BEST ISTANBUL DATA SCIENCE COMPETITION

Osman Dogukan KEFELİ[1], Enis ARICI[2], and Sanal VAROL[3]

[1]Faculty of Science and Letters, Istanbul Technical University
[2]Faculty of Aeronautics and Astronautics, Istanbul Technical University
[3]Faculty of Mechanical Engineering, Istanbul Technical University

June 2021

Abstract

In this study, we aim to estimate wind turbine power generation to eliminate the imbalances in the energy industry. We have been given datasets from three different providers and three different estimates. Our main goal is to surpass the given prediction's precision. Additional features of columns are generated with the help of existing features which improved the models' efficiency significantly. Three different regression models have been created with given data. The linear regression model gave the best result among all other models with data within the range from the beginning of 2021 till today. Daily forecast of power generation for 5th of June presented in the results section.

1 Introduction

According to World Bank, electric power consumption and production are increasing annualy [1]. About the EU directives, countries should produce more renewable energy and it should reach 45 percent by 2030. The wind is one of the most paid attention and widely using resources as a competitive and sustainable renewable energy resource. As the number of wind turbines increases worldwide, it is becoming a huge challenge to assess their performance and efficiency. Accordingly, companies are looking for ways to establish more optimized systems and with that, they want to increase their profits. Today, with the development of technology in the energy sector, it has become possible to make big profits for the company by analyzing the huge data we have obtained. In this study, the goal is to do the data analysis study using weather forecast information. We try to estimate the energy to be produced according to the changing weather conditions in the most accurate way. If it is more accurate, the more profit company will make. In the wind energy sector, the day-ahead market is involved in the creation of these profits. The day-ahead market is an electricity market that allows the participants to determine the amount of electricity to be produced and consumed, thus regulating the electricity reference price. The day-ahead market is formed on an hourly basis, starting at 00:00 every day, with periods lasting 24 hours. Until 12.30, market participants ask the prices of the next day for the day-ahead market. Offers are accepted between 13.00 and 13.30 hours. In a conclusion, the energy quantities and prices become fully optimized. Electricity producers perform these processes every day and try to maximize their profits by making their forecasts as close to reality as possible.

1.1 Power Curve

Power curve monitoring, a part of Condition monitoring systems (or CMS as well-known), can serve as a good indicator of wind turbine condition.Wind turbines capture the wind to produce electrical power. The theoretical power output of a wind turbine can be expressed as:

$$P = \frac{1}{2}\rho V^3 \left(\pi r^2\right) \cdot C_p \tag{1.1.1}$$

where C_p is the power coefficient of wind turbine, r is the the wind turbine's radius, ρ is the air density, and V is the wind speed.

2 Aim

We make the energy production and its cost based on the forecast of the weather conditions around the wind turbine. To determine this forecast most accurately, we received weather forecasts from 3 different companies. Based on these estimations, we will try to generate the closest estimation data to the produced wind energy. In this way, we will be able to optimize pricing that allows us to earn maximum profit. This case aims to forecast today's wind power production. We estimate the production rate as closely as possible, thus we can aim to get the maximum profit for our company.

3 Data

Datasets come from three different sources. Data is in the time-series domain. All providers contain meteorological properties such as wind speed, wind direction, temperature, and humidity. The first provider's dataset provides a full set of data between 1^{st} of January 2020 till today. The second provider's dataset presented with 8 percent non-random missing values in meteorological properties and less than 1 percent missing power estimates from the beginning of the dataset. The third provider's dataset has 8 percent missing values on pressure and humidity columns but they are negligible when all datasets are combined and taken as a whole.

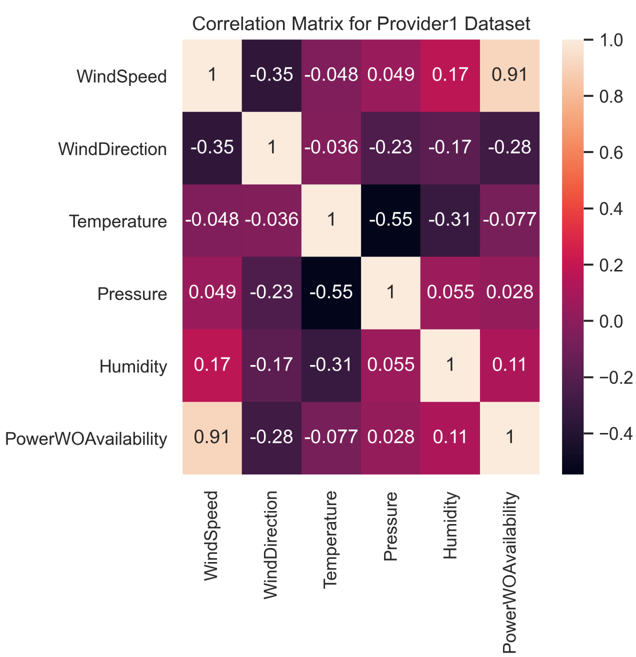

Figure 1: Correlation Matrix of Provider 1's Dataset

Here it is seen that provider 1's power availability has a strong corelation with wind speed as expected.

4 Analysis of the Data

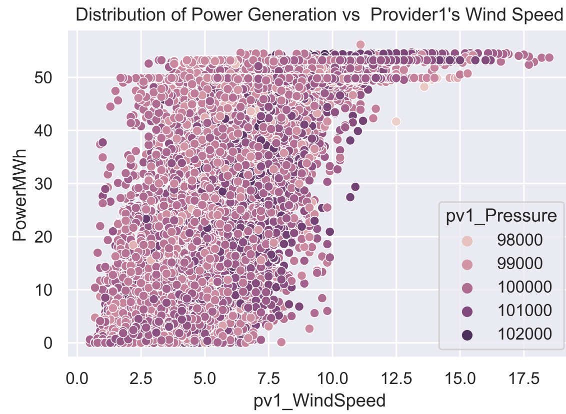

Figure 2: Distribution of Power Generation vs Wind Speed Colormapped Pressure

Due to the logarithmic relationship between pressure and density, the density increased more at increasing pressure values. Therefore, starting from the power density equation, production will increase at high pressures.

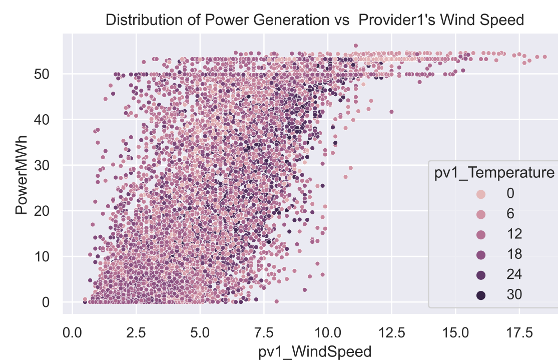

Figure 3: Distribution of Power Generation vs Wind Speed Colormapped Temperature

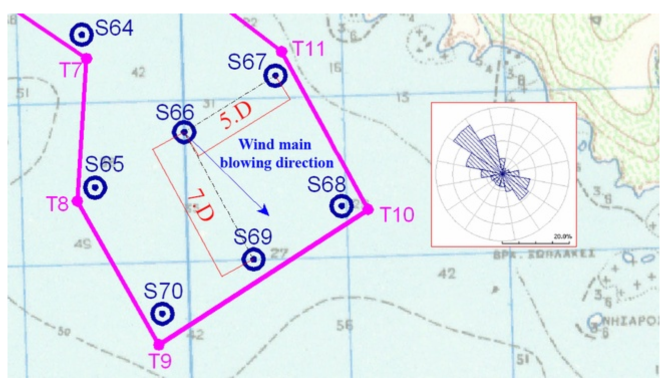

Figure 4: Micrositing [2]

The prevailing wind direction is taken into account when placing wind turbines. Turbines in the prevailing wind direction use 7 or 8 times the radius of the turbine, while in other directions it is taken 4 or 5 times.

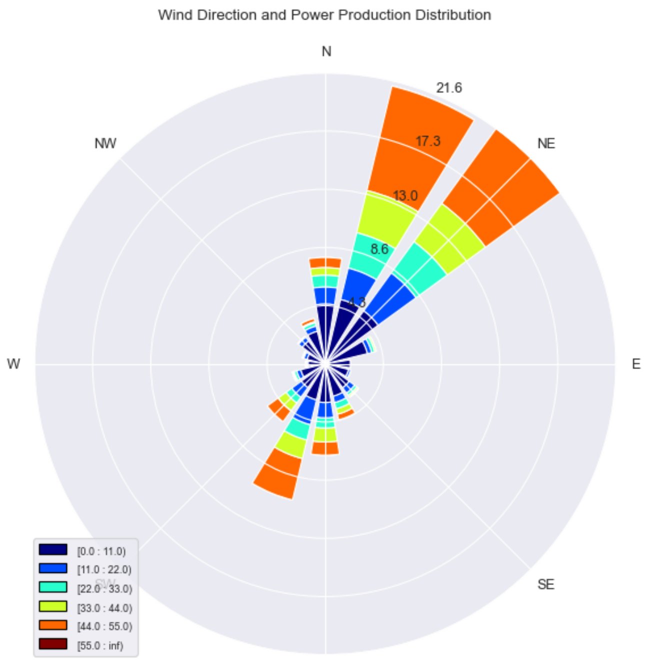

Figure 5: Rose Plot For Wind Direction and Power Availability

It is seen that most power is generated from winds that come from between north and north-east. The reason for this may be the sea effect in the prevailing wind direction and sea breezes. In addition, the turbines are expected to operate more efficiently in the prevailing wind directions, since the micrositing is structured according to the prevailing wind direction.

[hi]

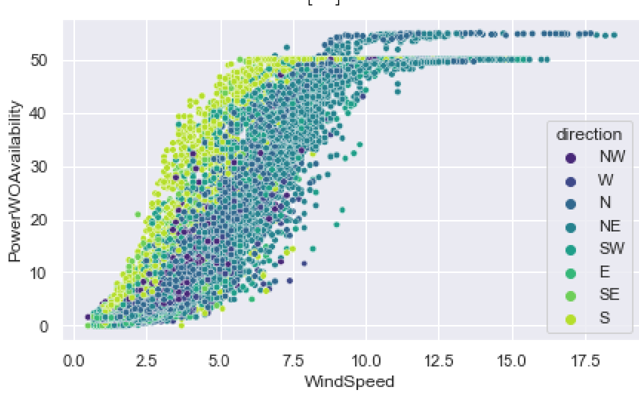

Figure 6: Power Availability's Relation With Wind Speed Classified by 8 directions

In this plot predicted power generations are higher for same wind speed when wind comes from south direction rather than other directions. The reason for this is that when the wind comes from the south, the average temperature is 5.2 °C, which increases the effect of air density in the power-law formula.

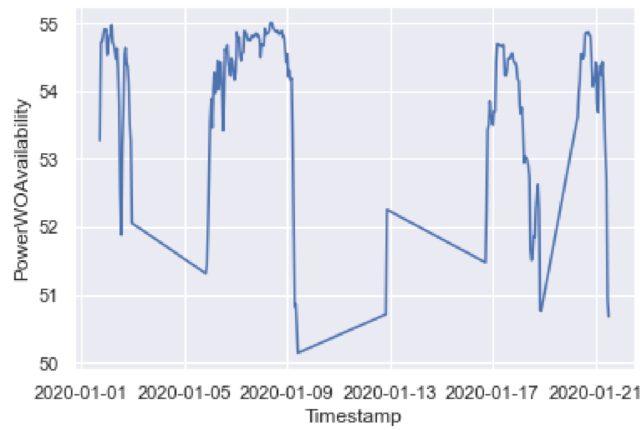

Figure 7: Data Where Max Power Estimates

In Figure 6 there is a gap in power estimates at 50 and 55 MWh. It is seen that this anomaly happened at the first 21 days of the data. Main reason for this is presumed a regulatory barrier or because some efficiency measures.

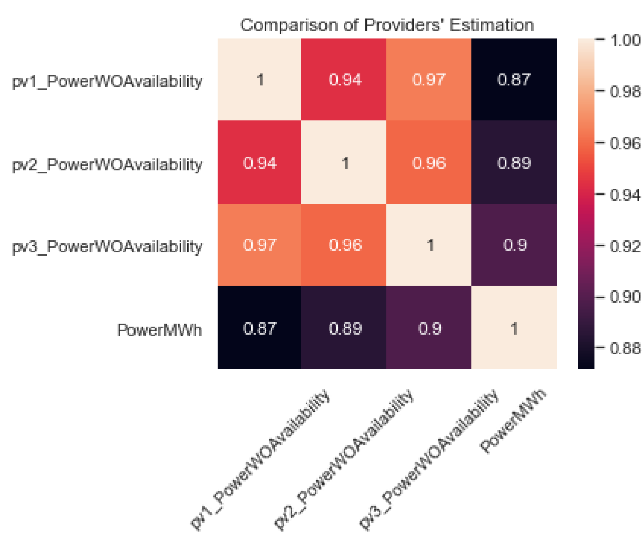

Figure 8: Correlation Between Provider's Estimates

Provider 3 had better predictions of power with respect to other providers. Assuming provider 3 posesses better numerical weather model than others.

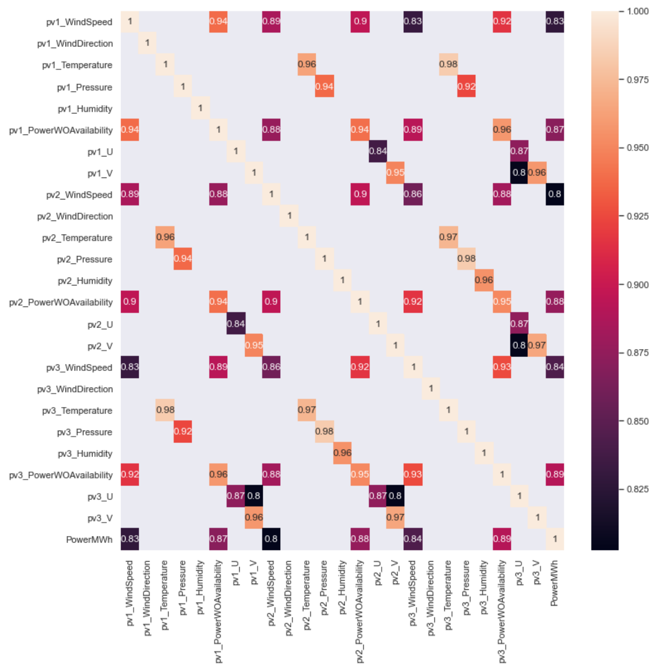

Figure 9: Spearman's Correlation Test (only 0.8 or higher)

Due to the nonlinear correlation between power generation and other meteorological properties, Spearman Correlation gives more precise results as seen.

Distancing between turbines is important due to the wind's attenuation between traveling.

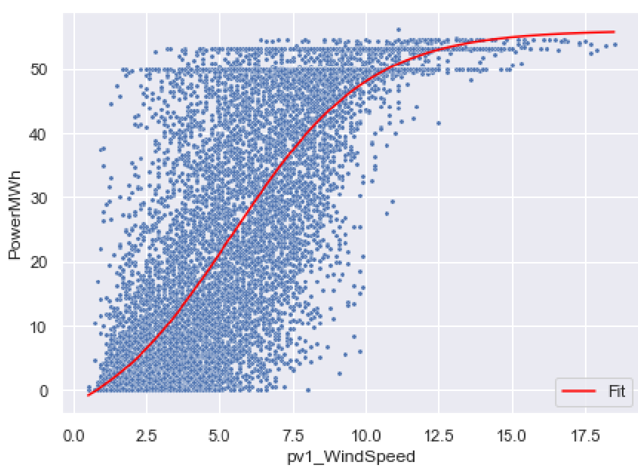

Figure 10: Sigmoid Fit to Power

Even the sigmoid fit function remains a rough estimate with wind speed taken as a parameter.

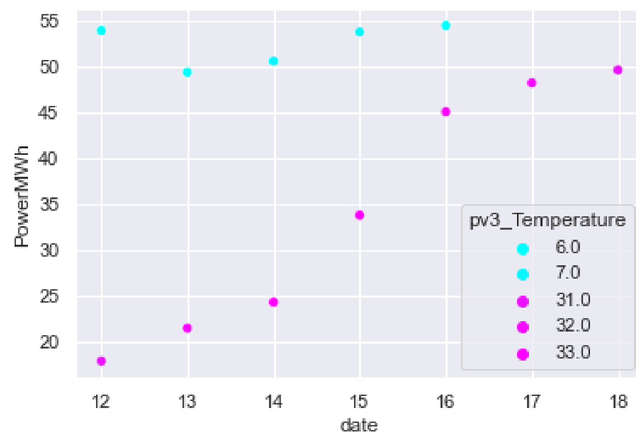

Figure 11: When turbines shutdown with effect of temperature

Wind turbines working between $-10\,^{\circ}\mathrm{C}$ and $30\,^{\circ}\mathrm{C}$ degrees. This graph shows that when temperature bigger than $30\,^{\circ}\mathrm{C}$ wind turbines shutdowned with brakes. Compared to the values at lower temperature values, it is seen that the capacity factor approaches 100 percent when the wind speed is 12 m/s.

5 Modeling

For producing a machine learning model with given datasets three models will be utilized. Linear Regression, XGBoost, Random forest regression will be compared with different input features. From existing features, we are able to produce more properties, change in pressure and change in temperature will be calculated and taken into account. Another feature that can be obtained is wind speed's separation of vector components $\overrightarrow{u}$ and $\overrightarrow{v}$. Since data is in time-series format, month and hour of the day parameters are added as extra features. Realizing that both the month and hour of the day are cyclic parameters. Therefore taking sinus and cosine of these features evolves the model. These features showed a significant performance boost in the results.

6 Conclusion

In conclusion, we have developed a better, more precise model that predicts the power generation from wind turbines according to the given actual generation. Comparison of the models that we have created linear regression model gave the most precise results when comparing root mean square errors.

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2} \qquad (6.0.1)$$

Models	RMSE
Provider 1 Estimates RMSE	10.442
Provider 2 Estimates RMSE	10.137
Provider 3 Estimates RMSE	9.533
Random Forest	9.620
XGBoost	9.926
Linear Regression	9.399

Above all linear regression model performed the best result.

7 Results

As the result, we compared our prediction model and provider prediction results with the 'Root Mean Score Error' metric until June 4, 2021. The results of our linear regression model yielded the best results in this test. Our predictions for power generation on hourly basis presented as follows:

2

date	model_linear
2021-06-05 00:00:00	4.57778
2021-06-05 01:00:00	3.0605
2021-06-05 02:00:00	3.16259
2021-06-05 03:00:00	3.81739
2021-06-05 04:00:00	2.93654
2021-06-05 05:00:00	2.96785
2021-06-05 06:00:00	1.91015
2021-06-05 07:00:00	0.742537
2021-06-05 08:00:00	0
2021-06-05 09:00:00	0
2021-06-05 10:00:00	0
2021-06-05 11:00:00	1.84901
2021-06-05 12:00:00	4.57209
2021-06-05 13:00:00	6.21882
2021-06-05 14:00:00	7.64528
2021-06-05 15:00:00	7.0303
2021-06-05 16:00:00	7.01672
2021-06-05 17:00:00	6.61451
2021-06-05 18:00:00	6.2123
2021-06-05 19:00:00	6.9837
2021-06-05 20:00:00	7.92631
2021-06-05 21:00:00	5.94339
2021-06-05 22:00:00	3.68892
2021-06-05 23:00:00	1.75043

Figure 12: Predictions for 5th of June

References

[1] Electric power consumption (kwh per capita). URL https://data.worldbank.org/indicator/EG.USE.ELEC.KH.PC.

[2] The siting of a wind park. URL http://www.aiolikigi.gr/en/e-learning/the-siting-of-a-wind-park/.