Columbia University

ENGI 4800 Data Science Capstone and Ethics

General Electric

Wind Power Forecasting

Authors:
Jonas Han [jh3877@columbia.edu]
Aditya Jadhav
[aaj2146@columbia.edu]
Gaurav Singh
[gs2938@columbia.edu]
Simran Lamba
[sl4228@columbia.edu]

Supervisors:
Dr. Andreas Mueller
Dr. Paul Ardis

Fall 2018

December 17, 2018

Contents

---

1 Abstract

In the last decade, wind farms have become increasingly commonplace across the U.S. as society has shifted its focus towards renewable energy and advancements in technology has enabled wind turbines to be cost effective. The cost effectiveness of investing in wind farms largely depends on the location of the turbines, as location greatly influences the amount of electricity generated over time. A key factor in determining the location of a potential wind farm is the weather patterns of that given location. Essentially, investors must be certain that a wind farm will produce enough electricity over time to make their investment worthwhile. In this paper, we will discuss our

approach to modelling and understanding the relationship between weather patterns and wind turbine output given a specific location. Weather data was collected from Weather Underground and a sample of on-shore turbine readings was provided to us by General Electric. We aligned these two separate data sources using autocorrelation and cross-correlation and then we utilized vector autoregression (VAR) and long short-term memory (LSTM) models to predict turbine output given weather conditions.

2 Introduction

Renewable energy sources constituted 17.12% of the overall energy demand in the United States in 2017, among which Wind Power is the second largest source for renewable energy with 254.25 TWh of energy produced. The U.S. Department of Energy has analyzed a scenario in which wind power meets 20% of the U.S. electricity demand by 2030, which means that the US wind power capacity would have to reach more than 300 GW. Being able to forecast wind power generation using climate data will help in better planning and management of energy resources in the country. In this project, we collaborate with General Electric to analyze and investigate the predictability of wind power and site feasibility from weather data for the turbines manufactured by GE. The power generation data from the wind turbines are proprietary, thus data has been redacted which could identify the turbines exactly. Time stamps on the power readings are missing, thus we attempt to align the timeseries from weather data and power generation data to triangulate the time stamps for the readings based on limited information present in the dataset.

In order to build a predictive model for power output from weather, we also use an open source data source in addition to GE Turbine data, to pressure test the hypotheses on forecasting of wind power using weather data. Different machine learning models are also built and compared in order to conclude the modelling which is most effective for forecasting wind power.

2.1 NOAA data vs USGS data

In Figure 1 three of the plots look at different turbine specifications like Rated Capacity, Rotor swept area and turbine height vs average daily wind speed for all the wind energy projects in the USGS dataset and the corresponding weather station data from NOAA. This was done to see if there are any intuitive trends like windy places have taller or higher capacity turbines. We observe no such trends. Most turbines come with standard specifications like 1500 kW capacity, certain standard values of rotor diameter (corresponds to swept area) that has no bearing with the 'windiness' of the area that they're deployed in. These factors are perhaps influenced more by the budget of the project. However, there is an obvious trend with swept area and rated capacities.

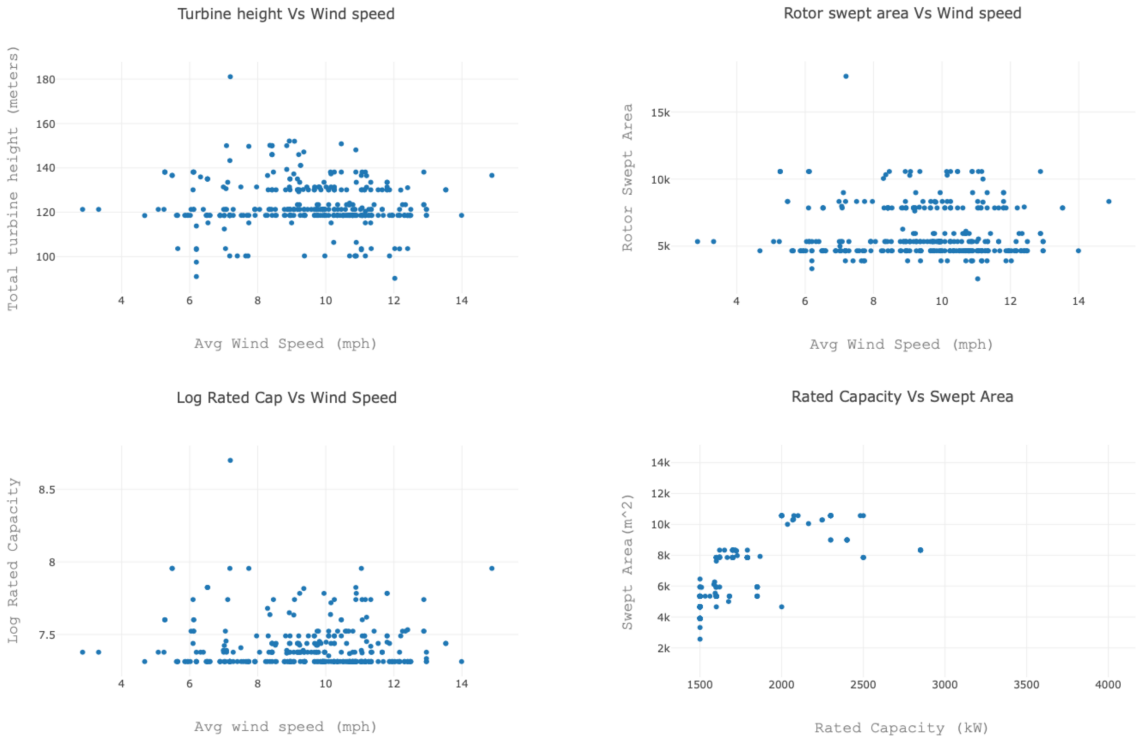

Figure 1

3 Related Work

3.1 Factors Affecting Wind turbine Efficiency

Wind power is primarily affected by the cube of the wind velocity, the area swept by the blade and the density of air. In other words Wind Power depends on: amount of air (volume), speed of air (velocity), mass of air (density) flowing through the area of interest[1]. Each of these individual factors like density can be affected by other contributing factors like ambient temperature, cloud coverage etc. while the swept area depends on the technical parameters like blade length, shape etc.

3.2 Wind Forecasting Based on Time-scale Classification

There are several approaches of classifying wind forecasting methodologies based on time-scale. Looking at classification systems defined in various literature descriptions[2,3], wind forecasting can be broadly divided into the following 4 categories:

Time Scale	**Range**
Super short-term	Few minutes to 1 hour ahead
Short-term	From 1 hour to several hours ahead
Medium-term	From several hours to 1 week ahead
Long-term	From 1 week to 1 year or more ahead

3.3 Wind Power Forecasting Methods

In recent years, many approaches have been implemented for wind turbine output forecasting. We can consider three main categories:

- **Physical**: These are mathematical models that use large amount of historical NWP(Numerical Weather Prediction) data like temperature, terrain and pressure. In general, these models require alot of computational resources and are recommended by meteorologists only for forecasts in large-scale.

- **Statistical**: This approach uses a massive training data-set to describe a relationship between input and output data. Statistical models are further categorized into: (1) times-series models and (2) machine learning, artificial intelligence models. The first includes uses simple auto-regressive techniques like AR, VAR, ARIMA, ARMA. The second one focuses on more recent papers trying to predict wind turbine output using algorithms like K-nearest Neighbors (KNN), Support vector Machine (SVM), Tree-based Gradient Boosting Machine (GBM, XGBoost), Artificial Neural Network (ANN) and Recurrent Neural Networks (RNN using LSTM or GRU).

- **Hybrid**: They can combine different approaches like combining physical or time-series models with ANNs.

For our project, we aim to explore statistical models including both time-series and machine learning approaches.

Erdem and Shi implemented four approaches based on autoregressive moving average (ARMA) method for a short-term wind prediction in 2011. These proposals achieved good results, but each has its advantages and disadvantages [4].

Nils, Justin and Oliver formulate the prediction task as regression problem and test different regression techniques such as linear regression, k-nearest neighbors (KNN) and support vector regression (SVR) in 2016. In their experiments, they analyze predictions at individual turbines level as well as entire wind parks. The most important result is that predictions with the highest accuracy are achieved for both setups with the SVR technique. They conclude that a machine learning approach yields feasible results for short-term wind power prediction and outperforms the traditional persistence model [5].

Also neural networks have been applied to wind power prediction, e.g., [6] focuses on the incorporation of time series components into existing machine learning models and evaluating how it will affect the performance in one-step-ahead and multi-step-ahead forecasting scenarios. The authors of this paper propose and apply two temporally dependent models based on neural networks for the wind energy forecasting problem: Autoregressive Artificial Neural Network (AR-ANN) and Recurrent Neural Network (RNN) using LSTM.

In this paper, we try to predict instantaneous wind turbine output essentially using two models: (1) Vectorized Autoregression (VAR) and (2) Long short-term Memory Recurrent Neural Network (LSTM RNN). We compare the predictions from these two models with the predictions from a baseline Random Forest model.

4 Turbine Readings Exploration

Our industry mentor, Dr. Paul Ardis, provided us with a sample of 86 on-shore wind turbines from a GE project in Kern County, CA. The sampled readings were taken in 2016 and contained 2,600 instantaneous power readings (kilowatt hours) for each of the 86 turbines. Paul was unable to provide us with crucial metadata which included: project(s) name, turbine model(s), turbine dimension(s), periodicity of readings, and timestamps of readings. The following data exploration and analyses will attempt to uncover some of these details regarding GE's sample of turbine readings.

4.1 Triangulating GE Project

Since there are multiple projects in the domain of location provided to us, we attempted to narrow down the exact project so that we can get the weather station closest to the turbines from which the readings have been drawn.

As evident from the distribution of the observations, we can identify the max capacity of the turbines. The spike at the end of the distribution suggests that the turbines had a max capacity of 1500 kW. Some erroneous measurements were recorded from the sensors which are above 1500 kW. The frequency distribution of the readings look uniform suggesting full range of power generation.

There are in total 10 GE projects in the Kern County.

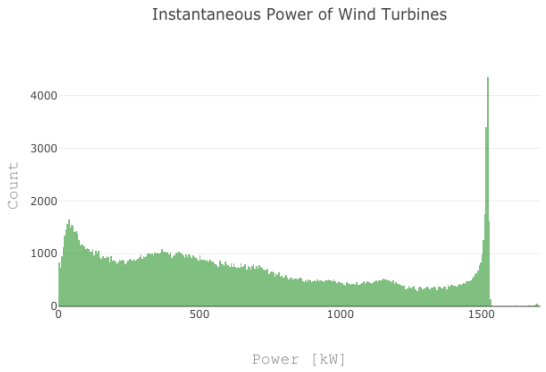

Figure 2

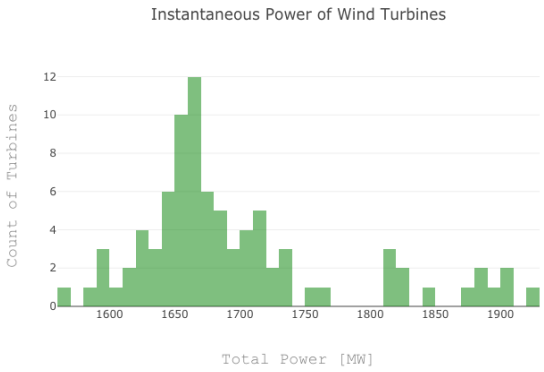

Figure 3

Based on the analysis above, we triangulate two GE projects with greater than or equal to 86 turbines, 1 model type and max capacity of 1500 kW. Alta I and Manzana Winds are the two projects which meet the filter criteria and both the projects have the same weather stations associated with them.

4.2 Relationship Between Turbine Readings

We first tried to gain an understanding of the relationship between each of the turbines. Below is a correlation matrix displayed as a heatmap. Note that all the turbines are at least fairly correlated ($r > 0.6$) with one another. Also, there appears to be two separate groups of highly correlated turbines. Specifically, turbine 1 through turbine 74 form a group of highly correlated turbines and turbine 75 through turbine 86 form another group of highly correlated turbines. The reason for this

pattern is that the latter group of turbines have interpolated values for approximately 320 readings. This is displayed in the time series plot adjacent to the heatmap. Note that turbines 77 and 80 have interpolated readings for readings 865 - 1185. In the interest of accuracy, we removed the latter group of 12 turbines (turbines 75-86) from our analyses and will only consider the readings from the first 74 turbines for the remainder of the paper.

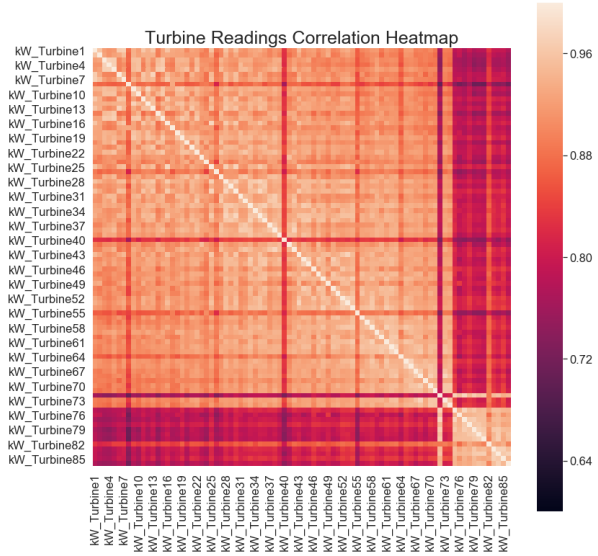

Figure 4

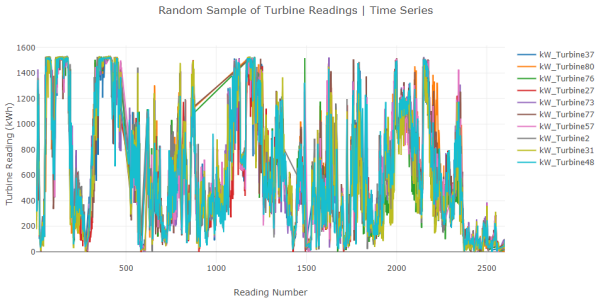

Figure 5

Next, we tried to gain an understanding of the variance across turbine readings in our dataset. The time series line graph displays a random sample of 10 turbines from the 86 turbines in our dataset. The key takeaway from the plot is that there is fairly low variance in the readings across all turbines. We generated this plot numerous times with different random samples and the graph remains largely unchanged. This strongly indicates that all turbines in our dataset are of the same model and dimensions, and in the same geographic vicinity. Thus, we will proceed with this assumption for the rest of our analyses.

5 Weather Data Exploration

5.1 Weather Underground Data Collection

Weather Underground (WU) is a weather data service that collects data through its own weather stations. The weather data available is recorded at 20 minutes interval which is at a better time resolution than that available through the NOAA API. After narrowing down the location of the turbines situated in the Kern County, the closest weather station available was Mojave, California. We scraped the data available on the website for the weather station from January 1, 2016 to June 30, 2016. Relevant fields from the data collected:

Variable	**Description**
Time Stamp	Time of the recorded reading

Temperature	Air temperature of the weather station in degree Fahrenheit
Dew Point	degree Fahrenheit
Humidity	Percentage Humidity
Wind	Direction of the wind or CALM for no wind
Wind Speed	in mph
Pressure	Air pressure in inches

5.2 Weather Data Analysis & Visualizations

Below are time series plots for averaged temperature and wind speed from WU data. Figures 6 and 7 show daily averages for temperature and wind speed respectively across 6 months starting from January to June. Temperature seems to follow a clear seasonal trend and steadily increases across the six months reaching 80 Fahrenheit and above starting from May. Wind speed on the other hand does not seem to follow a clear seasonal trend. Since it is hard to see the periodicity when plotting time series for all six months, we plotted time series for averaged temperature and wind speed only for a single month in order to get a more granular view. Figures 8 and 9 show daily averages for temperature and wind speed respectively for the month of January.

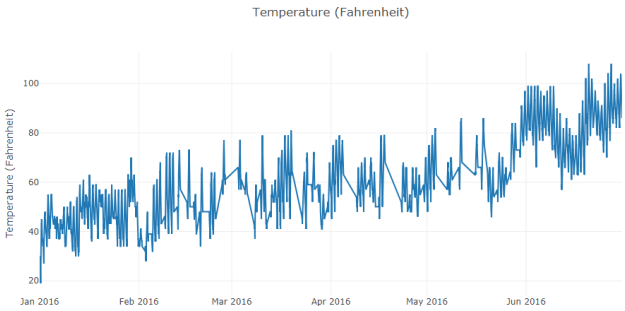

Figure 6

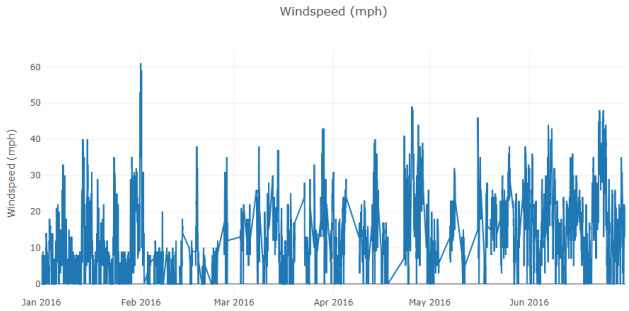

Figure 7

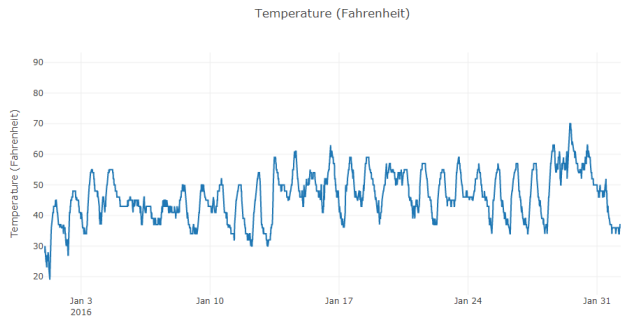

Figure 8

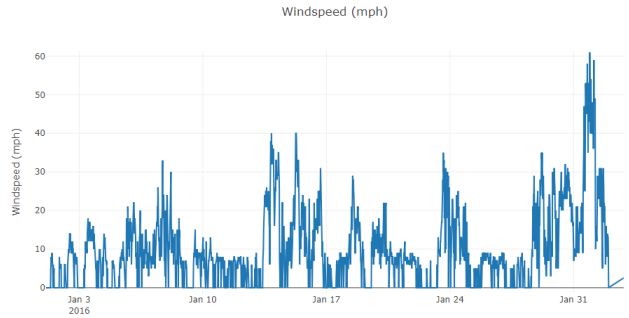

Figure 9

Further, we examine the distributions of temperature and wind speed by time of day. Figures 10 and 11 below exhibit temperature and wind speed respectively, averaged across 6 months for every 20 minute interval throughout the day. Wind speed appears to follow a steady rise and fall

throughout the day. Wind tends to be calmest during the early morning hours from about 5AM to 7AM and the strongest in the late afternoon and evening. Temperature too appears to follow an expected trend. Temperature seems to reach its peak around noon between 11PM and 3PM. It starts dropping during evening and is at its lowest during early morning between 4AM and 6AM.

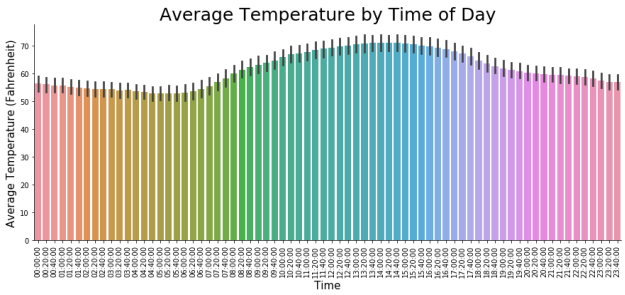

Figure 10

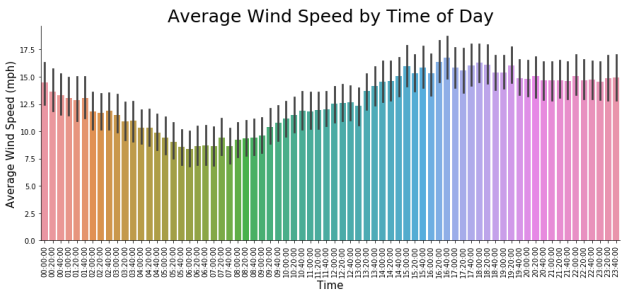

Figure 11

Below are the bar plots of temperature and wind speed respectively, averaged across months. Temperature follows a clear trend and keeps increasing, reaching its highest in June. It appears that average wind speed remains low during January to February and increases in the spring months of California.

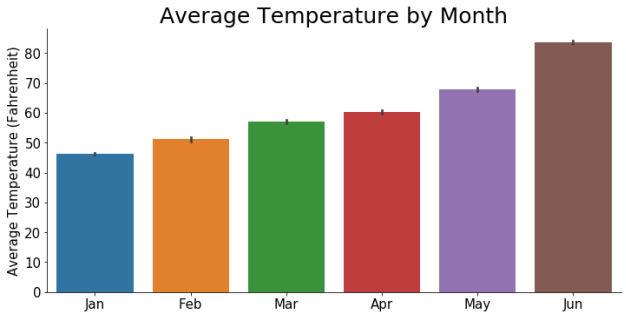

Figure 12

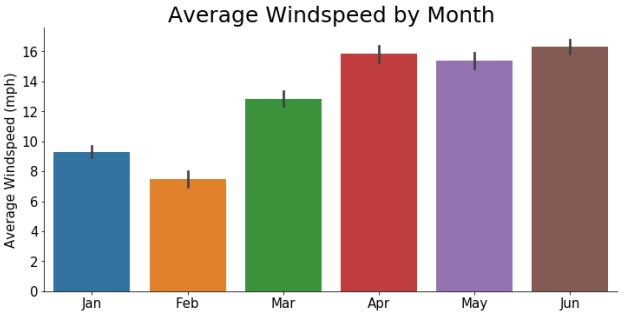

Figure 13

6 Uncovering Periodicity & Timestamps of Turbine Readings

As mentioned previously, the turbine readings that we received from GE were missing critical details necessary to align it to the weather data. Specifically, we needed to determine the periodicity and timestamps of the readings before we could proceed with any modelling. Below discusses our approaches to solving these issues.

6.1 Autocorrelation Analysis for Periodicity

We will first analyze the autocorrelation plots of the weather and turbine output data to determine the periodicity of the turbine readings. The autocorrelation plot for temperature acts as a baseline for standard cyclical patterns in time series data. As expected, we observe a very normal and steady pattern appear in the plot. There are peaks at time lags of 72 and 144 (1 time lag = 20 minutes), which translates to exactly 1 and 2 day lags, respectively. These peaks and valleys in the plot

correspond directly with the highs and lows in temperature within a given day. Wind speeds, on the other hand, do not exhibit such an ideal cycle. The wind speed autocorrelation plot does peak at a time lag of 72, but the correlation is significantly lower and the pattern quickly disappears after the initial peak. This indicates that wind patterns are remarkably less consistent than temperature patterns. We then plotted the autocorrelations for the turbine readings to hopefully find a similar initial peak that would represent a daily pattern in readings. Unsurprisingly, the autocorrelations for turbine readings are just as weak as they are for wind speed. There appears to be a slight peak at times lags between 250 to 300 and another larger peak at approximately 700 to 750, but the correlations are all quite weak. Furthermore, the peaks are significantly less smooth than the peaks found in the autocorrelation plots for temperature and wind speed. Due to the weaker pattern in wind speed and consequently, turbine output, it is extremely difficult to accurately determine the true periodicity of the readings. However, our best estimate is that approximately 250-300 readings represent a day's worth of data.

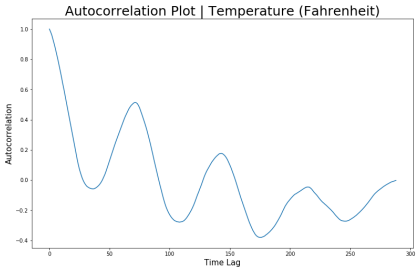

Figure 14

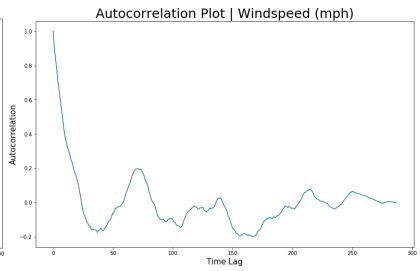

Figure 15

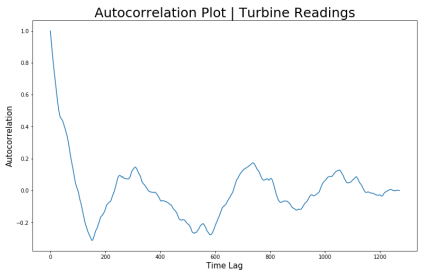

Figure 16

After relaying our findings from the autocorrelation plots to Paul, he informed us that the actual periodicity of the readings was 288 per day, which corroborates with our estimate of 250-300 readings per day. Under this assumption, the 2600 readings that we received would translate to approximately 9 days worth of data at 5 minute intervals. Now that we have confirmed the periodicity of the turbine readings, we will next utilize cross-correlation analysis and Pearson correlation with time delay to uncover the timestamps of the sampled turbine readings.

6.2 Cross-Correlation Analysis for Timestamps

Cross-correlation is widely used in the field of signal processing to determine where two signals or distributions have the greatest overlap. Intuitively, cross-correlation can be thought of as sliding a shorter signal along a longer signal to detect a notion of similarity between the two signals. Since we have 6 months of weather data and only 9 days of turbine readings, the longer signal is the weather data and the shorter signal is the turbine readings in our scenario. We are essentially trying to determine where the turbine readings distribution aligns best with the wind speed distribution. When the two distributions are closely aligned, this will produce a high cross-correlation dot product. Conversely, when the two distributions are not aligned, then a low cross-correlation dot product will result. Note that in the left plot below, several peaks appear at the end of January and at the end of April. Since the peaks in April are noticeably bigger than the ones in January, we will choose the highest peak in April as our starting timestamp. This peak with the greatest dot product corresponds to April 24th at 3:10AM.

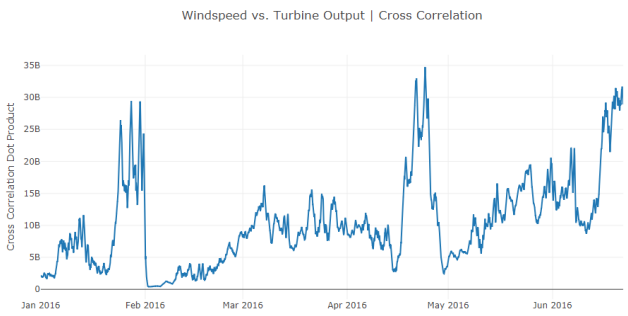

Figure 17

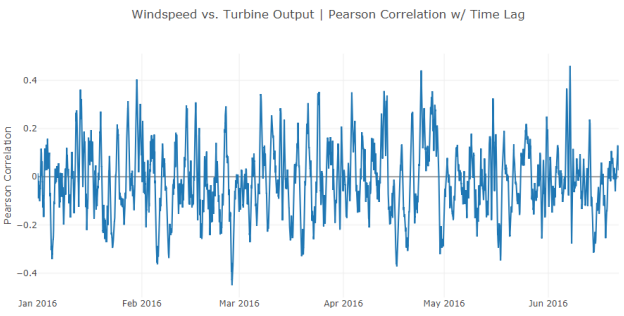

Figure 18

In the figure on the right, we are calculating the Pearson correlation coefficient instead of the cross-correlation dot product. Notice that there does not appear to be a single peak that is considerably larger than the others. However, the peak from the cross-correlation plot (April 24th at 3:10AM) coincides with one of the largest peaks in the Pearson correlation plot. The Pearson correlation between wind speed and turbine output with timestamps beginning on April 24th at 3:10AM is 0.44. We can now begin to build models as we have confirmed the periodicity and starting timestamps of the turbine readings.

6.3 Weather Data vs. Turbine Readings

Below are time series plots and scatterplots after setting the turbine readings to begin on April 24th at 3:10AM at 5 minute intervals. Note that the turbine readings appear to be moderately aligned with wind speed in the time series plot. However, the scatterplot indicates that there are still many instances where the wind is calm, but the turbines are generating electricity.

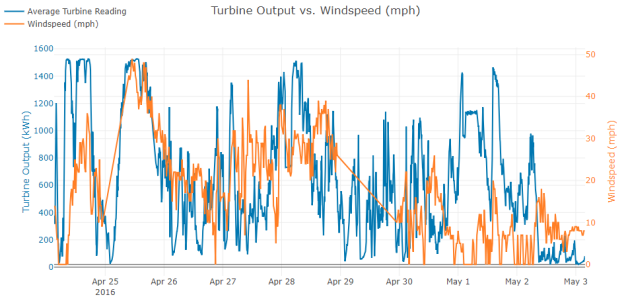

Figure 19

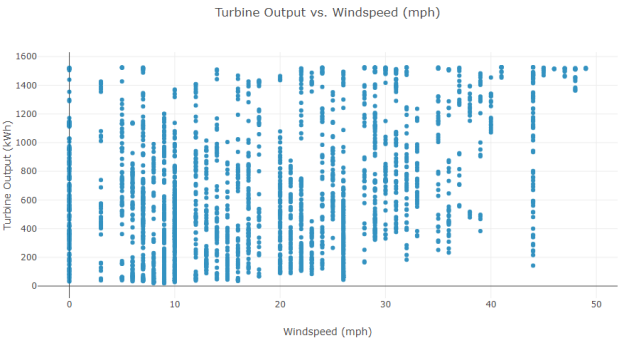

Figure 20

7 Turbine Output Prediction

In this section we attempt to predict the turbine output (as a fraction of the total rated capacity) given local weather parameters like wind speed, wind direction and temperature. As seen in the analysis preceding this section, we encountered issues with misaligned timestamps in the GE output data. Therefore, we looked for other sources of turbine output data that might be used to build a model. To this end, we found data made available by Sotavento Galicia, a company based in Spain.

They have provided hourly historical data going back 14 years from 2004 for an experimental wind farm with 24 turbines. Along with the total output of the wind farm, they have provided hyperlocal windspeed and wind direction measurements. We will compare models fitted on both data in this section. To make an appropriate comparison both output have been scaled to values between 0 and 100, denoting the percentage of grated capacity output by the wind farms.

7.1 GE Data model fit

For the GE data, our methodology was as follows: Based on the information we received from our industry mentor and subsequent analysis, we assumed equally spaced timestamps with a resolution of 5 minutes and that the data was reported for the year 2016. Next we found the alignment point with the highest cross correlation across the wind speed and turbine output time series. This point was found to be 2016-04-24 03:10:00 as seen in Figure 17.
The 2600 data points correspond to roughly 9 days of data from April 24th to May 3rd. Our methdology for training and testing the data was as follows: We trained the models on the first 80% of the data and evaluated it on the latter 20% as seen in Figures 21 and 22. In both cases, the wind speed, wind direction, temperature are the features used to predict the output. The Random Forest model will be used as baseline with which we compare performance of the time series based Vector Autoregression model.

Once we fixed the timestamps, we tried to fit Vector Auto Regression (VAR), Random Forest and LSTM models on this data. We used square root of the mean squared error as the comparison metric.

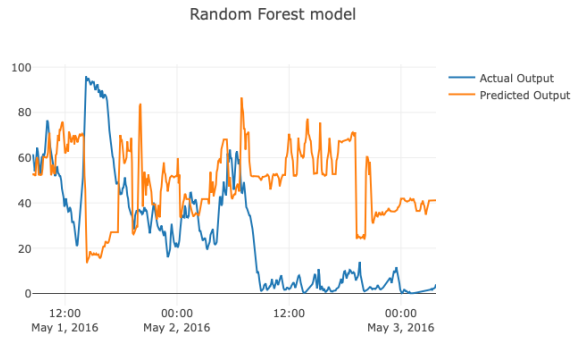

Figure 21

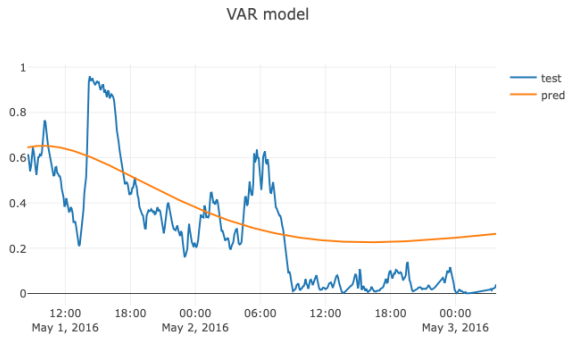

Figure 22

7.2 Sotavento Data model fit

Now we focus on the Sotavento Galicia data. We have about 14 years of hourly historical data which amounts to close to 120,000 data points with hyperlocal weather parameter values. This is a large enough dataset that could be used to train a Deep Learning model. To this end, we chose to experiment with Recurrent Neural Networks like LSTMs (Long Short Term Memory) and GRU (Gated Recurrent Unit) based architectures. The idea was to model the turbine output as a multivariate sequence prediction problem.

After exploring the data we found that for the years 2004 and 2008, the data had a lot of missing values. So we decided to drop them from the analysis. Subsequently, we trained the model on data for the years 2005 through 2016 and validated it against the 2017. Finally, the trained model was used to predict the output for 2018.

Following are results for the LSTM model for to separate seasons Spring and Fall. These seasons represent the time when the average wind speeds on an average are at a peak and trough. As we can see, the model predicts the directionality of the output time series.

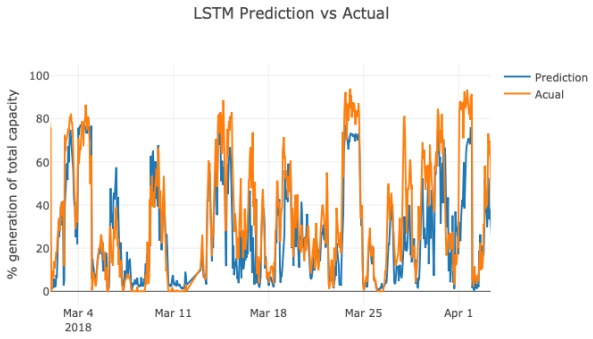

Figure 23

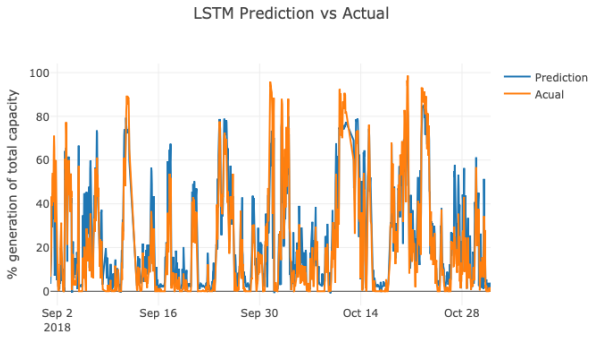

Figure 24

To document the performance of all the models across all datasets:

Dataset	Model	RMSE
GE	Random Forest	39.36%
GE	VAR	19.2%
GE	LSTM	36.67%
Sotavento	LSTM	13.22%

The LSTM model which performs the best and can predict an entire seasons worth of output given local weather parameters with a fair amount of precision, we could test it's efficacy on other wind farm locations, given local wind speed and direction. However, both data are aggregates for a collection of turbines and therefore do not have granular information on turbine specifications. We estimate that the model would benefit greatly from turbine information like turbine height, swept area and rated capacity. However, for site selection, an aggregated expected efficiency score would provide a good measure of scope for investment. In that regard, these current models that we built on the Sotavento data could be useful.

8 Challenges & Limitations

A major challenge that we encountered on this project was aligning the GE turbine readings with the weather data. We received the turbine readings with very little background information and metadata. Critical details of the data such as the periodicity and timestamps of the readings were

not accurately provided which severely hindered our progress and capabilities in the project. We were also led astray for a majority of the project in regards to the metadata.

Another significant limitation was the small sample size of turbine readings that was provided to us. Specifically, we were only given a sample of 86 homogeneous turbines for a single project in Kern County, California. Since the turbines all originated from the same project and were of the same model/dimensions, we were unable to to model the relationship between key turbine features (i.e. rotor diameter, rotor height, total height) and turbine output. Furthermore, we were unable to train our models on turbines from different geographic locations. That is, training on weather data and turbine readings from a single location greatly limits the capability and usefulness of our models. Lastly, the 2600 turbine readings that we received translated to only 9 days worth of data. There were not enough data points to accurately model the intraday fluctuations in wind speed as reflected in all of our GE-based models.

9 Conclusion & Future Work

In the arena of limited information present in the GE Dataset, we aligned the time series from GE Turbines power generation and weather data to build a predictive model for forecasting wind power. Temperature has more periodic patterns and varies very smoothly with time unlike wind, which is the primary driver for wind power generation. Post alignment of time series of GE power generation and climate data, we achieved a RMSE of 19.2%. To reinforce the finding that weather data can be used for predicting wind power, we built an LSTM model on open source dataset with RMSE of 13%.

In search for the right GE project, we correctly triangulated the wind project and maximized the correlation of the power readings from the project with the weather data scraped from Weather Underground. One key aspect of our results was to build a forecasting model in insufficient intelligence present. This also reflects the starkly contrasting setting of data science applications in industry vs working with ideal datasets. Our attempts to mitigate the challenges throughout were followed by bigger roadblocks. Landing on the precise dataset with appropriate hypotheses is more arduous than building a machine learning model on incomplete data.

In order to further enhance the model, the next logical step would be to inculcate turbine features in the model. Characteristics of turbines like rotor diameter and height are important drivers of power generation. More data can be sampled from the turbines to build a long term forecasting model, which may be used for transfer learning to discover new geographic locations for site selection of upcoming offshore wind turbine projects. We now understand that LSTMs work well on forecasting wind power output, in order to test the robustness of the model on GE Turbines, a wider time horizon of data with different regional variances maybe assimilated with also encompassing onshore wind turbines to understand the differences in power generation behavior of these turbines.

10 References

[1] Wind Power Fundamentals

[2] Soman,S.S.;Zareipour,H.;Malik,O.;Mandal,P., “A review of wind power and wind speed forecasting methods with different time horizons” North American Power Symposium (NAPS), 2010 ,PP.1-8

[3] Zhao, D.M., Zhu, Y.C. and Zhang, X. (2011) Research on Wind Power Forecasting in Wind Farms. Proceedings of the 2011 IEEE Power Engineering and Automation Conference, Wuhan, 8-9 September 2011, 175-178.

[4] E. Erdem and J. Shi, “ARMA based approaches for forecasting the tuple of wind speed and direction,” Appl. Energy, vol. 88, no. 4, pp. 1405–1414, Apr. 2011.

[5] Nils André Treiber, Justin Heinermann, Oliver Kramer, ”Wind Power Prediction with Machine Learning”, Published 2016 in Computational Sustainability.

[6] Rui Li, Pu Wang, Jingrui Xie, Alex Chien, Mustafa Kabul, ”Short-Term Wind Energy Forecasting with Temporally Dependent Neural Network Models”, SAS Institute Inc