FINAL PROJECT REPORT

Solar Power Potential Estimation in India

Thummala Milind Kesar (11841160)
Gundu Shreya (11840530)
Saksham Bhushan (11840980)
Ruchika Gaur (11840960)
Pratik Sanjay Patil (11840850)

Data Collection

DATASET 1

(AllSolarPowerPlantsData.csv)

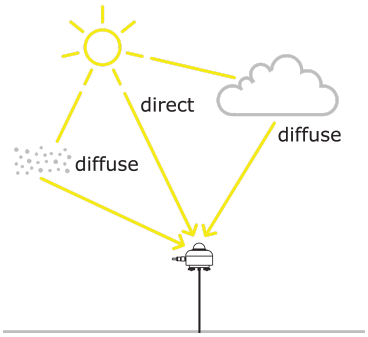
The dataset contains the data for all the solar power plants (in India) listed by the Government of India. This data includes the name of power plants, capacity, location (district and state), latitude, longitude, date of commissioning, average domestic electricity rates in Rs./KWh, per capita electricity consumption in KWh, monthly solar radiation, the monthly plane of the array.

I: 0	Name of Plant	Installed Capacity (MW)	Type	Location (District)	Latitude	Longitude	State	Date of Commissioning	Average domestic electricity rates in Rs./KWh	Per capita electricity consumption	Solar Rad Monthly	Solar Rad Annual	poa monthly
0	M/s. Sri Power Generation Ltd	1.0	solar	Dist Ananthapuramu	14.68333	77.60000	Andhra Pradesh	14-Jan-2012	9.95	1480	[6.619419574737549, 7.027534961700439 5, 7.0989...	6.08714	[205.20201110839844, 196.77098083496094, 220.0...
1	M/S. Amrit Jal Ventures Pvt Ltd	1.0	solar	Dist Ananthapuramu	14.68333	77.60000	Andhra Pradesh	7-Mar-2012	9.95	1480	[6.619419574737549, 7.027534961700439 5, 7.0989...	6.08714	[205.20201110839844, 196.77098083496094, 220.0...
2	M/s. Kishore Electro Infra Pvt Ltd	1.0	solar	Dist Guntur	16.30000	80.45000	Andhra Pradesh	13-Mar-2012	9.95	1480	[5.876765727996826, 6.448623180389404, 6.70761...	5.66221	[182.1797332763672, 180.5614471435547, 207.936...
3	M/s. Gajanan Financial Services Pvt\rLtd	1.0	solar	Dist Kurnool	15.82547	78.03012	Andhra Pradesh	14-Mar-2012	9.95	1480	[6.261613845825195, 6.7980875968933105, 7.0308...	5.98449	[194.1100311279297, 190.34645080566406, 217.95...

Columns

- *Name:* Name of the Solar Park or Name of the Company owning the Solar Plant.
- *Installed Capacity (MW):* Capacity of the solar plant in Megawatts.
- *Location (District):* Corresponds to the district in which the plant is located.
- *Latitude & Longitude:* Latitude and longitude of the district, obtained through Geocoder.
- *State:* State of the plant in which it is located.
- *Date of Commissioning:* Date of approval of the plant by the government.
- *Average domestic electricity rates in Rs./KWh:* Average domestic electricity per unit rate state-wise.

- *Per capita electricity consumption:* State-wise per capita consumption of electricity in (kWh per capita).
- *Solar Rad Monthly:* Month-wise solar radiation in kWh/m^2/day for each location of the already known solar plants. An array of values representing monthly data.
- *Solar Rad Annual:* Average of solar radiation annually in kWh/m^2/day for each location of the already known solar plants.
- *Poa monthly:* Monthly average of the Plane of Array Irradiance in kWh/m^2. Plane of Array Irradiance is the irradiation coming from the sun and also including the diffusion and reflection components of the irradiation. An array of values representing monthly data.



- *All sky insolation*: The average amount of solar radiation incident on a horizontal surface at the surface of the earth under all-sky conditions with the direct radiation from the sun's beam blocked by a shadow band or tracking disk.

We further grouped our dataset district-wise. The snapshot of the resulting dataset is as follows. Here we took the sum of Installed Capacity (MW) for each district. This dataset was used for modeling and some visualizations.
(This is Data.csv file)

	Unnamed: 0	Location (District)	Installed Capacity (MW)	Latitude	Longitude	Average domestic electricity rates in Rs./KWh	Per capita electricity consumption in KWh	Solar Rad Monthly	Solar Rad Annual	poa monthly	allsky_insolation
0	0	Aalo	0.05	28.170120	94.798230	4.0	703.0	[3.999358654022217, 3.6590754985809326, 3.6333...	4.204289	[123.98011779785156, 102.45411682128906, 112.6...	[3.36, 3.46, 3.39, 3.31, 3.27]
1	1	Abali	0.01	30.458560	78.250090	4.0	703.0	[5.87696647644043, 5.861727237701416, 6.584075...	5.881289	[182.1859588623047, 164.12835693359375, 204.10...	[4.68, 4.69, 4.6, 4.68, 4.46]
2	3	Adilabad	228.00	19.666670	78.533330	9.5	1896.0	[6.148787975311279, 6.589134216308594, 6.79743...	5.831581	[190.6124267578125, 184.49575805664062, 210.72...	[5.17, 5.07, 5.14, 5.02, 4.98]
3	4	Agar Malwa	58.77	23.825024	76.072969	6.5	1084.0	[5.8297810554504395, 6.8476409912109375, 7.099...	5.992356	[180.72320556640625, 191.73394775390625, 220.0...	[4.98, 5.02, 5.02, 4.98, 4.76]
4	5	Ahmedabad	10.51	23.027760	72.600270	5.0	2378.0	[6.179914951324463, 6.704948425292969, 7.12015...	6.023422	[191.57736206054688, 187.73855590820312, 220.7...	[5.34, 5.22, 5.26, 5.23, 5.16]

DATASET 2
(jagalur_hourlyv3.csv)
The dataset contains time series data for Jagalur Hybrid power plant power production for every 15 min from 1st January 2014 to 31st December 2014.
(These datasets have been obtained through mailing respective authorities and private firms recently.) (Dataset2)

Columns

- *time:* Date and time(mm-dd-yyyy hh:mm) of data collected.
- *Values:* Solar energy produced in that hour (MW)

	index	time	Values
0	0	01-01-2014 00:00	0.000
1	1	01-01-2014 01:00	0.000
2	2	01-01-2014 02:00	0.000
3	3	01-01-2014 03:00	0.000
4	4	01-01-2014 04:00	0.000
5	5	01-01-2014 05:00	0.000
6	6	01-01-2014 06:00	24.172
7	7	01-01-2014 07:00	76.397
8	8	01-01-2014 08:00	122.814
9	9	01-01-2014 09:00	154.061
10	10	01-01-2014 10:00	163.537
11	11	01-01-2014 11:00	161.615
12	12	01-01-2014 12:00	152.405
13	13	01-01-2014 13:00	131.113
14	14	01-01-2014 14:00	102.619
15	15	01-01-2014 15:00	60.183
16	16	01-01-2014 16:00	14.744
17	17	01-01-2014 17:00	0.000
18	18	01-01-2014 18:00	0.000

DATASET 3

We also obtained day to day weather-related data of Jagalur for the year 2014, to correlate it with the production data.

The data was obtained by visual crossing weather API.

(This is weather_data.csv)

Columns

The columns include the date, maximum temperature, minimum temperature, the average temperature of that day, wind chill, heat index, precipitation, snow depth, wind speed, wind gust, visibility, cloud cover, relative humidity, and weather condition.

	Name	Date time	Maximum Temperature	Minimum Temperature	Temperature	Wind Chill	Heat Index	Precipitation	Snow Depth	Wind Speed	Wind Gust	Visibility	Cloud Cover	Relative Humidity	Conditions
0	Jagalur, KA, India	01-01-2014	26.4	17.1	21.2	NaN	NaN	0.0	NaN	7.6	NaN	6.3	30.0	67.49	Partially cloudy
1	Jagalur, KA, India	01-02-2014	26.2	17.7	21.4	NaN	NaN	0.0	NaN	7.6	NaN	6.3	43.8	64.43	Partially cloudy
2	Jagalur, KA, India	01-03-2014	28.7	16.7	22.3	NaN	27.7	0.0	NaN	9.4	NaN	6.3	31.3	57.97	Partially cloudy
3	Jagalur, KA, India	01-04-2014	29.4	17.4	22.3	NaN	27.9	0.0	NaN	5.4	NaN	7.0	25.0	52.18	Clear
4	Jagalur, KA, India	01-05-2014	30.1	18.4	23.2	NaN	28.3	0.0	NaN	9.4	NaN	7.0	17.5	48.39	Clear

Method of Data Collection

- The list of all the solar power plants in India has been obtained by scraping the annual renewable energy plant report which is made available by the government.
- The average domestic electricity rates of the states for the year 2020 are collected from the “Bijli Bachao” website.
- Per capita, electricity consumption for the year 2018-19 in (KWh) is collected from the press information bureau.
- The latitude and longitude of the district in which the solar plant is located are collected using Geocoder and using these coordinates the average monthly Solar Radiation is obtained using the PVWatts Solar API.
- Production details of Jagalur Solar Plant have been obtained through mailing respective authorities and private firms recently.
- The weather dataset was obtained by using visual crossing weather API.

Data Cleaning

For DATASET 1

We have done the visualization only for the rows which had the district identified. For certain power plants, there was no district information available and for certain plants the district information was incorrect. For these power plants, we have not included them in the analysis yet but have used the coordinates of their respective state to identify them.

There was also the problem that names for certain districts were appearing twice (due to some manual errors (spelling mistakes etc)). We had to go through the list and correct such mistakes.

The plants for the state of MP had a lot of missing data, we had to manually look and label this data.

Since our analysis was at a **district level we grouped this dataset by district** (Data.csv) and used this for remainder of the modelling etc.

For DATASET 2

The raw dataset was very inconsistent, as there were some repetitive data and for some rows, the data was missing. So, we cleaned the dataset, grouped the data on an hourly basis, and arranged it in a day by day hourly manner.

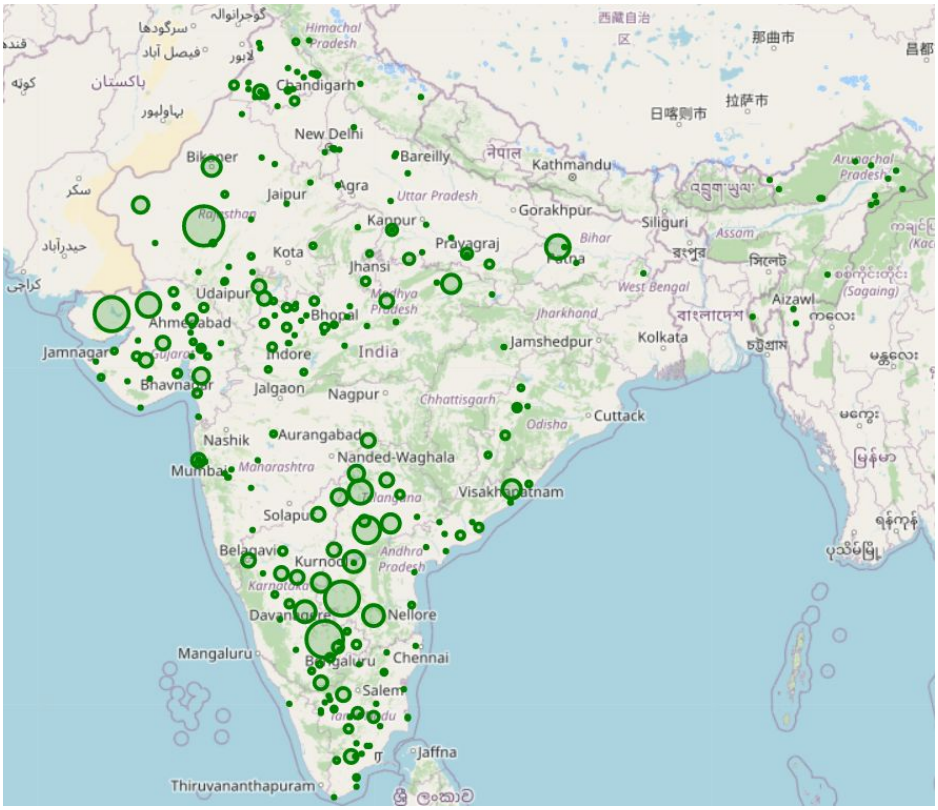
For DATASET 3

The weather dataset was already cleaned and consistent.

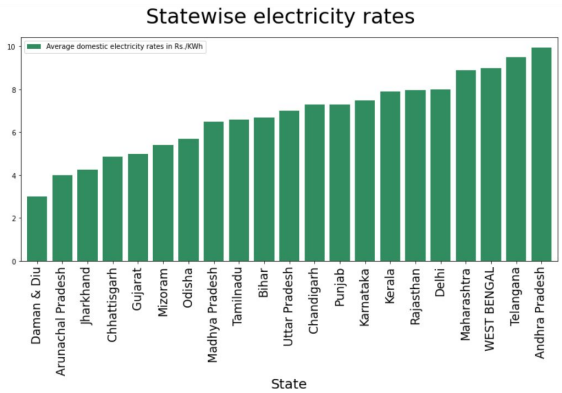
Initial Visualizations

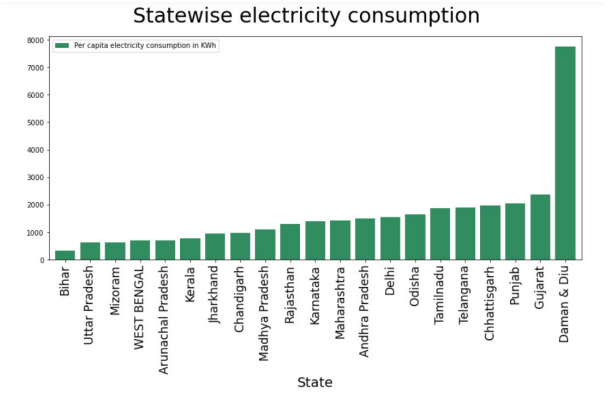
After the collection and cleaning of the data, we have plotted the following graphs to understand the data better.

- Map plot of India to show the distribution of solar power plants along with their capacity (Radius indicating the installed capacity w.r.t. each district).

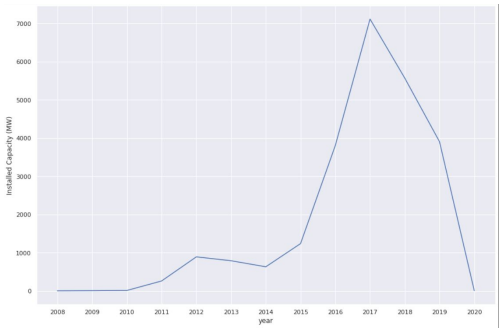


- Bar plot showing state-wise per Capita Electricity Consumption.
- Bar plot showing state-wise average domestic Electricity Rates.

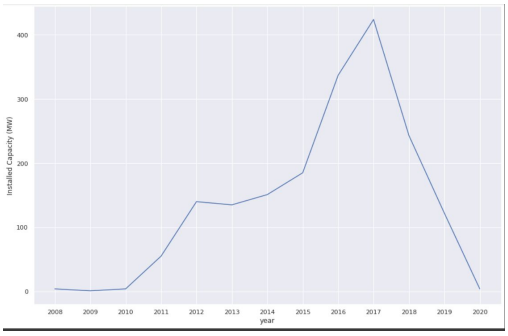




From the above graphs, we can infer that the average domestic electricity rates are highest in Andhra Pradesh and the Per capita electricity consumption is highest in Daman & Diu.

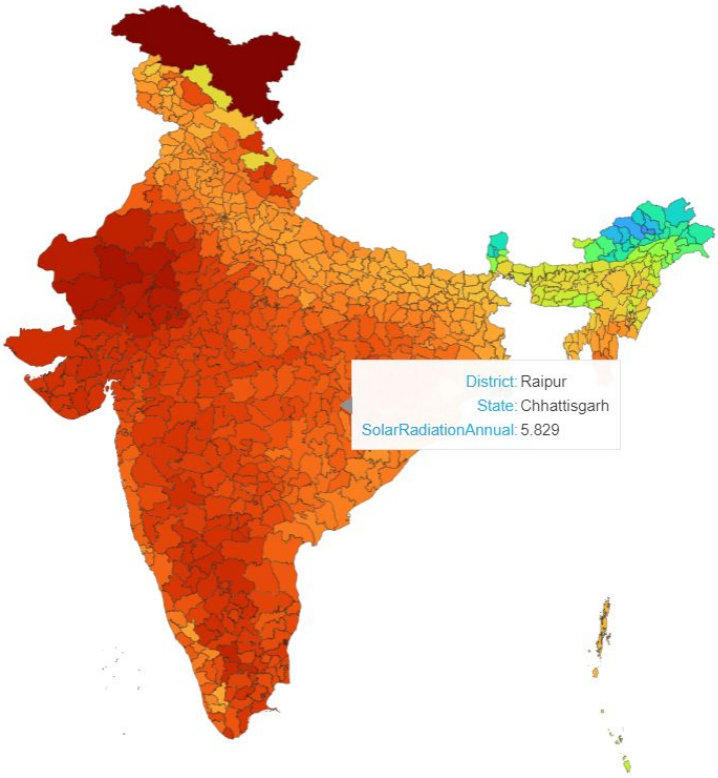


Installed Capacity Installed (year-wise)



Number of plants set up (year-wise)

- Map plot of India showing the annual solar radiation of that district.
  (Yellow < Orange < Red)(solar radiation)

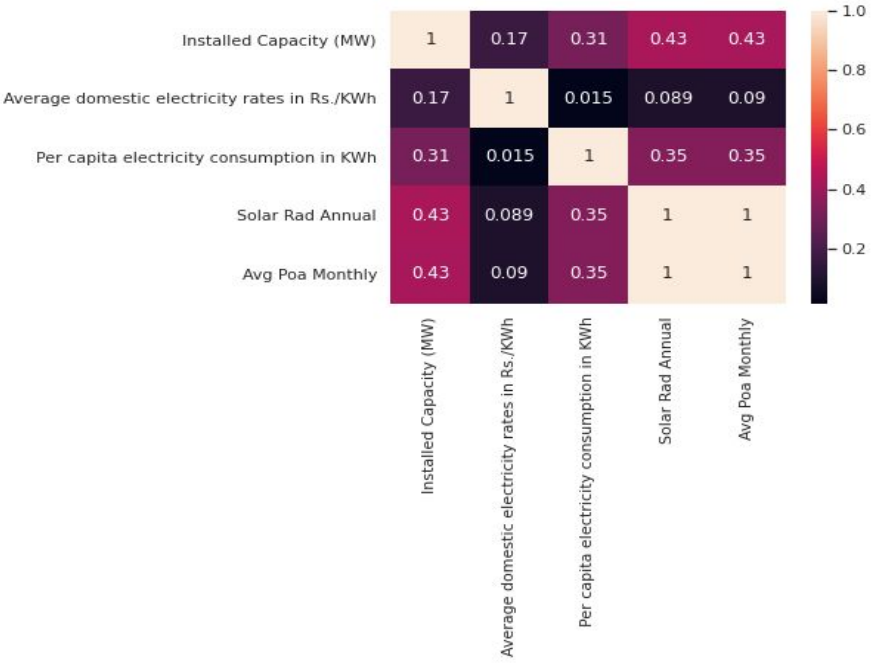


District-wise Solar Radiation in India

- Correlation Heatmaps of Installed Capacity (MW) with predictor variables are as follows

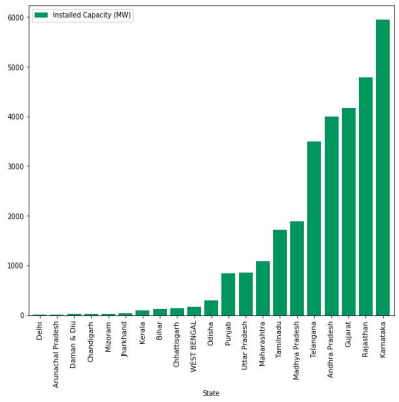
Pearson Correlation

Spearman Correlation

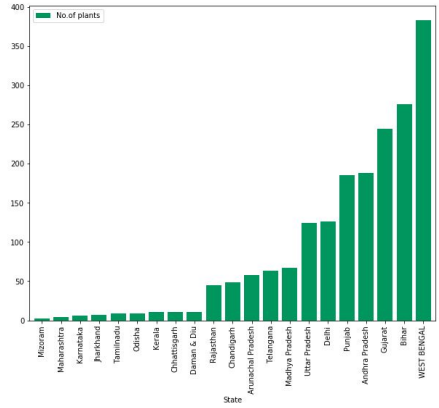


Spearman correlation values are higher than Pearson correlation values because of its less sensitivity towards outliers. Also, Pearson correlation is a parametric test whereas Spearman is a non-parametric test. Since our dependent variable does not have a

Gaussian distribution (which is discussed in the modeling section) Pearson is not suitable for our dataset.



State-wise installed capacity.



State-wise count of the plants.

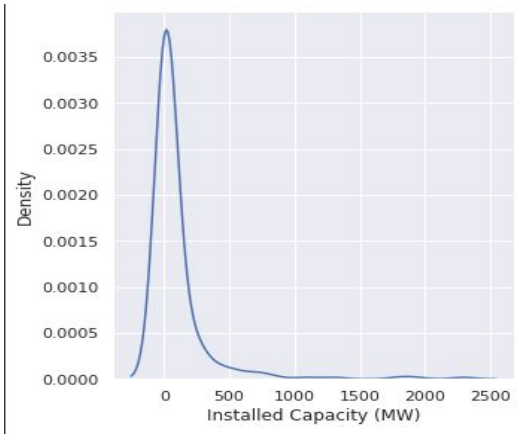
Modeling

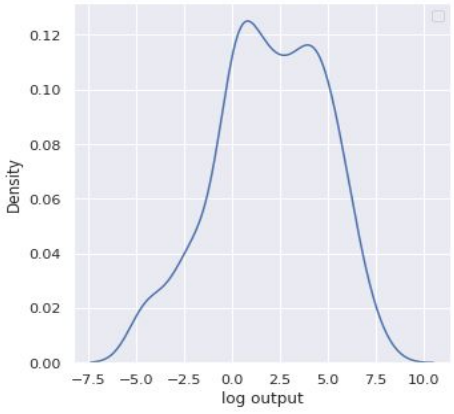
(Use Data.csv for modelling)

Regression Models

We have tried different approaches to make a regression model to predict the Installed Capacity (MW) using features such as Average domestic electricity rates in Rs./KWh, Per capita electricity consumption in KWh, Annual solar radiation, Avg Average of otherwise POA, and all-sky insolation.

Before we get into the various models we can see the distribution of Installed Capacity (MW) is log-normal based on the following plot. (We verified this using KS-test after removing certain plants which had installed capacity less than 0.5MW (the p-value obtained was 0.13)



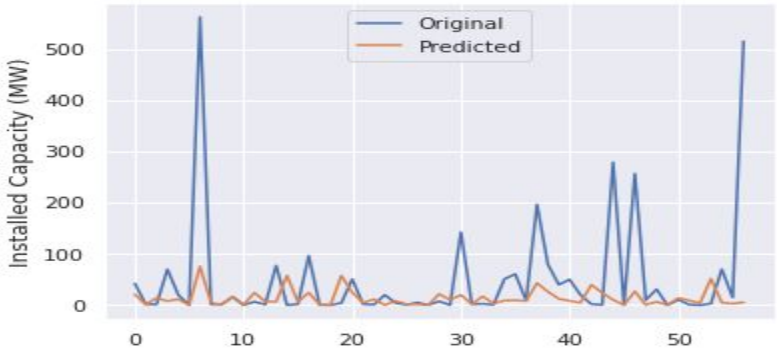


KstestResult(statistic=0.06510099672972569, pvalue=0.17386841155391364)

So instead of predicting the Installed Capacity directly, we tried to predict the log(Installed Capacity) as this data would have a somewhat normal distribution.

Multivariate Polynomial Regression

First, we made a simple multivariate polynomial regression model for which we obtained the following results:



The degree found to the best fitting was 2. However, as is evident from the plot the R2 score for this model was very low around 0.25-0.27 after cross-validation with k-fold. Clearly, this was not a good fit.

Generalized Linear Model

Next, we tried a generalized linear model where we used the Gaussian family with log link for sm.GLM (as our distribution was lognormal for the target variable). This method is a little different from above as taking the log output directly may affect certain results and hence GLM is more suited for models having such distributions in the exponential family.

The summary of the model was as follows:

```
==============================================================================
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  163
Model:                            GLM   Df Residuals:                      157
Model Family:                Gaussian   Df Model:                            5
Link Function:                    log   Scale:                         44719.
Method:                          IRLS   Log-Likelihood:                -1101.0
Date:                Tue, 17 Nov 2020   Deviance:                   7.0208e+06
Time:                        02:27:22   Pearson chi2:                 7.02e+06
No. Iterations:                    44
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -22.0273      9.175     -2.401      0.016     -40.009      -4.045
x1             0.2930      0.246      1.189      0.234      -0.190       0.776
x2             4.0222      2.883      1.395      0.163      -1.629       9.673
x3          1926.6959    738.359      2.609      0.009     479.539    3373.853
x4         -1922.1957    735.743     -2.613      0.009   -3364.225    -480.166
x5            18.6009      5.285      3.520      0.000       8.243      28.958
==============================================================================
```

(Note:- For training this regression model we removed very small plants with a capacity less than 0.5MW to improve results)
R2-score :- 0.296 (after 5 fold validation)
Hence we can see that this model is definitely better than the previous approach but again it is not good enough.

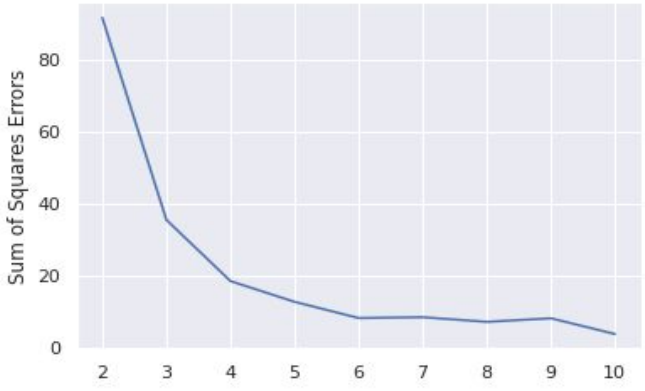
Random Forest Regressor
Using a random forest regressor we were able to obtain an R2 score of 0.38 with 5 fold validation. This was our best regression metric, however, it was still pretty low.

Hence we can see that regression techniques don't really fit very well on our dataset, maybe because there are several more features involved in predicting Installed Capacity and more reliable data sources to get solar radiation.

Classification Approach

Since it was not possible to predict the exact value of the Installed Capacity, we instead tried a classification approach where we gave class labels to each of the districts based on Installed Capacity (MW) and then tried to predict the class labels for new data. In simpler terms, the dataset was divided into large, medium, small plants and we predicted what class new districts will fall in.
In order to create classes we used two approaches:

- **Unsupervised Labellings (k means clustering)**:
  In this approach, we tried to assign labels to our dataset based on the K-means clustering algorithm. To find the best number of clusters we used the elbow method and silhouette score.



  We can see that the optimal number of clusters is 3 from the elbow method. Hence we used these class labels for predicting.

- **Manual Labels :**

The different classes were very unbalanced in the unsupervised labeling approach, so we instead assigned labels by a simple function with thresholds such that:

Capacity	Label	Count
Capacity < 5 MW	0	124
5 MW =< Capacity < 100	1	101
100 =< Capacity	2	58

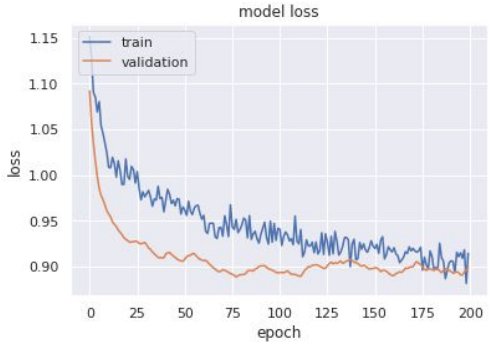
We chose these thresholds after trying out various thresholds to maximize accuracy.
We can see that the labels too are a bit equally distributed as compared to the previous case and the classification algorithms gave better results.

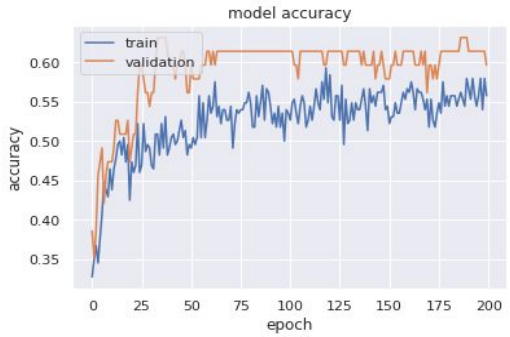
Logistic regression with Cross-Validation:
First, we tried logistic regression with 5 fold validation. For this, we obtained an accuracy of around 0.55.

Simple Neural Network with Dropout:
Next, we trained a simple neural network with the following structure:

```
model = Sequential()
model.add(Dense(16, input_dim=5, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(12, activation='relu'))
model.add(Dense(3, activation='softmax'))
```

We used “relu” activation and for the last layer softmax. We used Dropout to tackle overfitting which we first encountered. We used the optimizer as Adam and the loss function as categorical cross-entropy. The training-validation loss and the training-validation accuracy we obtained were as follows:





Accuracy is: 66.66666666666666

Micro F1 score is: 0.6666666666666666
Macro F1 score is: 0.6075036075036075

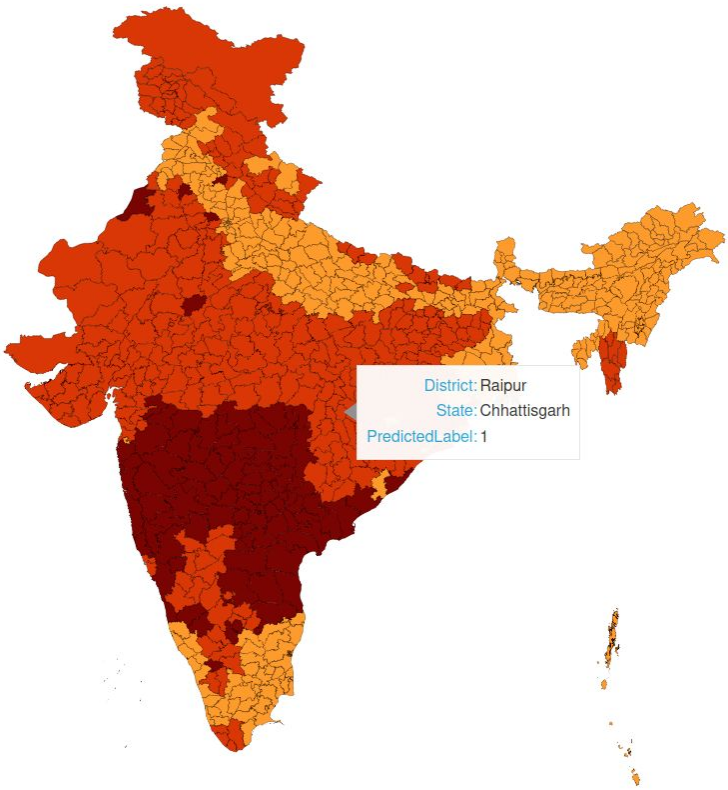
The accuracy varied between 60 % to 66%. We used this as our **final model** to predict the potential of various districts in India.

We can see that definitely classification algorithms are more suitable for our dataset as there doesn't seem to be any direct function mapping our input features to the output variable (Installed Capacity (MW)). It is more suitable for guessing whether the districts are suitable for small, medium, or large plants (ranges of Installed Capacity (MW)).

Plotting the Result

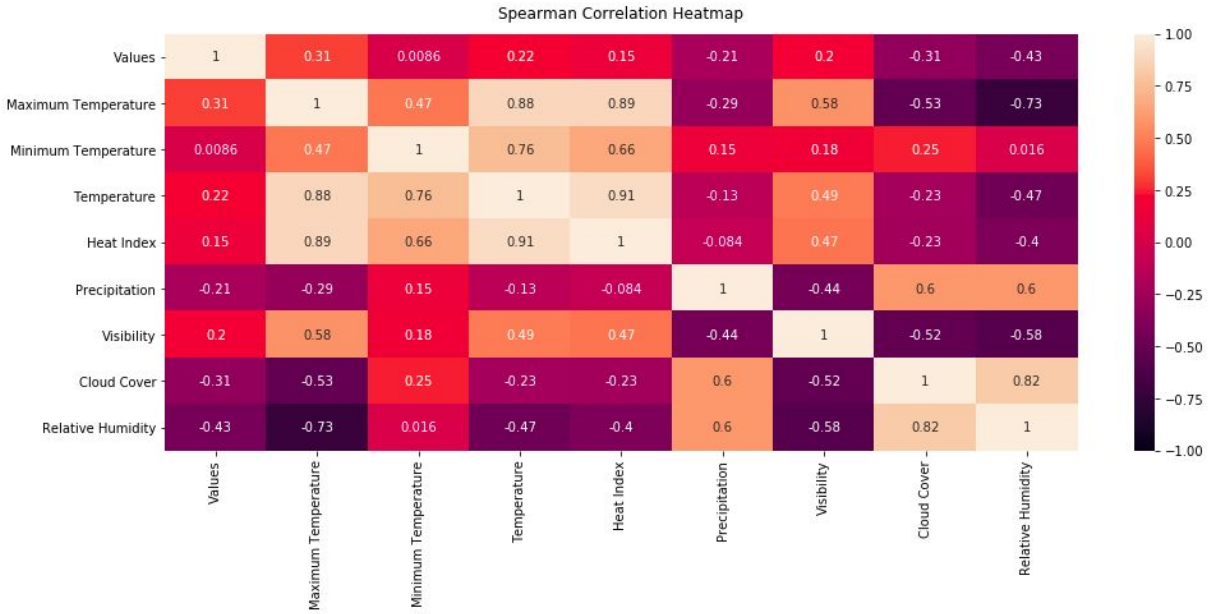
In the below graph we have shown the label predicted by our model for the districts of India.

As stated above, label 0 is colored orange, label 1 is colored red, and label 2 is colored dark red/brown.



Correlation with weather data

We correlated the day-wise production of one particular solar power plant (Dataset 3) with the weather. The results obtained were as follows:



We can see from the Spearman correlation heatmap that the production values correlate with relative humidity by -0.43. This is quite intuitive because with the increase in humidity there will be high chances for rainfall which in turn reduces the amount of production.

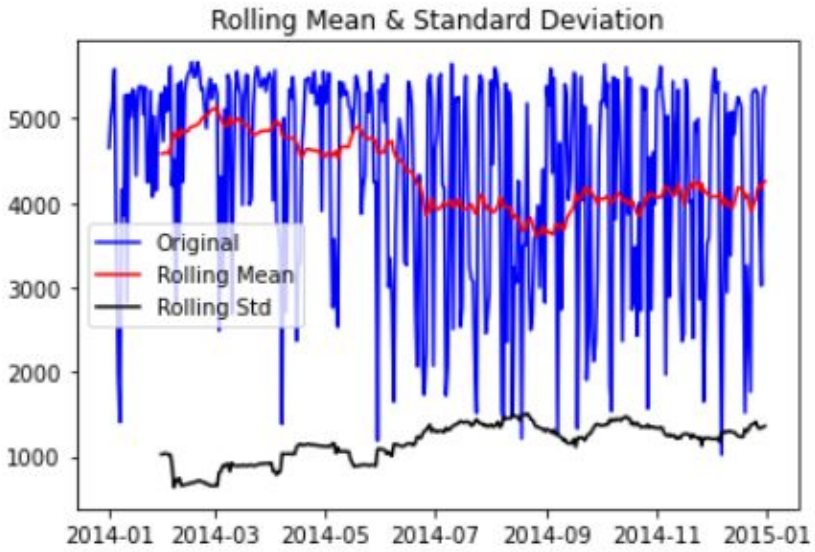
Spearman correlation gave better correlation values when compared to Pearson correlation as it is less sensitive to outliers than Pearson.

ARIMA Model on Time Series data

Before we can build a model, we must ensure that the time series is stationary. We used the two primary ways to determine whether a given time series is stationary.

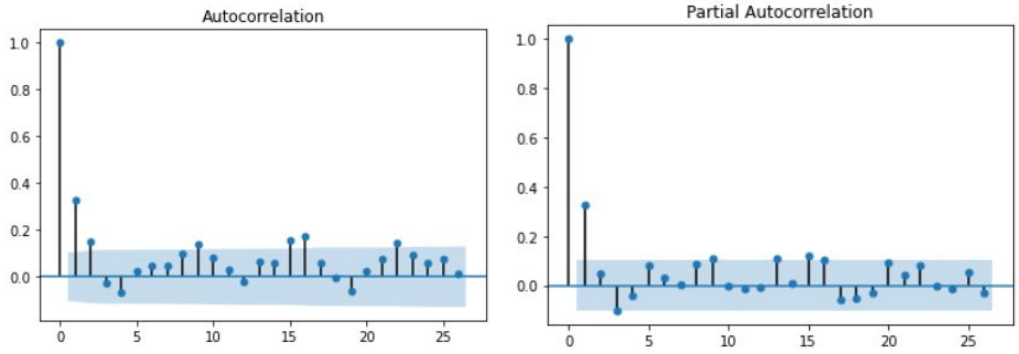
1. **Rolling Statistics**
2. **Augmented Dickey-Fuller Test**

The following is the Rolling Statistics and Augmented Dickey-Fuller test on our cleaned time series dataset. As we can see that the p-value is very low hence we can assume the time series to be stationary and hence it is relevant to apply the ARIMA model.




```
ADF Statistic: -13.548132607769483
p-value: 2.4388770970713854e-25
Critical Values:
        1%: -3.4484434475193777
        5%: -2.869513170510808
        10%: -2.571017574266393
```

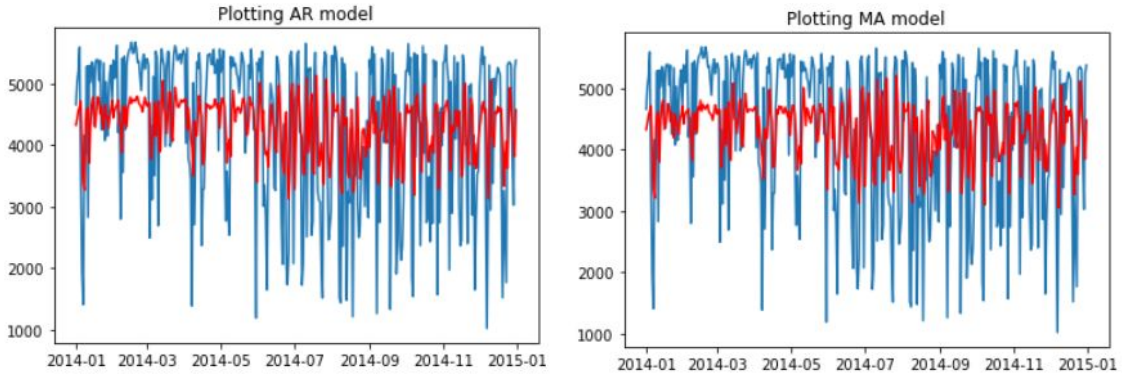
Now using ACF and PACF to figure out the best order of the ARIMA model



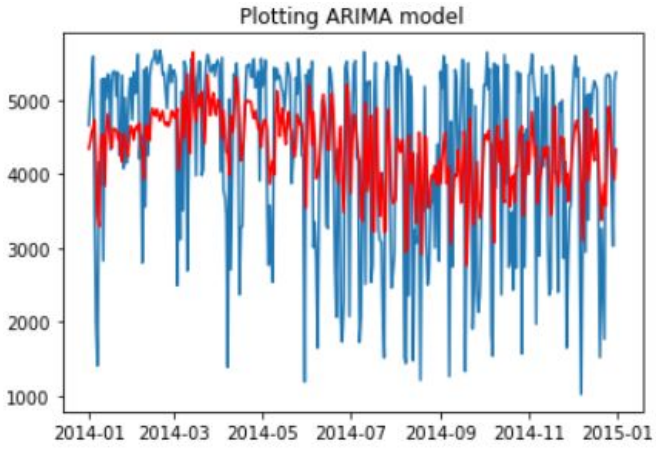
We get the value of p = 6, d = 0, q = 5

ARIMA model is the combination of **Auto Regression (AR)** and **Moving Average (MA)**

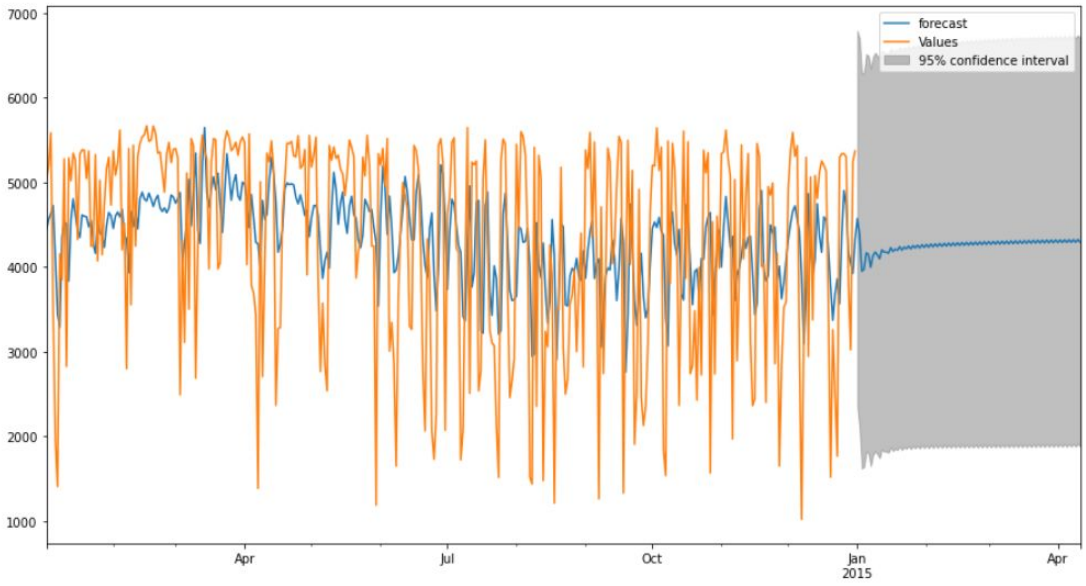
Note: The red line is predicted and the blue line is the actual value.



Now plotting the ARIMA model -



The final output of the ARIMA model prediction for the next 100 days
i.e. 1st January 2015 to 10th April 2015



Results and Conclusions

We can see that making precise predictions for the value of Installed Capacity (MW) using regression techniques is not very suitable for our dataset either because we need to include more features or maybe there is no such mapping to Installed Capacity (MW) based only on certain input features. Classifiers to assign labels to the district (based on Installed Capacity (MW)) seem to work better which is expected because intuitively here we are trying to just tell which range the Installed Capacity (MW) falls in.

We also performed correlation with the weather for the day-wise production of one particular solar power plant and the results have been explained in that section.

We also made a simple ARIMA model to estimate the production of one particular solar power plant.

Future Scope

- Collecting more data for adding more input features for improving the prediction of Installed Capacity (MW) in the regression models
- Optimizing the classifiers further for improved results
- Trying out more Time Series models for better estimation of the day-wise production of solar power plant
- Collecting and analyzing production data from more power plants to generalize the time series model.

Problems We Faced

- We had to spend a lot of time on data collection and cleaning initially as we had to collect data from multiple sources and had to do a lot of cleaning
- The API calls were limited for both the solar radiation and weather data. We could not get access to some other input features which we wanted to add such as land cost, production costs, etc.
- Since the dataset was not very suitable for regression analysis we had to try out various models and go through distribution to improve accuracy

Contributions

Member Name	Data Collection and Cleaning	Visualization	Regression and Classification Analysis	Time Series Analysis
Thummala Milind Kesar	Yes	Yes	Yes	Yes
Pratik Sanjay Patil	Yes	Yes	Yes	
Saksham Bhushan	Yes	Yes		Yes
Ruchika Gaur	Yes	Yes		Yes
Gundu Shreya	Yes	Yes		Yes

Files and Code

The notebooks attached in the submission folder are as follow:

- SolarPowerPotentialEstimation_Cleaning&Visualization:- For scraping the data from various sources and making the CSV data files (Don't need to run this as we have already made the CSV files)
- SolarPowerPotentialEstimation_Modelling :- For Regression and Classification tasks (use Data.csv (District wise accumulated dataset)). For predicting use FinalPredicted.zip which is attached in the link as it was too large (contains all district data))
- SolarPowerPotentialEstimation_TimeSeries :- For time series analysis. The dataset used for this is jagalur_hourlyv3.csv
- Data.csv - District Wise Accumulated dataset used for regression analysis and classification (This is DATASET 1 grouped by district)

- jagalur_hourlyv3.csv - Detailed Production of one solar power plant on an hourly basis (Dataset 2)
- weather_data.csv - This is the file corresponding to Dataset 3.
- Since some of the datasets were too large we added them in the drive and shared their link. This link contains the '.shp' file of the Indian districts and also one CSV file, which contains the predicted labels for the Indian districts. Along with the plots. Link. It also contains certain plots made using these files.

References

- https://www.bijlibachao.com/news/domestic-electricity-lt-tariff-slabs-and-rates-for-all-states-in-india-in.html
- https://pib.gov.in/PressReleseDetailm.aspx?PRID=1592833
- http://cea.nic.in/reports/others/planning/rpm/Plant-wise%20details%20of%20RE%20Installed%20Capacity-merged.pdf
- https://pvwatts.nrel.gov/pvwatts.php
- https://developer.nrel.gov/docs/solar/pvwatts/v5/
- https://www.visualcrossing.com/weather-api
- http://www.populationu.com/india-population
- https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
- https://towardsdatascience.com/machine-learning-part-19-time-series-and-autoregressive-integrated-moving-average-model-arima-c1005347b0d7
- https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/
- https://towardsdatascience.com/random-forest-in-python-24d0893d51c0
- https://www.statsmodels.org/stable/index.html