Analytics Vidhya Jobathon

Assessment

The Approach

By Nikhil Londhe

Thank you Analytics Vidhya for providing such an opportunity. Thank you to the teachers, the institutions, organisations, colleagues, classmates, friends, family, the people who have been a part of the journey.

Overview

- "Green" wants to achieve the target of 95 percent renewable energy usage of the total energy usage.
- The data for past nine years is provided.
- The prediction for energy consumption for the next three years is to be made, for "Green" to understand about the energy scenario.
- The Dataset is related to the Time-Series Data category.
- The time shall be separated hourly.

Brief on the approach

The approach I followed here, in brief, is as follows.

- The data preparation was to be done, hence, Data Pre-processing, and Feature Engineering, was performed in the initial phase.
- Model Building was performed then, where a suitable model was selected, and further tailored according to the task.
- Then, the selected model was fit on the train data, and predictions were obtained for the test data.
- In the final phase, a file was made with the IDs and results arranged.

Data pre-processing and Feature Engineering

The following points were met with while programming.

- The "datetime" column has Dtype "object", it shall be converted to DateTime datatype.
- The column "energy" has null values.
- The number of null and non-null values from the "energy" column were obtained.
- There are 1900 null values in the column.
- The corresponding row of every null value was obtained.

- Conversion of the datatype of the "datetime" column to DateTime format was to be done.
- 'parse_dates' parameter was used.
- Made a table with only null values from the "energy" column, and rest columns as it is.
- The output showed that the null values are somewhat randomly occurring regardless of the date and time features.
- Hence, the null values shall be removed in this case, instead of any other alternative.

- This shall preserve the integrity of the dataset, and be of utility for predictions based on detailed features such as month, day, hour.
- All the values are made "non-null" after dropping.
- The energy consumption can depend on the year, month, day, hour of the day, minute, second, more or less.
- The dependency of consumption on seconds of the minute of the hour can be chosen to be negotiated based on the model building time.

- Since the test data contains time only up to the minute mark, and not the seconds hence, the train data shall contain accordingly, however.
- The day of the week - this factor is considered.
- Dropped the "datetime" column, as to make the data succinct.
- Bifurcation of attributes - a.k.a. features, independent variables, feature columns, X - and response variable - a.k.a. dependent variable, target variable, y.
- As the attributes each have different range of values from one another, scaling shall be done to bring them on one scale.

- Data preparation is completed.

Discovery of the previous phase

- Missing values – NaN, were dealt with.
- The date and time attributes were handled.
- Apart from this pre-processing, feature engineering was provided.

The Model

- The task is a regression task.
- Model is fitted, the model used is LGBMRegressor, from lightgbm.
- Test Dataset is incorporated.
- Test data has been prepared accordingly.
- Accordingly, test data has been scaled.
- From the model, predictions are obtained.
- For submission file, file, as per format, is made.

Path to the Model

- There are Machine Learning algorithms which I have worked with, including K Nearest Neighbours, Linear Regression, Polynomial Regression, Linear Discriminant Analysis, Support Vector Machines, Decision Tree, Random Forest, Gradient Boost, Bootstrap, Ensemble, XGBoost, lightgbm.
- The lightgbm model has shown to be performing good, and was selected as the model in this case.
- Further, as the task was a regression task, LGBMRegressor was selected.

This is about the approach, for now, thank you.