SMARTWATT INTELLIGENCE
Smart Energy Applications

Rodrigo Pinto
22/01/2018

# Wind Power Forecasting

## Introduction

The goal of this project was to develop models capable of forecasting the wind power generated by two different eolic parks. The provided data sets contain meteorological forecasting and in-house collected data.
This project was divided in five parts:

1. Data Cleansing
2. Feature Engineering
3. Exploratory Data Analysis
4. Feature Selection
5. Data Modeling

## Data Cleansing

In this part we checked the data for missing values and to see if all the data is in the right format. We found that the following variables were completely missing: **measured_wind_gust**, **lightning_risk** and **wind_gust**, therefore, we removed them. For the park with id 51 we also found that the **Power** was missing for the first 1454 rows, so we removed them. The rest of the data was in the desired format, no changes were required.

## Feature Engineering

From the variable **predictiondate** we extracted two new features: **hour** and **month**.

## Exploratory Data Analysis

During this phase we analysed different types of charts in an attempt to better understand our data (all the charts are available on the IPython notebook).

By observing the box plots of the power for each hour and month we observed that there are certain hours and months during which there is an increase in Power, this way validating our choice of extracting the hour and month as a feature.

By analysing the scatter plot of the measured wind speed vs power we observe a linear relationship, but it is worth mentioning that there is a minimum wind speed required for the park to generate power. There is also a point where the power remains constant even with an increase in the wind speed. This happens because each park is only capable of generating a certain amount of power. With this information we know we have to define a maximum value for the output of our models.

By analysing the correlation between each feature we found that there were no features extremely correlated.

## Feature Selection

We used recursive feature elimination and a 10-fold validation to rank the features by importance and to select the best number of features.
For the park with id 04 the following features were found to be optimal: **wind_speed**, **wind_direction**, **temp**, **hour**.
For the park with id 51 the following features were found to be optimal: **wind_speed**, **wind_direction**, **temp**, **hour**, **month**.

## Data Modeling

The first step for modeling our data was to perform a 85/15 split on the training set to obtain two sets: training and test. The training set, along with a 5-fold validation technique, was used to select and tune the best models. The test data was never used in the optimization process, it was used just to check the true performance of the model on out-of-sample data.
The features on the training set were normalized to have zero mean and unit variance. We used the computed means and variances to normalize the test set as well. The target variable was also scaled between 0 and 1.
Since this is a regression problem, we chose the RMSE as the metric to evaluate the models performance.

We initially selected a set of candidate models: **Lasso; Elastic Net; Kernel Ridge; Gradient Boosting Regressor; Support Vector Regressor; XGBoost Regressor (XGB); Light GBM Regressor (LGBM).**

For those candidate models, we performed a 5-fold validation on the train set, and concluded, that for the both parks, the **XGB** and the **LGBM** tend to perform better. Therefore, we selected the **XGB** and the **LGBM** models for the hyper-parameter phase.

For each model, the hyper-parameter tuning was performed using Bayesian Optimization.
After we fine tuned each model, we experimented stacking them in an attempt to reduce the RMSE, that is, we averaged their outputs. In table 1, we show the results obtained during the 5-fold validation phase with the tuned models and with the stacked model. We found that for both datasets, the stacked model performed the best. Table 2 shows the result of the stacked models on test set. We believe that the models performed well on the test set, making them ready to deploy.

Table 1: Validation results.

| Id: 04 | | Id: 51 | |
|---|---|---|---|
| Model | Average RMSE | Model | Average RMSE |
| XGB | 0.130603 | XGB | 0.104178 |
| LGBM | 0.129097 | LGBM | 0.105558 |
| Stacked(XGB+LGBM) | **0.128902** | Stacked(XGB+LGBM) | **0.101308** |

Table 2: Test results.

| Id: 04 | | Id: 51 | |
|---|---|---|---|
| Model | RMSE | Model | RMSE |
| Stacked(XGB+LGBM) | 0.120047 | Stacked(XGB+LGBM) | 0.129326 |

## Attachments

All the files used for this project are in the repository:
https://github.com/rodrigomfw/SMARTWATT