DS 5110 Project Proposal: Electrical Price Prediction

Weipeng Zhang, Soham Shinde, Ziyue Wang, Mingni Luo

Summary

Forecasting in energy markets is identified as one of the highest leverage contribution areas of Machine Learning towards transitioning to a renewable based electrical infrastructure.

The dataset contains hourly data for Energy Generation, Consumption, Pricing and Weather from years 2014 to 2018 for Spain. There are two csv files: *energy_dataset.csv* and *weather_features.csv*. Both of them are time-series data having sizes (35064 x 29) and (178396 x 17) respectively. [1]

The goal of the project is to predict the electrical price by the time of the day using generation, consumption and weather data and also analyze what features influence electrical price the most in Spain. After data wrangling and feature engineering, a regression model would be implemented which performs the best in accordance with the project goal.

Proposed Plan

Exploratory Data Analysis will be done in order to examine the structure of the dataset and to find potential relationships between the features. The process of Model Selection will be carried out by selecting the best one by comparing and validating with various parameters and choosing the final one.

Methods like Linear Regression, Ridge and Lasso Regression, Random Forest Regression and Support Vector Regression will be used. Different variations of the above methods will be implemented to identify the one that performs the best for the given dataset.

With feature engineering and more complex models along with boosting methods, such as LightGBM and XGBoost, the gap between the predictions and the actual values will be narrowed.

Preliminary Results

The feature correlation matrix (Fig.1) tells us that the feature '*price actual*' is correlated with '*generation fossil gas*', '*generation fossil hard coal*', '*generation hydro pumped storage consumption*' and *'wind_speed*'.

In order to get a rough estimate of how far our model can get, the dataset was simplified by removing categorical features and then was used to train several regression models. Mean-Squared Error(MSE) of test data was used to evaluate the performance of the trained models. The MSEs of models are around 200. By visualizing the predicted result of ensembling LR, Lasso, Ridge, SVR and RF, (Fig.2) we can conclude that though the precision is not that high, these models are able to predict the trend of electrical price correctly.

References

[1] https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather

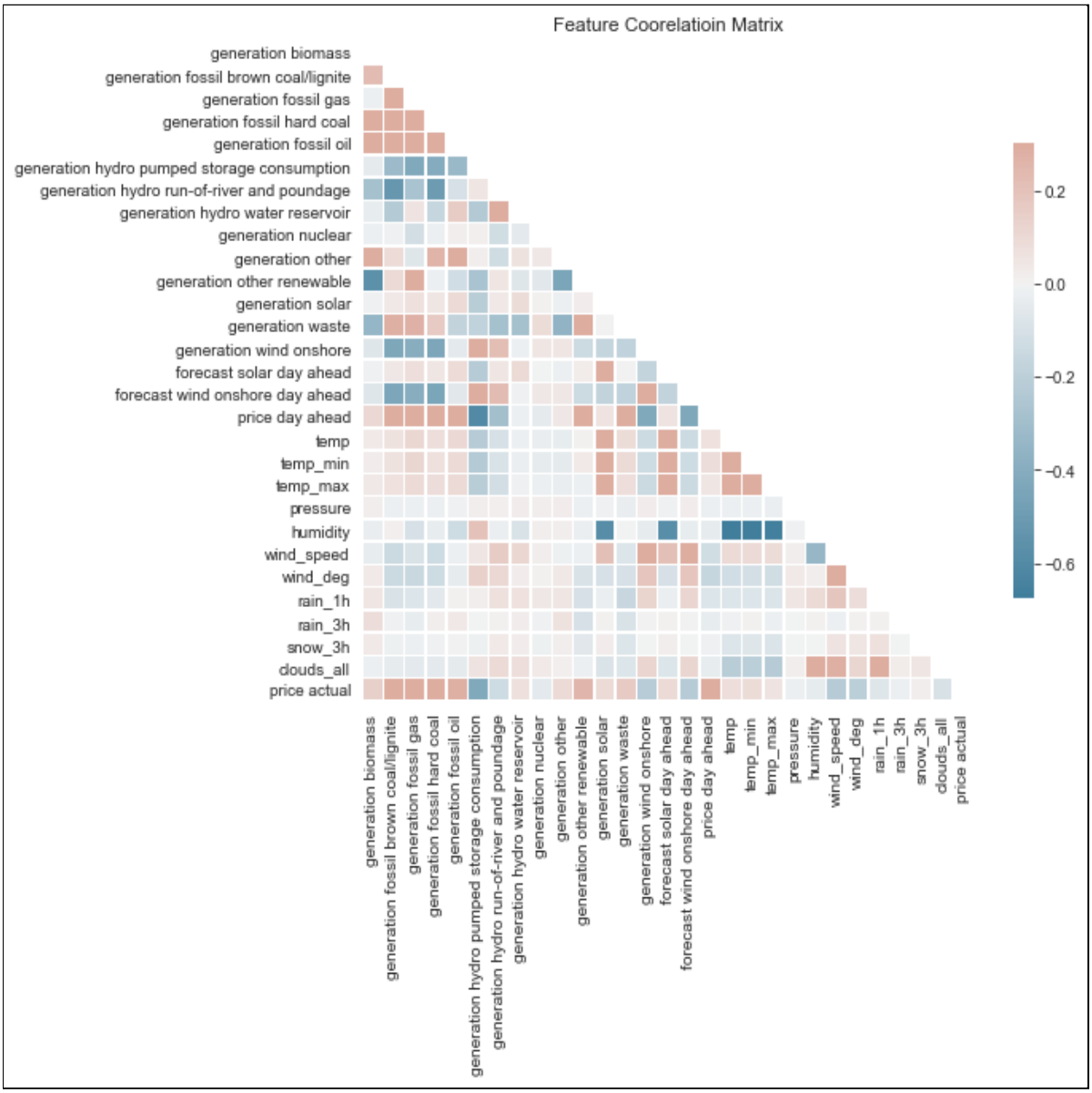
Feature Coorelatioin Matrix
generation biomass
generation fossil brown coal/lignite
generation fossil gas
generation fossil hard coal
generation fossil oil
generation hydro pumped storage consumption
generation hydro run-of-river and poundage
generation hydro water reservoir
generation nuclear
generation other
generation other renewable
generation solar
generation waste
generation wind onshore
forecast solar day ahead
forecast wind onshore day ahead
price day ahead
temp
temp_min
temp_max
pressure
humidity
wind_speed
wind_deg
rain_1h
rain_3h
snow_3h
clouds_all
price actual
0.2
0.0
−0.2
−0.4
−0.6

Fig.1

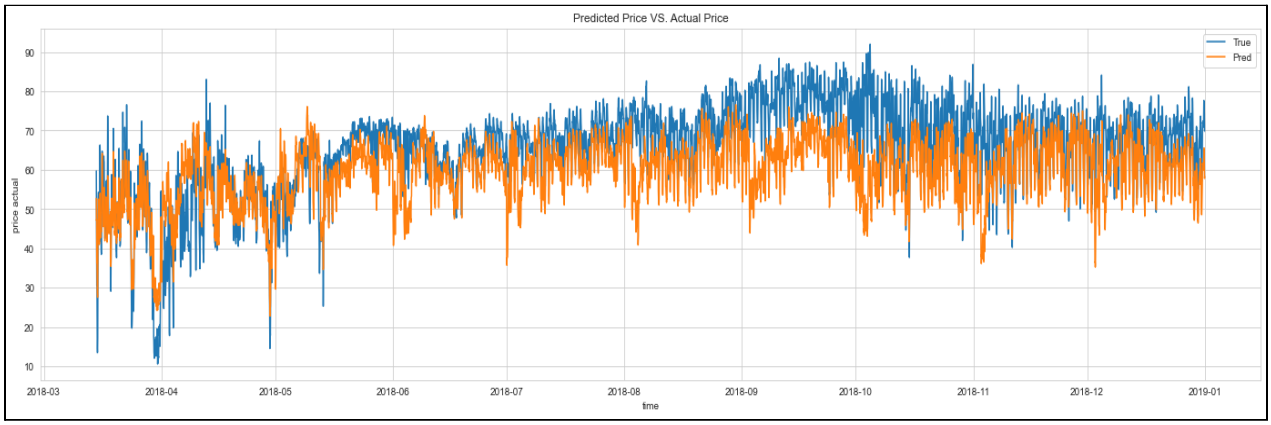
Predicted Price VS. Actual Price
True
Pred
price actual
time
2018-03
2018-04
2018-05
2018-06
2018-07
2018-08
2018-09
2018-10
2018-11
2018-12
2019-01

Fig.2