### Project Report: Wind Energy Forecasting

**Introduction**

The way our societies generate, transport and consume electricity is undergoing significant transformation. As communities around the world seek to massively decarbonize their infrastructure in efforts to curb climate change, new technologies and systems will be adopted to replace traditional fossil fuel infrastructure. The electricity grid is transforming from a centralized model, wherein electricity flows unidirectionally from large centralized power plants to consumers, to a distributed model featuring many more energy producers and distributed energy resources (DERs). Renewable energy technologies like solar power and wind power will play an increasingly large role in electricity generation. Wind power, especially, is poised to see significant growth in the coming decades, with the advent of offshore wind technologies and the low costs associated with wind power generation. While wind power is appealing as a zero-carbon solution, it does have one drawback in that it is intermittent (i.e. the wind does not blow 100% of the time). To incorporate high penetration rates of renewable energy sources like wind power, the future electricity grid will require accurate energy forecasts to be operated effectively and reliably.

Thus, the goal of this project is to use machine learning to accurately predict hourly wind power generation at 7 wind farms, based on historical wind speeds and wind directions. The client for the project is a vertically-integrated electric utility company that must incorporate higher amounts of wind power generation into its electricity portfolio. Should the project be successful, the client will have a model that can accurately forecast wind power generation based on wind speed and wind direction. This model could then be applied to future wind data to forecast wind power generation. If the client is able to accurately forecast wind power generation, it can make better informed decisions with regards to electricity transmission sizing, sizing of nearby DERs, sizing of energy storage systems, and other issues related to renewable energy integration in a smart grid or microgrid.

**Dataset Description**

The data used for the project is provided by the Institute of Electrical and Electronics Engineers (IEEE), Power & Energy Society, and retrieved through the [Kaggle](https://www.kaggle.com/c/GEF2012-wind-forecasting) database. The dataset is a time series dataset with historical power generation, wind speeds and wind directions, for the time period from July 2009 to December 2010.

**Data Cleaning Process**

*Reading the files*

The raw dataset is comprised of several csv files. The pandas library is imported as **pd**, and used to import the csv files using **pd.read\_csv**.

*Converting dates and times to DateTime objects*

The dataset features a ‘date’ column, featuring dates and times in integer (int64) format as YYYYMMDDHH, where YYYY = year, MM = month, DD = day, and HH = hour. To be more useful, these values are converted from int64 into [DateTime](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.to_datetime.html) objects using **pd.to\_datetime**. Furthermore, these DateTime objects are converted into a standardized [ISO 8601](https://en.wikipedia.org/wiki/ISO_8601) format for convenience. A function called **convert\_to\_iso** is defined, to convert a datetime in int64 format to a DateTime object in ISO 8601 format. The function **convert\_to\_iso** is applied to the ‘date’ columns in the dataset.

*Creating a column for modified dates, ‘mod\_date’*

In addition to the ‘date’ column, the dataset features an ‘hors’ column, featuring hour values in int64 format. The values in ‘hors’ ranges from 1 to 48, representing the number of hours-ahead being forecasted. For example, let’s say it is July 1st, 2009 at 12am. At this time and date, there is forecast data associated with hors=1. This forecast data thus refers to the forecast for the following time:

(July 1st, 2009, 12am) + (1-hour-ahead forecast) = July 1st, 2009, 1am

Thus, a new column, ‘mod\_date’, is created to feature the DateTimes that include both the original datetimes (‘date’), and the hour-ahead forecasts (‘hors’).

‘date’ + ‘hors’ = ‘mod\_date’

In the equation above, values for ‘date’ are in DateTime format. Values for ‘hors’ are originally int64 but are converted to timedelta format using **pd.to\_timedelta**.

*Creating a column for forecast categories, ‘forecast\_cat’*

Forecasting data is split into the following four categories:

* Category 1: 1-hour to 12-hour ahead data
* Category 2: 13-hour to 24-hour ahead data
* Category 3: 25-hour to 36-hour ahead data
* Category 4: 37-hour to 48-hour ahead data

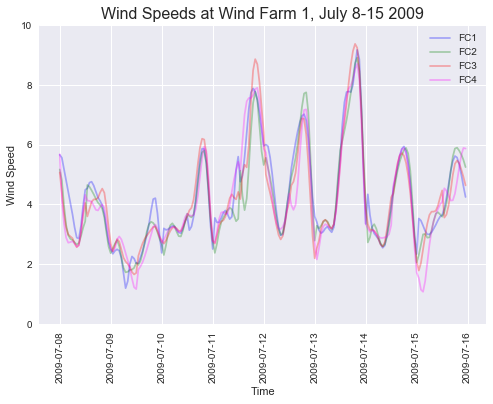
Thus, a new column ‘forecast\_cat’ is created, featuring a value ranging from 1 to 4 for each data row, corresponding to the appropriate forecast category shown above. Boolean selection is achieved using [.loc](http://pandas.pydata.org/pandas-docs/version/0.22/generated/pandas.DataFrame.loc.html).

*Merging* *wind data with training data*

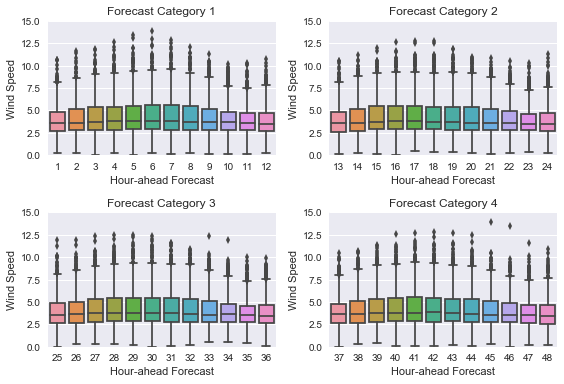
Then, the wind data is merged with the wind power generation training data using [pd.DataFrame.merge](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.merge.html). The specific merge method is ‘left’, specified by the ‘how’ argument, and is analogous to a LEFT OUTER JOIN [SQL Join](https://www.codeproject.com/Articles/33052/Visual-Representation-of-SQL-Joins). Thus, the left outer join in this situation returns all of the rows for which there is both wind speed, direction data and wind power generation data available. Rows that have wind power generation data but no wind speed, direction data available are not included.

**Exploratory Data Analysis [**[**Jupyter Notebook**](https://github.com/jon-lo/wind-energy-forecasting/blob/master/Wind%20Energy%20Forecasting%20Data%20Story.ipynb)**]**

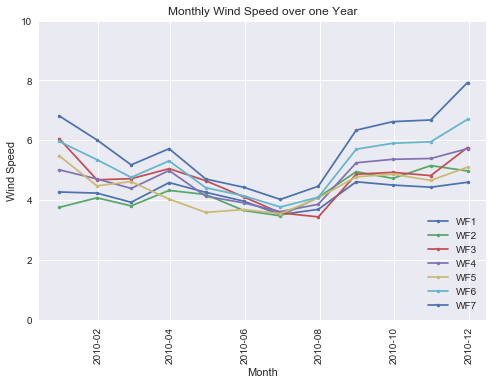
Plotting wind speed over the course of one week for each wind farm suggests that wind speeds tend to follow a diurnal curve with lower wind speeds at night and higher wind speeds in the afternoon to early evening.



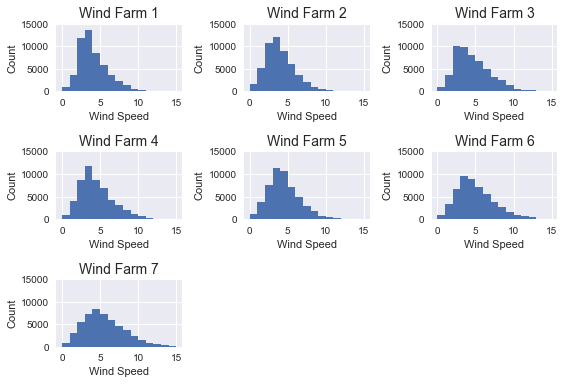
A visual comparison of the wind speeds of different forecast categories suggest that shorter and longer forecasts for wind speed are of similar magnitude. Inferential statistics can be used to confirm if there is a significant difference between wind speeds of different forecasts.



A visualization of mean wind speeds by month indicates that wind speeds tend to be lower during the summer and higher in the fall (September to December).

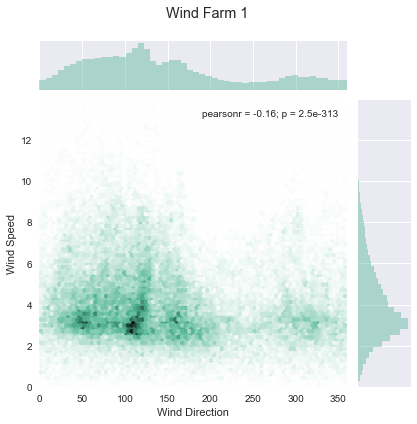


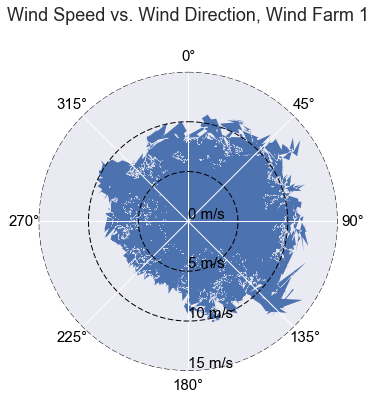
The distributions of wind speeds at all wind farms were unimodal but the shape of the distribution was slightly different for each wind farm. This may be due to geological or other reasons that are not included in the available data. The wind farms are likely independent from each other (the data collected at one wind farm has no effect on the data collected at another wind farm).



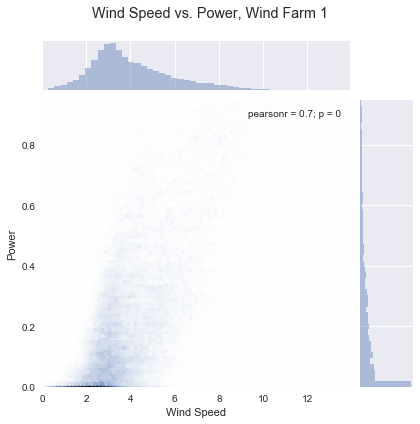
Across all wind farms, there was either no or weak correlation between wind speed and direction. Some wind farms had concentrated winds in one particular direction (Wind Farm 1, Wind Farm 4) whereas other wind farms had concentrated winds in two directions (Wind Farm 2, Wind Farm 6, Wind Farm 7). Wind Farm 3 and Wind Farm 5 had more winds from 0-250 degrees and less wind from 250-360 degrees.

These differences in distributions may be due to geological or other reasons that are not included in the available data. The wind farms are likely independent from each other (the data collected at one wind farm has no effect on the data collected at another wind farm). See figure below for wind speed versus wind direction at Wind Farm 1. For more figures for other wind farms please see [here](https://github.com/jon-lo/wind-energy-forecasting/blob/master/Wind%20Energy%20Forecasting%20Data%20Story.ipynb).

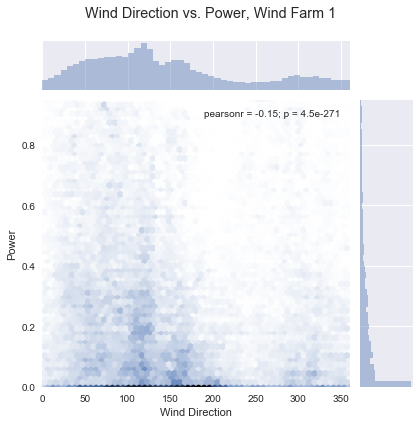




Across all wind farms, there was a strong correlation between wind speed and power, with Pearon's r-values ranging from 0.67-0.81. This correlation is expected given that the theoretical power output of a wind turbine generator is proportional to the cube of the wind speed. See figure below for wind speed versus power at Wind Farm 1. For more figures for other wind farms please see [here](https://github.com/jon-lo/wind-energy-forecasting/blob/master/Wind%20Energy%20Forecasting%20Data%20Story.ipynb).



Across all wind farms, there was a no correlation between wind direction and power. See figure below for wind direction versus power at Wind Farm 1. For more figures for other wind farms please see [here](https://github.com/jon-lo/wind-energy-forecasting/blob/master/Wind%20Energy%20Forecasting%20Data%20Story.ipynb).



**Inferential Statistics [**[**Jupyter Notebook**](https://github.com/jon-lo/wind-energy-forecasting/blob/master/Wind%20Energy%20Forecasting%20Inferential%20Statistics.ipynb)**]**

Recall that there are four forecast categories:

Forecast Category 1: 1-12 hours ahead

Forecast Category 2: 13-24 hours ahead

Forecast Category 3: 25-36 hours ahead

Forecast Category 4: 37-48 hours ahead.

A visual comparison of the wind speeds of different forecast categories suggest that they are similar across forecast categories. Inferential statistics are used to compare the wind speeds of different forecast categories to ascertain whether or not there are any significant differences between shorter and longer forecasts. In other words, we are interested to see if the wind speeds of one forecast category are different from the wind speeds of other forecast categories. For each wind farm, we can compare the mean wind speed of Forecast Category 1 (μ1) to the mean wind speed of Forecast Category 2-4 (μ234). The null hypothesis in this case would be that the mean wind speed of Forecast Category 1 is equal to the mean wind speed of Forecast Category 2-4:

μ1 = μ234

The alternate hypothesis would then be that the mean wind speed of Forecast Category 1 is not equal to the mean wind speed of the other Forecast Categories:

μ1 ≠ μ234

Similarly, we can test the fullowing null hypotheses:

|  |  |
| --- | --- |
| μ2 = μ34 | Comparing forecast category 2 to longer forecast categories (3, 4) |
| μ3 = μ4 | Comparing forecast category 3 to longer forecast categories (4) |
| μ4 = μ123 | Comparing forecast category 4 to shorter forecast categories (1, 2, 3) |
| μ3 = μ12 | Comparing forecast category 3 to shorter forecast categories (1, 2) |
| μ2 = μ1 | Comparing forecast category 2 to shorter forecast category (1) |

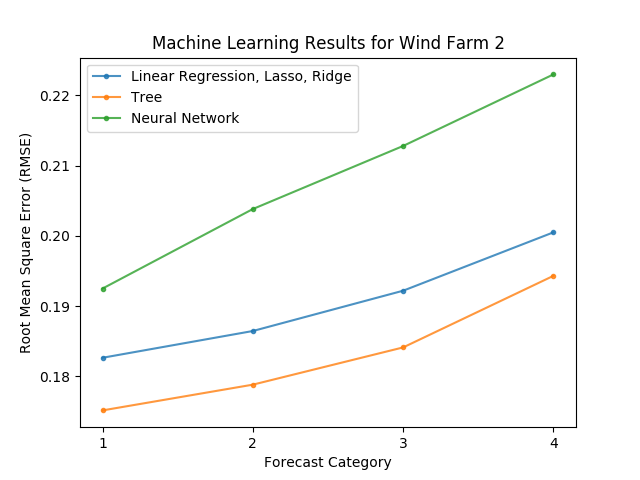
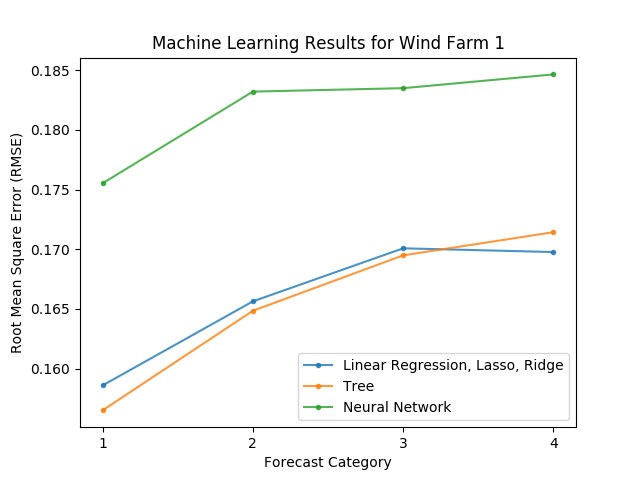
See table below for results of two-sample z-tests for the above null hypotheses. We assume a critical value of Z = 1.96 for an α =0.05 level of significance. Z-scores that were greater than 1.96 or -1.96 are highlighted in red, below.

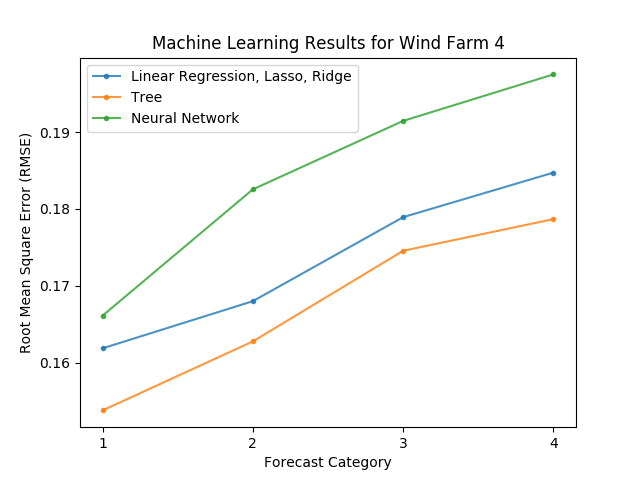
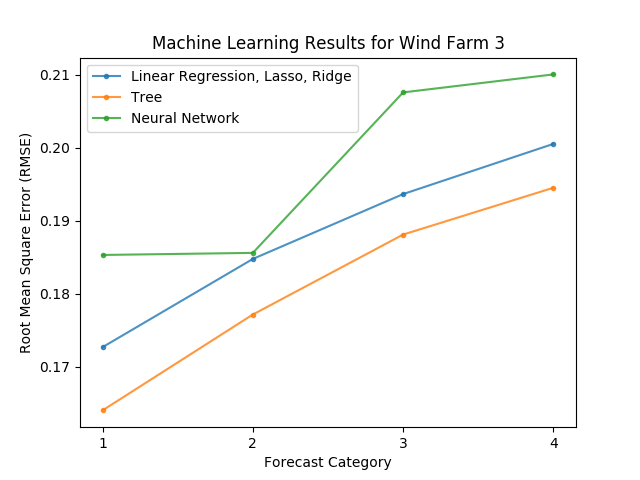
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Forecast Category Comparison** | **Z-score for Difference in Mean Wind Speed between Forecast Categories at each Wind Farm (WF)** | | | | | | | **Average Z-Score** |
| **WF1** | **WF2** | **WF3** | **WF4** | **WF5** | **WF6** | **WF7** |
| μ1 = μ234 | 1.799 | 2.973 | 2.088 | 2.845 | 2.911 | 2.322 | 2.015 | 2.422 |
| μ2 = μ34 | 0.625 | 1.700 | 0.459 | 1.400 | 0.790 | 0.059 | -0.110 | 0.703 |
| μ3 = μ4 | 0.137 | 0.310 | 0.448 | 0.897 | 0.411 | 0.639 | 0.629 | 0.496 |
| μ4 = μ123 | -1.013 | -2.077 | -1.284 | -2.351 | -1.684 | -1.331 | -1.138 | -1.554 |
| μ3 = μ12 | -1.191 | -2.379 | -1.027 | -1.774 | -1.667 | -0.776 | -0.522 | -1.334 |
| μ2 = μ1 | -1.117 | -1.463 | -1.435 | -1.520 | -1.932 | -1.860 | -1.707 | -1.576 |

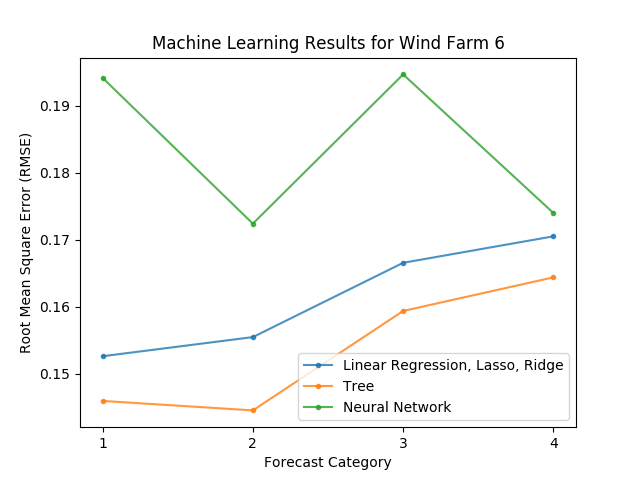
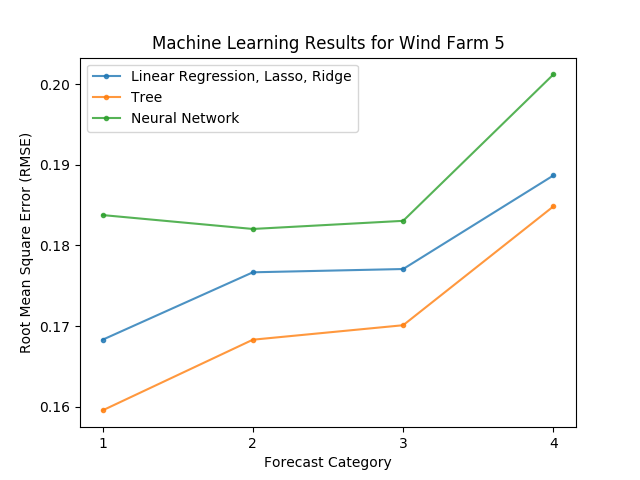
Comparing forecast categories to each other, it was found that mean wind speeds of Forecast Category 1 (1-12hrs ahead) tended to be significantly different from longer Forecast Categories (13-48hrs ahead), with 95% confidence. Z-scores for this comparison were in excess of 1.96 (95% confidence threshold) for all wind farms except Wind Farm 1, which had a Z-score of 1.798917, which still corresponds to 92.8% confidence. Another interesting conclusion is that Z-scores seem to vary across all wind farms, with no apparent pattern. This may be due to geological or other features not accounted for in the data set.

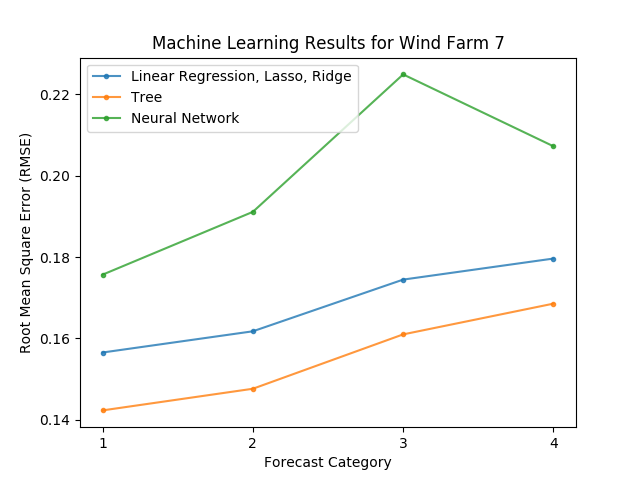
**Machine Learning Models**

Predicting wind power production based on wind speed and wind direction is a regression problem. We model a dependent variable (wind power) based on two independent variables (wind speed and wind direction). In total, five machine learning models were trained and tested on the data: Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression and Neural Network Regression. Root Mean Square Error (RMSE) was the metric used to evaluate prediction accuracy; an RMSE closer to a value of 0 is indicative of higher predictive accuracy. Generally, Decision Tree Regressors performed best, producing the lowest RMSE values across all wind farms. Neural Network Regressors performed worse. Also, Lasso Regression and Ridge Regression produced virtually the same results as standard Linear Regression. This makes sense given that there are only two independent variables at play; Lasso and Ridge are useful when there are many more feature variables involved, and it is necessary to incorporate variable selection. Linear/Lasso/Ridge Regression generally did not perform as well as Decision Tree Regression but always outperformed Neural Networks.









**Conclusion**

A variety of Machine Learning Models were trained and tested on wind speed, wind direction and wind power production data. Prediction accuracy was evaluated using Root Mean Square Error (RMSE). Decision Tree Regressors performed best as compared to Standard Linear Regression, Lasso Regression, Ridge Regression and Neural Networks. Therefore, Decision Tree Regression is recommended for future predictions of wind speed, wind direction and wind power production data at the various wind farms.