.
- This will enable better regulation of fossil fuel based power generation.
- The project aligns with our sustainable development goals to apply ML models to solve real world problems.

Solar Panel [Wikimedia Commons]

Literature Review

- Prediction models for smart homes for advanced planning of electricity consumption
- Compared Least Linear square regression and SVM considering the correlation between various weather observations

[1] N. Sharma, P. Sharma, D. Irwin and P. Shenoy, "Predicting solar generation from weather forecasts using machine learning," 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm), 2011, pp. 528-533, doi: 10.1109/SmartGridComm.2011.6102379.

Solar Power Plant [Wikimedia Commons]

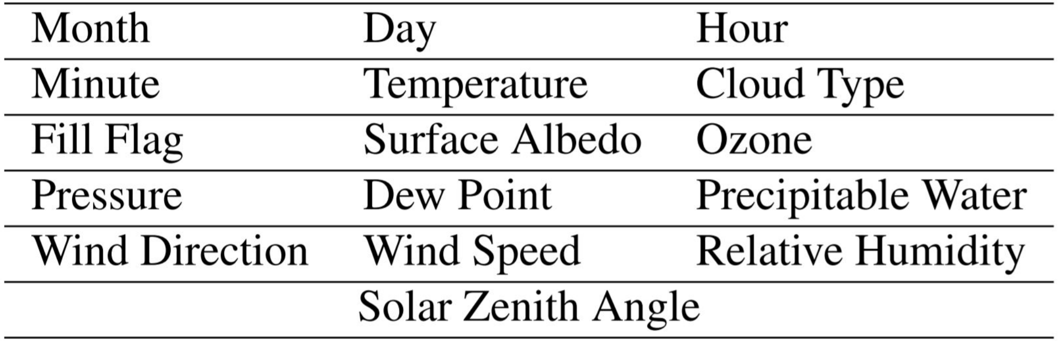
Literature Review

- Forecasting renewable resources to benefit power systems
- Then the paper has focused on ANN. To select the most effective features they used - correlation & sensitivity analysis.
- After obtaining the best features, they have used different models like ANN, MLR & persistence forecasts model and compared their performance.
- The paper has also mentioned that removing night hours from the initial dataset has helped to slightly improve the model performance.

[2] M. Abuella and B. Chowdhury, "Solar power forecasting using artificial neural networks," 2015 North American Power Symposium (NAPS), 2015, pp. 1-5, doi: 10.1109/NAPS.2015.7335176.

Dataset Description

- Task : Given weather and time features, predict solar intensity.
- Input Features : Weather & Time

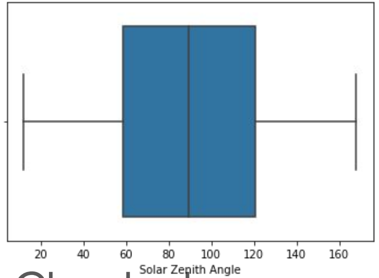


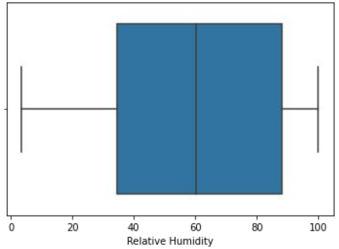
Month	Day	Hour
Minute	Temperature	Cloud Type
Fill Flag	Surface Albedo	Ozone
Pressure	Dew Point	Precipitable Water
Wind Direction	Wind Speed	Relative Humidity
Solar Zenith Angle		

- Numerical Output : Solar Intensity (‘GHI’ in watts per square meter)

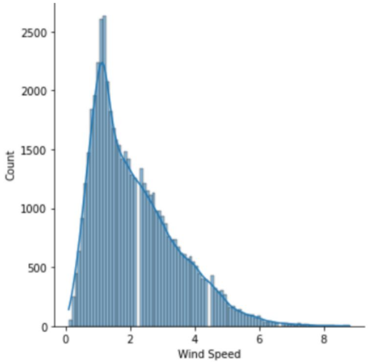
EDA

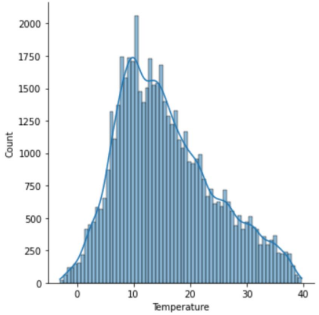
- Outlier detection using box plots



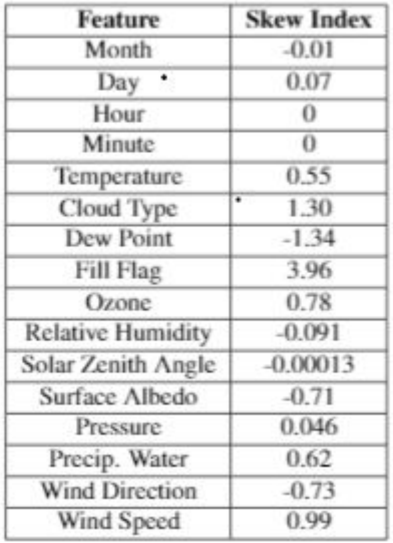


- Check skewness using histograms





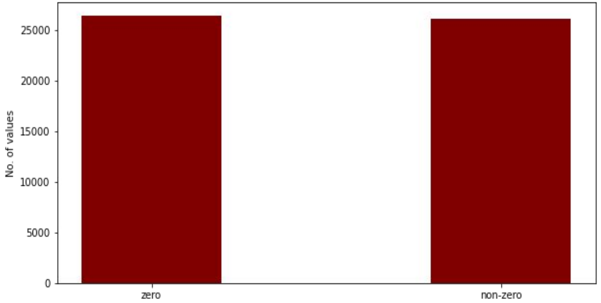
Skew Index for each feature

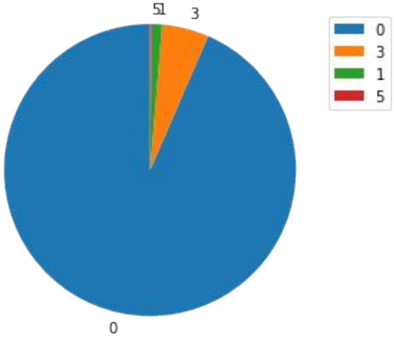


Feature	Skew Index
Month	-0.01
Day	0.07
Hour	0
Minute	0
Temperature	0.55
Cloud Type	1.30
Dew Point	-1.34
Fill Flag	3.96
Ozone	0.78
Relative Humidity	-0.091
Solar Zenith Angle	-0.00013
Surface Albedo	-0.71
Pressure	0.046
Precip. Water	0.62
Wind Direction	-0.73
Wind Speed	0.99

EDA

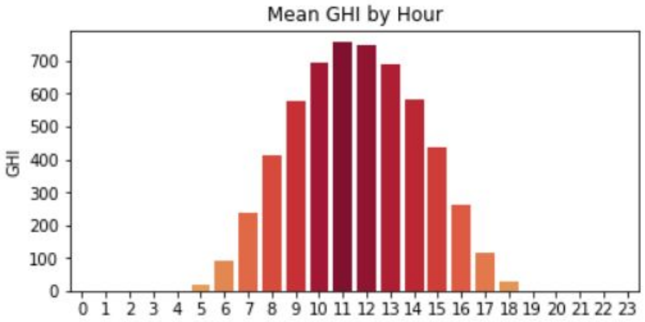
- Depicting Sparse Dataset & Fill Flag Distribution

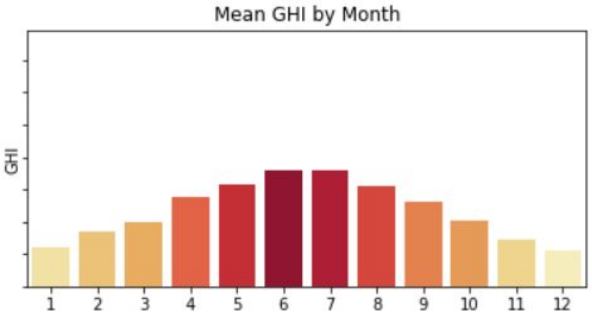


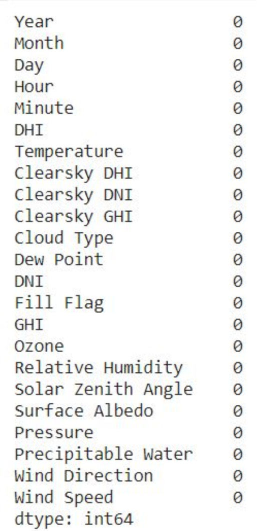


'N/A': 0, 'Missing Image': 1, 'Low Irradiance': 2, 'Exceeds Clearsky': 3, 'Missing Cloud Properties': 4, 'Rayleigh Violation': 5

- Mean GHI by Hour & Mean GHI by month & NaN values



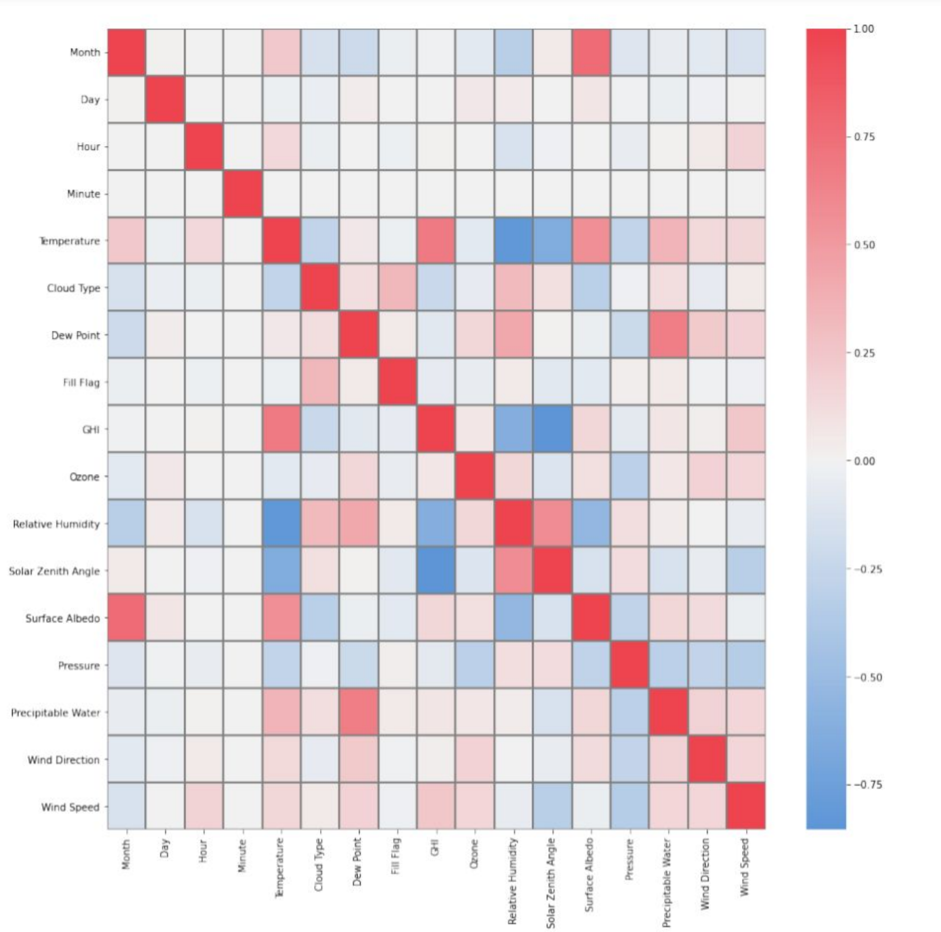




```
Year                  0
Month                 0
Day                   0
Hour                  0
Minute                0
DHI                   0
Temperature           0
Clearsky DHI          0
Clearsky DNI          0
Clearsky GHI          0
Cloud Type            0
Dew Point             0
DNI                   0
Fill Flag             0
GHI                   0
Ozone                 0
Relative Humidity     0
Solar Zenith Angle    0
Surface Albedo        0
Pressure              0
Precipitable Water    0
Wind Direction        0
Wind Speed            0
dtype: int64
```

Data Preprocessing

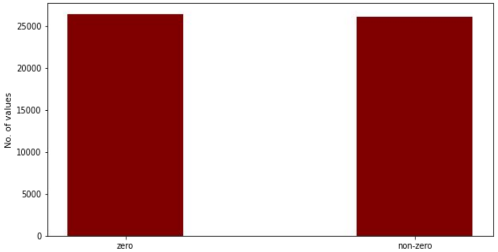
- Based On Correlation
- Based on Skewness
- Outlier Detection.
- Remove NaN
- Features that do not add any information
- 48-Hour Mapping of the dataset
- Data Normalization



Methodology

- Split dataset into 80:20 train test split.
- We use RMSE scores to measure performance of prediction models

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$



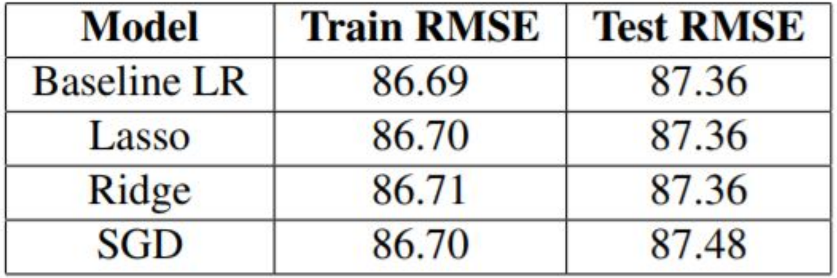
- **NOTE** : We have a sparse dataset with around 50% GHI values equal to 0.
- The 0 GHI values are corresponding to the time before sunrise and after sunset.
- We train our model using the dataset corresponding to the non-zero GHI values only.

Baseline Models (Methodology)

- Apply techniques :
  - Linear Regression
  - Lasso Regression
  - Ridge Regression
  - SGD Regressor
  - Polynomial Regression (n = 2, 3, 4 and 5)
- Compare Train-Test RMSE of different models.

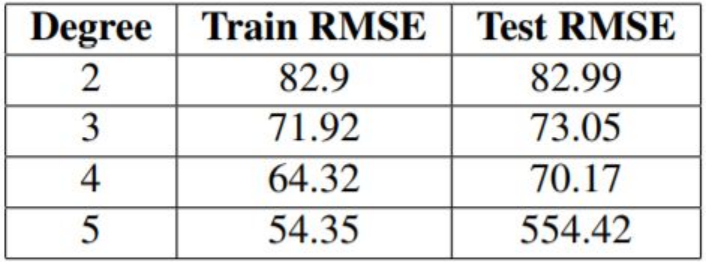
Results & Analysis(Bias - Variance Tradeoff)

(A) Without **Feature Expansion**



Model	Train RMSE	Test RMSE
Baseline LR	86.69	87.36
Lasso	86.70	87.36
Ridge	86.71	87.36
SGD	86.70	87.48

(B) **Polynomial Regression** Without **Feature Expansion**



Degree	Train RMSE	Test RMSE
2	82.9	82.99
3	71.92	73.05
4	64.32	70.17
5	54.35	554.42

Polynomial Regression reduced both the train & test RMSE without causing much overfitting till n=3, depicting that underfitting was occurring in our initial models. At degree 5 there was substantial overfitting as our model was too complex. So, we get our best performance for degree = 3 in polynomial regression case.

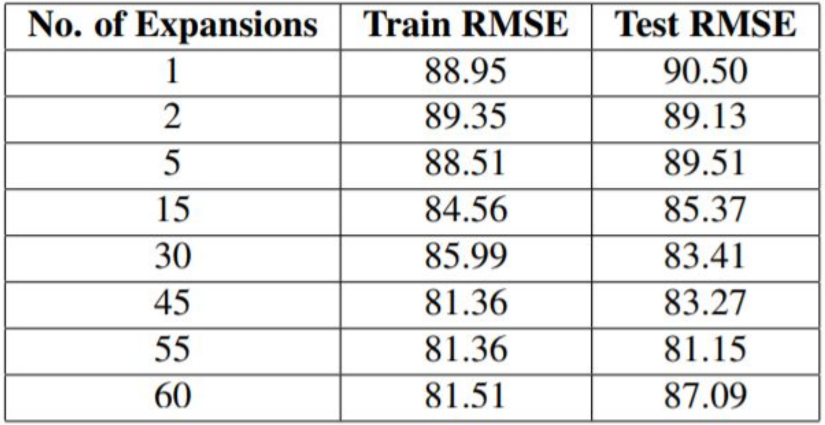
Feature Expansion (Methodology)

- Use data from previous time points as new features.

- Apply techniques :
  - Linear Regression
    - Without PCA
    - With PCA
- Principal Component Analysis (PCA) to reduce dimensions

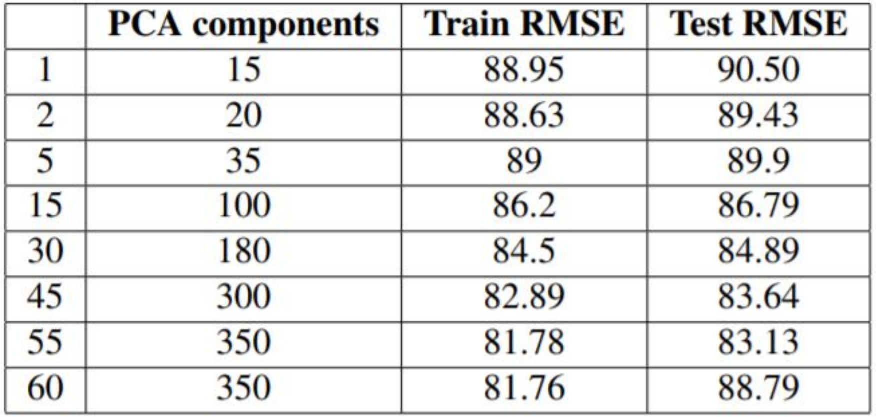
Results & Analysis

<u>(C) **Feature Expansion** w/o **PCA**</u>



No. of Expansions	Train RMSE	Test RMSE
1	88.95	90.50
2	89.35	89.13
5	88.51	89.51
15	84.56	85.37
30	85.99	83.41
45	81.36	83.27
55	81.36	81.15
60	81.51	87.09

<u>(D) **Feature Expansion** with **PCA**</u>

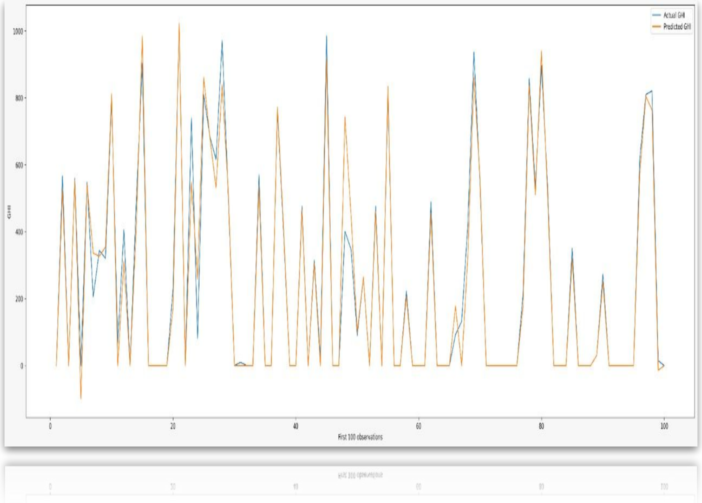


	PCA components	Train RMSE	Test RMSE
1	15	88.95	90.50
2	20	88.63	89.43
5	35	89	89.9
15	100	86.2	86.79
30	180	84.5	84.89
45	300	82.89	83.64
55	350	81.78	83.13
60	350	81.76	88.79

Feature expansion improved the performance as we went on increasing the no. of expansions. However, it made the model complex by increasing the dimensions. To counter the issue, we used PCA which reduced the dimensions without compromising on the performance. **NOTE:** As we increased the expansions by 60, we observed an increase in the variance due to overfitting.

Conclusion

- Baseline models show similar results
    - Linear Regression does NOT overfit
- Polynomial regression shows best result at n = 3. (n means degree)
    - Overfitting from n = 4
- Feature expansion without PCA improves over baseline LR till 55 feature expansions
- Applying PCA after feature expansion improves performance further by
    - Reducing model complexity ensuring minimal information loss while reducing the dimensionality of the dataset.



Polynomial Regression without feature expansion at n=3 shows best performance among all the models that we have used so far.

Timeline

- We were able to follow our timeline mentioned in the proposal.

What we have done?	What we are planning to do?
EDA and Data Preprocessing (Finished on 20th September)	Regression Trees (by 24th Oct)
Baseline Models - LR / Ridge / Lasso / SGD without feature expansion. (Finished on 4th Oct)	SVM Regression (by 30th Oct)
Feature Expansion of our dataset and applied LR models; PCA on feature expansion and applying LR models (Finished on 7th Oct)	K Means Clustering (by 6th Nov)
Polynomial Regression (Finished on 10th Oct)	Study application of various advanced methods to improve our RMSE scores (by 12th Nov)
Final Report (Finished on 12th Oct)	Final Report (by 20th Nov)

Member Contribution

Rishi Singhal	● EDA ● Polynomial Regression ● Feature Expansion ● SGD Regressor
Naval Kumar Shukla	● Data Pre-processing ● Baseline LR ● Lasso, Ridge ● PCA
Udit Narang	● Data Pre-processing ● Feature Expansion ● Polynomial Regression ● SGD Regressor
Kushal Juneja	● EDA ● Baseline LR ● Lasso, Ridge ● PCA

Thank you