PROJECT EXERCISE #2

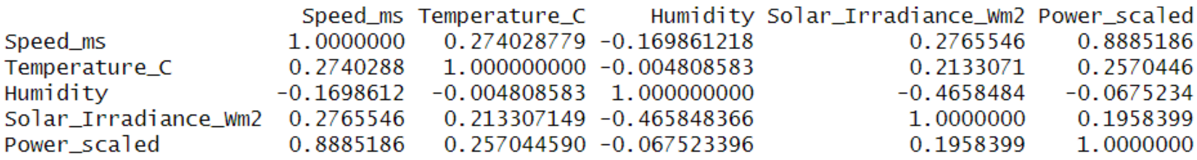
Data Bosses
Anne Christine Domercant
Tina Truc Hoang
Tasnim Quayum

Contribution:

Anne Christine	Knn, trial model, code merge, report
Tina Hoang	Polynomial, Random Forest, report
Tasnim	Linear model 1, exponential model, report

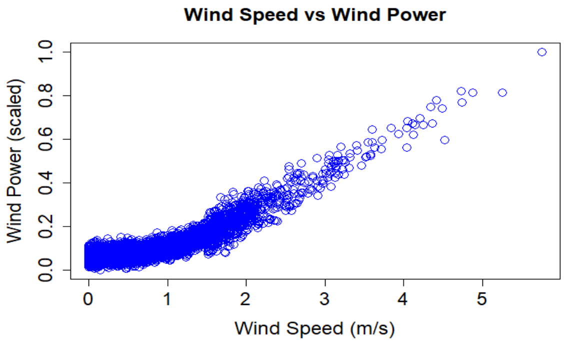
Course: Industrial Informatics (14:540:485:01)
Instructor: Dr. Ahmed Aziz Ezzat
Rutgers, The State University of New Jersey

Task 1: Preliminary analysis:



```
                     Speed_ms Temperature_C     Humidity Solar_Irradiance_Wm2 Power_scaled
Speed_ms            1.0000000   0.274028779 -0.169861218            0.2765546    0.8885186
Temperature_C       0.2740288   1.000000000 -0.004808583            0.2133071    0.2570446
Humidity           -0.1698612  -0.004808583  1.000000000           -0.4658484   -0.0675234
Solar_Irradiance_Wm2  0.2765546   0.213307149 -0.465848366            1.0000000    0.1958399
Power_scaled        0.8885186   0.257044590 -0.067523396            0.1958399    1.0000000
```

When performing our analysis of the data and its relationships, we are able to make conclusions about different variables. When looking at the correlation values of wind power versus other environmental factors, speed has the highest correlation at a value of 0.8885186. Because of the high correlation value, the basis of our regression model will use the relationship between speed and power. The high correlation also gives strong support for attempting a linear model, as a high correlation shows a high linear relationship. When examining the plot of wind speed vs wind power for the data observed, we are able to visually determine that a linear relationship or exponential relationship may be sufficient to establish the relationship between the two variables, which is the reasoning behind why those two models are attempted. Another determination made from examining the graph is the concept of a cut-off wind speed. When visually inspecting the plot, the wind cut-off speed appears to be approximately from 0 to 0.5 m/s, which is what we use in the "trial model.



There is a low correlation between temperature and power, and solar irradiance and power at correlation values of 0.25and 0.19 respectively, so we also considered using these environmental factors as part of our models. One find that we found during our analysis is that there is a strong correlation between wind power and its lagged values at a value of 0.8199444. However, because the test data does not have time associated with the points, we are unable to use this relationship when creating our regression model.

Task 2: Model building:

Regression Model	Train_RMSE	Test_RMSE
Linear	0.042	0.042
Exponential	0.064	0.073
Polynomial	0.029	0.029
Random Forest	0.015	0.031
Knn	0.025	0.032

When creating our models, we decided to use k-fold cross validation as our method. Since our dataset is large, we felt it would be the ideal method to help reduce the arbitrary factor of selecting training and testing data sets. We used a k value of 5 when developing our models.

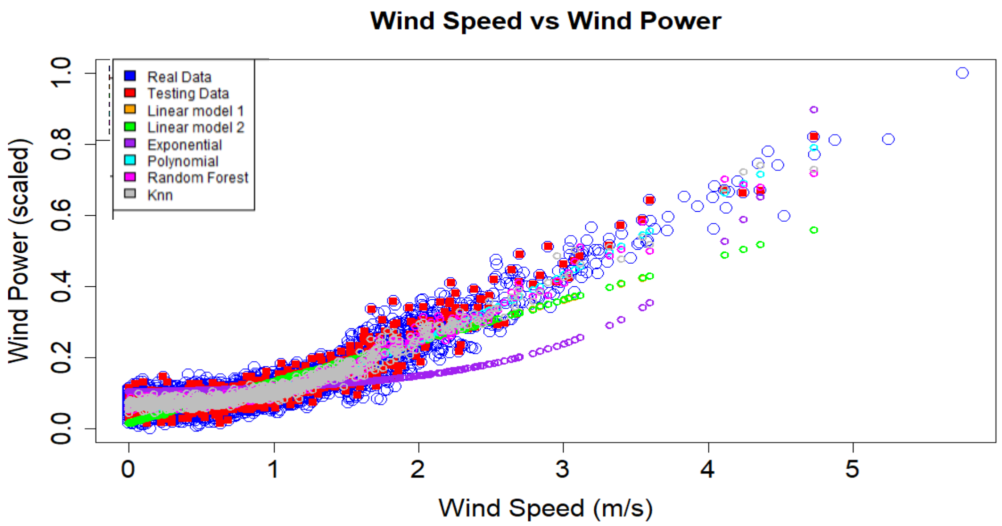
i) The first model attempted was the linear model between wind speed and wind power because of the high correlation values calculated when performing the initial analysis.

ii) The second model attempted was an exponential model. When visually examining the wind speed vs wind power plot it appears to be exponential. The exponential model does not appear to represent the model as well as the linear model.

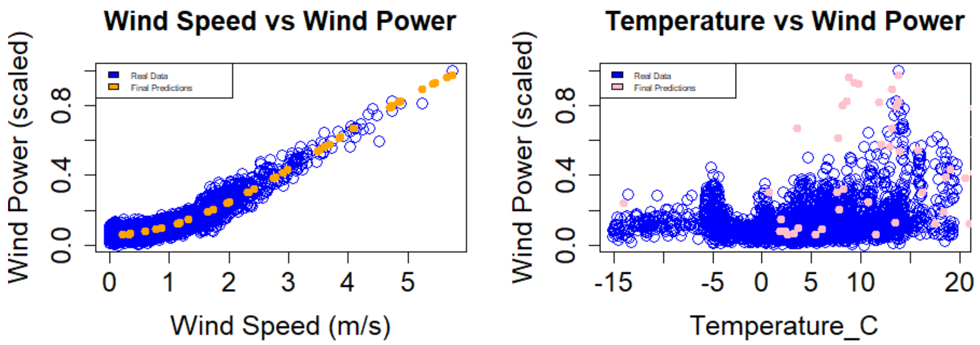
iii) The polynomial regression model is an extension of the linear regression that can capture nonlinear relationships between the environment variables and wind power. From task 1, we want to predict the wind power based on temperature and wind speed along with their second and third-order interactions.

iv) For random forest, we use temperature and wind speed as the predictors to fit the model. By including interactions between temperature and wind speed, we allow the model to capture potentially complex relationships between these two features and their joint effect on the target variable.

v) K-nearest neighbor was the final model tested. While there was success in using parametric approaches such as polynomial, our team was interested in investigating if a non-parametric approach would be most suitable to create a model that predicts wind power. To create the model, we use the variables temperature and wind speed.



Task 3: Final Prediction



Polynomial model:

```
new_pmod = lm(Power_scaled ~ Temperature_C + I(Speed_ms^2) + I(Speed_ms^3), data = df_reshuffled)
```

After observing the RMSE from five regression models from two datasets (one without cut-off, and another with cut-off), we compare the errors between training and testing data. Although Random Forest and Knn model performs better than Polynomial in Training data, their RMSE in testing data are higher than that of Polynomial model. Therefore, the most suitable model is Polynomial regression.