Project Report

Dt.30-06-2017

* **UMESHA NANDEESH**

The Project consisted of three problems listed below.

1. Evaluate the ability of forecasting Agencies like NCRWF and NOAA to fit the requirement of Algo Engines.
2. Build the Solar Irradiance forecasting model
3. Study on Price Forecasting and Build the basic model using SVM for Price forecasting using the Kaggle Thesis
4. **Evaluate the ability of forecasting Agencies like NCRWF and NOAA to fit the requirement of**

**Algo Engines.**

* The major agencies that were providing the forecast data were NOAA and NCMRWF.

**DATA EXTRACTION:**

* We need to extract only wind speed at 10meter and 80meter as well as wind direction this is obtained from ugrd and vgrd fields in the file
* Based on farm location we need to fine nearest location for which data is available (0.25 deg X 0.25 deg)
* Forecast data is stored in the following format:
* Forecast Data Store

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Source ID | Location ID | Forecast date | Forecast  Time | Forecast for time | Point Lat | Point Long | Height of obs | UGRD | VGRD | Wind speed | Wind dir. |
| 1 | 4 | 2015-07-03 | 09:00:00 | 03:00:00 | 22 | 70 | 10 | 3.14 | 0.00 | 3.142 | 0 |
| 1 | 4 | 2015-07-03 | 09:00:00 | 03:00:00 | 22 | 70.25 | 10 | 0.37 | 1.56 | 1.602 | 183.5519 |
| 1 | 4 | 2015-07-03 | 09:00:00 | 03:00:00 | 22.25 | 70 | 10 | -2.50 | -1.06 | 2.718 | 311.4195 |
| 1 | 4 | 2015-07-03 | 09:00:00 | 03:00:00 | 22.25 | 70.25 | 10 | -1.62 | -0.05 | 1.618 | 185.3851 |
| 1 | 4 | 2015-07-03 | 09:00:00 | 03:00:00 | 22 | 70 | 80m | -5.98 | -0.68 | 6.022 | 329.9809 |
| 1 | 4 | 2015-07-03 | 09:00:00 | 03:00:00 | 22 | 70.25 | 80m | 6.48 | 1.57 | 6.67 | 44.22 |
| 1 | 4 | 2015-07-03 | 09:00:00 | 03:00:00 | 22.25 | 70 | 80m | -5.25 | 3.97 | 6.58 | 33.91 |
| 1 | 4 | 2015-07-03 | 09:00:00 | 03:00:00 | 22.25 | 70.25 | 80m | -2.96 | -0.45 | 2.99 | 342.5843 |

* Each site will be given a location ID and latitude/longitude coordinates of each location will be stored in the source entity data store
* Based on the coordinates of the location, the Top/Bottom Latitude and Left/Right Longitude will be calculated in order to download & extract grid wind speed/direction data for the given location
* Calculate:
  + Bottom Lat = Round down (Location Latitude) to closest multiple of 0.25
  + Top Lat = Round Up (Location Latitude) to closest multiple of 0.25
  + Left Long = Round down (Location Longitude) to closest multiple of 0.25
  + Right Long = Round Up (Location Longitude) to closes multiple of 0.25
* **STEPS TO DOWNLOAD:**
  1. Go to the link- <http://nomads.ncep.noaa.gov/>
  2. Select: Dataset GFS 0.25 degrees – GRIB Filter
  3. Select the date for sample.
  4. Select the file according to forecasting hour. (.f000 for immediate forecast, .f001 for 1hr from then….)
  5. Select the levels 10m and 80m
  6. Select the variables UGRD , VGRD
  7. Select a sample latitude and longitude
  8. Select the URL for Web programming option.
  9. Then download. You will get an URL. By observing that the following code was written for automatic extraction of Data.
  10. The code is given in the Data Extraction folder with the name

“ wind\_data\_extract\_code\_for\_.foox\_autolocated.R ”

* 1. The code will automatically extract the data into various .csv files into different folders location-wise.
  2. The sample output files are given in the Data Extraction folder with the name “Sample output csv files”
  3. In the same way We extract the solar Data for the solar farm locations using the following code.

“solar\_data\_extraction\_autolocate\_TCDC\_code.R”

* 1. The NCMRWF data was available from the farm.
  2. The NCMRWF data was available in the format as given in the sample folder: “Sample NCMRWF data files ”
  3. That Data was Processed into the required format using the code: “NCMRWF\_sorting\_for\_append.R”
* After Acquiring the NOAA and NCMRWF data we will sort it out in the required format.
* Then append all the csv files into a single csv.
* Then the Root Mean Square Error for 4 Months was found out for NOAA,NCMRWF and Algo Engines.
* The values were found to be:
  + - Algo Engines – 26.5%
    - NCMRWF – 30.76%
    - NOAA – 38.9%
* Since the Percentage Errors of NCMRWF and Algo Engines were almost equal **NCMRWF was preferred as a backup data.**
* The plus point of NCMRWF was that the forecast was predicting on time but the Algo Engines forecast was delayed.
* The detailed Comparision results for each farm is given in the following .csv file. “All\_Stats.csv ”

1. **Build the Solar Irradiance forecasting model.**

* There were two machine learning Algorithms used to build the model – **The Artificial Neural Networks** and **Support Vector Machines**.
* **THE ARTIFICIAL NEURAL NETWORK:**
* The Solar irradiance values will be recorded during the sunlight hours only. So the forecasting has to be done only for those slots which are operating(sunny).
* The solar irradiance data for the Aditya Shakti farm was used for the analysis.
* The data extracted was for 2.5 months from mid-March to May.
* The input Data file is given in the following: “ farm\_data\_AS\_2.5\_months.csv” in the folder Solar Model.
* The Temperature Data for the same time was also obtained from the data base and the input is given in the file : “temp\_data\_AS\_2.5\_months”
* The Approach is as follows:
  + The time slots(24) from 5:30 AM in the morning to 5:30 PM in the evening was considered as the operating slots each of half-an-hour.
  + The model was trained with 7 days of data which contains the 24 slots for each day making the input a 24X7 matrix.
  + This was done using the following code: “ANN\_model\_with\_no\_temp.\_data.R ”
  + The error was found to be 20.76%.
  + To improve the accuracy the hourly temperature data was also given as input adding an extra 12 slots of temperature for each day making the input a 36X7 matrix.
  + This was done using the code: “ANN\_model\_with\_hourly\_temp\_data.R ”
  + The error was reduced to 19.66%
  + The accuracy of this approach was not so good as compare to that of the Average model that Algo Engines is currently implementing at 16.63%.
  + So this approach was not useful
* **SUPPORT VECTOR MACHINES:**
* The input data was the same for this as that of the previous approach.
* The approach is as follows:

1. **Continuous Input:**
   * The time slots(24) from 5:30 AM in the morning to 5:30 PM in the evening was considered as the operating slots each of half-an-hour.
   * The model was trained with 7 days of data which contains the previous 24 slots from the time we are forecasting making the input to be a 24X168 matrix.
   * Here each slot is predicted differently from the other unlike the ANN.
   * This approach was chosen to account the recent past data of solar irradiance.
   * This was analysed using the code: “SVM\_model\_with\_no\_temp\_data(24 slots input)(2hrs ahead prediction).R ”
   * The model was trained with the prvious 7 days data i.e., the previous 168 slots from the time of prediction.
   * The error of this model was found to be 19.37%.
   * The model was tested with the previous 3 days training also.
   * The model was also tested with the last 72,36 slots as input also those models didn’t prove to be efficient.
   * The other approach was training the model with Input of all 48 half hour slots of the day(keeping zero values for night times) .The error was quite high at 35%
   * The model was not performing well at the beginning of a day.So, we trained the model with the data from 7 AM to 5 PM only for each day. But the performance of the previous(from 5:30 AM to 5:30 PM) approach was better than this. This approach had an average error of 27% whereas the previous model had 19.5%.
   * Afterwards, the model was trained with the data of the slots prior to the slot we are forecasting. Ex: If we are forecasting at 10 AM slot we will train the model with the slots-6,6:30,7....9:30 for the previous 7 days. The error was 25.28%.
   * Decreasing the training Data leads to a significant decrease in the error to 18.4%.
   * The range of operating hours was found by observing the minimum start time and the maximum end time of the previous 3 days.
2. **Same Slot input:**
   * The input was changed to consider the same slot past data of the slot we are predicting.
   * The SVM was trained with the same slot data of the last three days and a slot prior to the slot we are predicting making it a total of 6 inputs. The data of the slot prior to the predicting slot of the same day was also given making the input variable count to 7.
   * The model was trained with the past 3 days data.
   * This was done using the code: “SVM\_model\_with\_no\_temp\_data(6 slots input)(2hrs ahead prediction)(same slot and previous slot last 3 days input).R”
   * The error was significantly reduced to 16.32% which is just 0.3% less than the Average model implemented by the algo engines.
   * Hence the Average model has outperformed all the Machine Learning techniques when it comes to solar irradiance techniques.

* (The other codes are also given in the “other codes” folder for reference with appropriate names. )

1. **Price Forecasting Model Using Support Vector Machines**

**AREA DIVISION:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Bid Area** | **Region** | **States covered under Bid Area** |
| 1. | N1 | North Region | Jammu and Kashmir, Himachal Pradesh, Chandigarh, Haryana |
| 2. | N2 | North Region | Uttar Pradesh , Uttaranchal, Rajasthan, Delhi |
| 3. | N3 | North Region | Punjab |
| 4. | E1 | East Region | West Bengal, Sikkim, Bihar, Jharkhand |
| 5. | E2 | East Region | Orissa |
| 6. | W1 | West Region | Madhaya Pradesh |
| 7. | W2 | West Region | Maharashtra, Gujarat, Daman and Diu, Dadar and Nagar Haveli, North Goa |
| 8. | W3 | West Region | Chhattisgarh |
| 9. | S1 | South Region | Andhra Pradesh, Telangana, Karnataka, Pondicherry (Yanam), South Goa |
| 10. | S2 | South Region | Tamil Nadu, Pondicherry (Puducherry), Pondicherry (Karaikal), Pondicherry (Mahe) |
| 11. | S3 | South Region | Kerala |
| 12. | A1 | North East Region | Tripura, Manipur, Mizoram, Nagaland |
| 13. | A2 | North East Region | Assam, Arunachal Pradesh, Meghalaya |

* The analysis was done for the Gujarat state i.e., Western 2 (W2) region.
* The temperature and the Wind speed data was obtained from the Patan farm site.
* The one year data was downloaded for price from the IEX website for Gujarat region.
* The data was converted into the following format.

“price\_data.csv”

* The hour variable was in the format “00:00 - 00:15” it was converted to “00:15” for each half an hour in R. (ex: 00:00,00:30,01:00,…..23:30).(00:30 represents 00:00-00:30 slot)
* The holiday data was collected from the official government website of Gujarat state and converted into the required format and merged to the price data frame. – “holiday\_data.csv”
* The day of the week was converted to interger numbers from 1-7 i.e., 1-Sunday,2-Monday,… 7-Saturday.
* The day of the year is written from 1-365/366.
* The hour of the day is in 0.5 to 24 which means 0.5 slot is the slot for 00:00 – 00:30.
* The temperature and wind speed data was obtained from the Patan farm located in Gujarat.
* The sample of Temperature and Wind speed Data is as follows: “temp\_ws\_data.csv”
* The data was converted into an appropriate format as given in the following excel file:

“model\_data.csv”

* The input and the output parameters were given in the table below:
* **Input Parameters:**

|  |  |  |
| --- | --- | --- |
| Parameter | Input Layer | Output Layer |
| Hour of the day | Sine and Cosine of the hour number of the day(0.5 to 24) in terms of (2π/24) for each half an hour | Load value for the half an hour slot to be predicted. |
| Day of the week | Sine and Cosine of the day of the week in terms of (2π/7) |
| Time of the year | Sine and Cosine of the day of the year in terms of (2π/365) |
| Temperature | Average daily temperatures in terms of moving point averages of 5,24,48 hours |
| Wind Speed | Average daily Wind Speeds in terms of moving point averages of 5,24,48 hours |
| Holidays | Binary input of 0 (non-Holiday) and 1 (Holiday) for the Bank holidays (except Christmas) |
| Previous Load | Loads for the same half-hour the day and week before. |

* The model was trained with 360 successive days of Data.
* The testing was done for 80 days.
* The following Code was used: “price\_forecast\_model\_svm.R”
* The Mean Absolute Percentage error for this model was found to be 11.05 %
* **KEY POINTS TO NOTE:**
  + The data used for the Temperature and Wind speed was the exact values, but while we forecast we have to use the forecasted temperature and wind speed Data.
  + This may introduce some more errors into the findings.
  + The Wind speed effect will not be there much in Gujarat since most of the production is from the solar only. This was checked by training the model without the wind speed data.
  + The error was found to be 10.84%
  + The model was checked with not giving any temperature or wind speed the error was 10.27 %.
* The Model has to be further analysed for the source of the error and where the errors are more and why.
* The detailed summary of the price forecasting paper is also given in the folder.