Tasneem Hossain

Topic:

**Challenge: Help ENGIE improve wind power production! By Engie**

ENGIE is a French multinational electric utility company, which is called one of the largest wind power producers of France and Belgium. It has 4553 MW installed throughout the world and is one the first wind power producer to have an installed capacity of 1730MW.

Wind Power plays a vital role in present and future climate changes because of its renewability. Therefore the development of wind power is put first on the concern of Engie. So they offered an open competition to public through challengedata.ens to improve wind power production meaning they want to optimize further. As the challenge goal statesthey want to detect abnormal divergence between expected production and actual production.

So formally, the main goal of this challenge is: to predict the active power produced by four wind turbines from several operational parameters; these parameters come from measurement sensors installed inside these turbines. While wind speed anemometers are customarily installed on the nacelle of each turbine, the related measurements are often considered to be of poor reliability for monitoring the operational performance; for that reason, the challenge focuses on predicting active power without considering wind speed data in the list of explanatory variables. (<https://challengedata.ens.fr/en/challenge/45/help_engie_improve_wind_power_production.html>)

The data set is provided, which contains the training input, output and the test input. Therefore the first thing I did was generated the data to see and evaluate the features and labels and target. The first thing I figured was that each attribute from the features or the X variables had 4 columns. So, I took only one column for each attribute, that reduced my columns to almost 24 from 78 columns. I also ignored the ID column from both X and Y (Features and Labels). The next was, the string value the Mac\_Code column, which is one of the most important key because it shows the turbine number (WT1, WT2, WT3, WT4). I replaced the values with number 1-4, so basically changed the string to numeric value because machine do not understand anything other than numeric values. I also scaled my features to reduce dominancy of any columns because I had values from -400 something to 20000, so dominancy could have hampered my data set a lot if I had not scaled it. I also removed the nan values (no values) from my dataset and converted all infinite values to finite. In addition, I also reduced my data set to 10% of actual dataset because the actual dataset has 6 hundred thousand instances which is not possible to run in my machine.

Now that my dataset is ready, I plotted just a plain y graph to see how the values are standing and to make sure if its regression or classification. Even though when I checked the dataset initially I figured it out it will be regression because the output is continuous but just to be surer I plotted the graph.

Since my problem falls under regression, I chose DecisionTreeRegressor as my first method, and calculated the mean and std scores and the checktraintest scores. The scores came out pretty good, with low gap between train and test scores. This shows that the model does not suffer from high variance which is a plus point and in this context, this is very important because the error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict. So, the goal of my challenge perfectly fits this criterion as I am trying to find the actual production and the estimated production. In addition I repeated the same procedures with Random Forest Regressor and Extra Random Forest, firstly to check if the scores are changing drastically or not but the scores came with more reduced gap between train and test, which is expected and is better; then secondly to see the pattern for the feature importance, in other words to see which features commonly came out as important for the three model. This is needed for my evaluation on future implementation because while doing this challenge I researched on which factors or features mainly affect the production scale and learned that location, temperature affects a lot.

The next method I chose was Linear Regression to get a perfect hold of the r2score and my r2 score came out to be 0.89 which is almost close the professional participants currently ranking. It also proves that the model explains all the variability of the response data around its mean.

Lastly, I used LinearSVR and reported the score, coefficient and intercept and enjoyed the more flexibility in the choice of penalties and loss functions.

All in all, if I have to choose 1 best algorithm it will be difficult because I think for my problem, and the way I approached it , the main 3 algorithms, Regressors, Linear Regression , LinearSVR they all had good performance in different aspects. For DecisionTreeRegressors,RFT,ERT, (I have categorized these 3 algorithms as one, as I used them in parts and steps to pull out the important features) they performed really well in pin pointing the important features that might help in present production scale and also in future implementations and also the scores were appealing. For Linear regressions, the r2 score reported was good enough to stabilize the problem alone and bring it to par. For LinearSVR the flexibility in the choice of penalties and loss functions made it through.

In conclusion, I would say for future implementation, such enormous features are not necessary and the most promising one from my research and through my algorithm was Hub Temperature and Converter Speed.

Works Citation <https://challengedata.ens.fr/en/challenge/45/help_engie_improve_wind_power_production.html>