Project Report
Predict Energy Behavior of Prosumers

Mattias Kimst, Johhana Laane
Repository - https://github.com/MattiasKimst/enefit-prosumers-kaggle-ids

Task 2

Our project is from the Kaggle competition arranged by Enefit. Enefit, as one of the biggest energy companies in the Baltic helps customers to find the best green energy solution. Currently Enefit is facing an imbalance problem, meaning it doesn't have accurate results on how much clients produce and consume and therefore more energy is produced than needed and consumers won't use all the energy. The increasing number of prosumers and their unpredictable energy use is causing logistical and financial problems for energy companies. This competition aims to tackle the issue of energy imbalance.

Our business goal is to improve the accuracy of energy predictions and consequently reduce the imbalance and associated costs. Business success criterias include:

- reduce the operational costs
- improve grid reliability
- integrate prosumers efficiently into the energy system

Currently, Enefit is attempting to solve the imbalance problem by developing internal predictive models and relying on third-party forecasts. However, these methods have proven to be insufficient due to their low accuracy in forecasting the energy behavior of prosumers.

Our inventory of resources include:

- Data provided by Enefit of prosumer energy behavior including 15 datasets
- Team consisting of 2 people
- Software which will be used is Jupyter Notebook online at kaggle.com and Github for storing notebooks, data and other files related to the project

Enefit has a wealth of data on prosumer energy behavior that can be used to develop more accurate predictive models. The requirement is to develop a predictive model using the provided Python time-series API. The final submission deadline is 31st January 2024. The code requirements are:

- CPU Notebook <= 9 hours run-time
- GPU Notebook <= 9 hours run-time
- Internet access disabled
- Freely & publicly available external data is allowed, including pre-trained models
- Submission file must be named submission.csv and be generated by the API.

The main risk is the potential for inaccurate predictions, which could lead to increased imbalance costs. The contingency plan would be to continue refining the model based on feedback and results.

Terminology:

1) Prosumers - consumers who also generate energy
2) Energy imbalance - refers to the situation where the expected energy use doesn't line up with the actual energy used or produced
3) Mean Absolute Error - a metric used to measure the average absolute differences between predicted values and actual values in a set of observations
4) Linear Regression - statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.
5) Decision Tree - machine learning algorithm used for both classification and regression tasks. It works by recursively partitioning the dataset into subsets based on the values of input features, leading to a tree-like structure where each node represents a decision based on a feature, and each leaf node represents the predicted outcome.
6) Random Forest - method based on constructing multiple decision trees during training and outputting the class or mean prediction of the individual trees for classification or regression tasks.
7) LGBM - Light Gradient Boosting Model - a framework for implementing gradient boosting algorithms. Gradient boosting is an ensemble learning technique where weak learners (typically decision trees) are trained sequentially, and each new tree corrects the errors made by the previous ones.

The cost involves the time and resources spent on developing and refining the predictive model. The benefits include reduced operational costs, improved grid reliability, and efficient integration of prosumers into the energy system.

The data-mining goal is to develop a predictive model that can accurately forecast the energy behavior of prosumers. Success criteria include:

- Mine the available data to identify patterns and factors that influence prosumer energy behavior
- Ensure the model scales well as the number of prosumers or the complexity of data increases.
- Develop a model that is interpretable, allowing stakeholders to understand the factors influencing energy behavior predictions.
- Ensure that the predictive model can integrate seamlessly with existing energy management systems or tools used by prosumers.
- The success of the data-mining effort would be measured by the accuracy of the predictive model, as evaluated by the Mean Absolute Error (MAE) formula.

Task 3

Gathering data

To predict the energy behavior of prosumers, we definitely need data on their energy consumption and production. We also need data on other factors that may affect their energy behavior, such as weather, seasonality, weather forecasts, location and prices. We need both historical data, the longer period the better and current data which the prediction would be based on.

Sufficient data is provided by the competition organizers, including data about prosumers capacities, their locations, electricity and gas prices, weather forecasts and historical weather . For some tasks we will use additional datasets generated by Kaggle collaborators, for example mapping location coordinates to counties.

Mostly the data provided is relevant, however it is given in different dataframes so that we need to join those dataframes to use for training a prediction model. Dataframes could be joined on dedicated columns in dataframes such as data_bloc_id. After we have joined the dataframes, we should drop all such columns that do not provide information on factors affecting energy production/consumption such as data_block_id. 6 dataframes are provided: client.csv, electricity_prices.csv, forecast_weather.csv, gas_prices.csv, historical_weather.csv, train.csv with a total size of 1.11GB. All those datasets will be used. Precise descriptions are given for the datasets, including descriptions for all columns: https://www.kaggle.com/competitions/predict-energy-behavior-of-prosumers/data.

Also, the locations in datasets given by organizers are expressed using longitude and latitude, which we need to map to counties, for that task we will use a csv dataset generated by a kaggle user MikeOMa:

https://www.kaggle.com/datasets/michaelo/fabiendaniels-mapping-locations-and-county-codes/

Describing data

The datasets are in .csv format. They vary on the number of rows because for example historical weather data contains measurements of weather multiple times an hour, electricity prices are given hourly, gas prices daily. While joining the datasets, such aspects should be considered. The datasets have enough information e.g rows to train a model on them.

Exploring data

Kaggle provides useful plots and statistical measures like Mean, Std deviation, quantiles, number of missing values, min, max for exploring the data, finding abnormalities and possible problems with data. While checking those we didn’t find any significant problems with data given by Enefit, the data quality seems to be good.

Verifying data quality

As stated in the previous answer, we didn’t find any significant issues with quality of data, all the required data is accessible, there aren’t many missing values or possible incorrect values, all data seems to be from trustworthy resources. We do not consider finding data from alternative resources except one dataset for matching coordinates to counties.

Task 4

Step 1: Understanding the problem and the data. Exploring the data using descriptive statistics, visualizations, and correlation analysis to get a sense of the distribution, relationship, and trend of the variables. (Member 1: 7.5h Member 2: 7.5h) => 15h

Step 2: Preprocessing the data. Transforming the data by joining the dataframes, dropping inappropriate columns, where necessary applying scaling, converting values to categorical etc to make the data suitable for modeling. Split the data into training and validation sets (test set is given separately) (Member 1: 3h Member 2: 2h) => 5h

Step 3: Building and evaluating the models. Choosing the models that are appropriate for this data, the time-series prediction, mainly XGBOOSt or LGBM. Choosing hyperparameters using cross-validation. Training the models on the training set and evaluating them on the test set using the MAE metric. (Member 1: 9h Member 2: 11h) => 20h

Step 4: Making improvements to the models based on results from the previous steps. Comparing the performance of different models and selecting the best one. (Member 1: 5h Member 2: 5h) => 10h

Step 5: Testing and submitting the model. Testing the final models on the test set and generating the prediction file in the format required by the competition. (Member 1: 1h Member 2: 1h) => 2h

Step 6: Making a poster and giving an overview of the most important results. (Member 1: 3h Member 2: 5h) => 8h