FORECASTING HOUSEHOLD HOURLY ELECTRICITY CONSUMPTION RATE

Introduction to Data Science (LTAT.02.002)

Patricia Kika Obinwanne

Wei-Chieh Wang

Chigozie Nkwocha

Contents

Business Understanding

Identifying the Business Goals

Background

With the ever-increasing complexities in the global economy from the invasion of Ukraine by Russia, electricity prices have skyrocketed and consumers around the world are looking for options to reduce electricity costs. In particular, this invasion has put European electricity consumers, especially the Baltic nations like Estonia, in a peculiar situation with increased uncertainties about sustained availability and a high cost of energy. As a result, electricity consumers are now seeking better ways of reducing electricity costs and conserving energy.

Doing this would ultimately translate to a reduced environmental footprint since electricity contributes to the carbon footprint released into the environment. One way would be to seek an alternative source of energy through the use of solar panels and other various forms of electricity generation to reduce costs. A major concern is that seeking alternative energy sources would lead to the business losing money from more customers withdrawing their patronage. Hence, there's a need to step in. The major question that needs to be answered is this: "How can Enefit as a business step in to turn this into a win-win situation for both her and her customers? How can they help households reduce their energy costs and environmental footprint and still be in business?"

One way would be to optimise their energy usage by developing a predictive model that is capable of forecasting the electricity consumption for a household per hour of the day. Optimising their energy usage will enable them to reduce electricity costs by making them aware of their energy usage per hour of the day and in turn, they can control the smart devices in the home. Accurately forecasting household-level energy consumption is a critical prerequisite for more sustainable energy usage and this could be a game changer in making available more disposable income for households for other purposes.

Business goals

The main goal of this project work is to develop an energy consumption machine learning model that is capable of predicting electricity consumption for a single household for the next

seven days. We intend that this project will enable consumers to understand their electricity usage, identify the hour(s) of the day when usage is high or low, and control usage.

Business success criteria

A success criterion of this project is that in the end, we can get very close to the actual hourly consumption (reduced errors), measured in terms of the mean absolute error (MAE). We need to get as close to the actual consumption rate as possible to determine the model's reliability in helping to save money for households, reduce carbon footprint in the environment and keep Enefit in business.

Inventory of resources

The project will be undertaken by a team of 3 members. We have available a historical dataset containing a household's hourly electricity consumption used between Sept 1st, 2021 to 24th August 2022. This dataset also includes the weather and electricity price. A holdout test set is kept to be used for forecasting. We also have Google Colab and Python data wrangling and modelling packages such as pandas, matplotlib, numpy, scikit-learn, and so on, to clean and visualise data and build and test our model.

Requirements, assumptions and constraints

The data is provided by Enefit and hosted publicly on the Kaggle competition page and it is assumed that appropriate access has been granted. The project assumes there is no immediate drastic change in climate conditions. It is also assumed that the hourly prices per electricity consumed for the next seven days are static. It is expected to be completed in approximately three weeks.

Risk and contingencies

In-person meetings for discussion pose a huge problem since the group members belong to different departments and have varied schedules, and as a result, a Whatsapp group was created. To synergise collaboration, a Google Colab file has been created and each member granted access to modify parts of the project but with the instruction that proper documentation is made for easy understanding to the other members.

Terminology

Below data terms are used;

- temp - Air Temperature (°C)
- dwpt - The dew point in (°C)
- rhum - The relative humidity in per cent (%)
- prcp - The one-hour precipitation total in mm
- snow - The snow depth in mm
- wdir - The wind direction in degrees (°)
- wspd - The average wind speed in km/h
- wpgt - The peak wind gust in km/h
- pres - The sea-level air pressure in hPa
- coco - The weather condition code
- el_price - the electricity price in Estonia on that hour (€/kWh)
- consumption - the electricity consumption (kWh)
- mae - Mean Absolute Error
- rmse - Root mean squared error
- mape - Mean Absolute Percentage Error
- ARIMA – AutoRegressive Integrated Moving Average

Costs and benefits

Not applicable

Defining Data Mining Goals

Data-mining goals

We aim to use data mining and visualisation to understand the electricity consumption of the household. We hope to determine the relationship between electricity consumption, price and weather data, understand the consumption distribution of the household, the hourly consumption rate over time, as well as, the times of the day when electricity usage is high. We also hope to develop a classical ARIMA model and at least two machine-learning models to forecast the hourly electricity consumption for the next 7 days, compare their performance and in the end, present a report of findings and present results to Enefit management.

Data-mining success criteria

The evaluation metric for the project is the MAE (Mean Absolute Error) of the machine learning models developed. The evaluation will also assess the predictive improvement of developed models compared to the classical time series model.

Data Understanding

Gathering data

Data Requirements

- Hourly Electricity consumption data and Price
- Weather data such as the temperature for that hour, the dew point, snow depth, wind direction, rainfall, relative humidity, sea-level air pressure, average wind speed and weather condition

Verification of Data Availability

Data exists and is accessible for use

Definition of Data Selection Criteria

The data (historical and test set) will be available on Kaggle as comma-separated values (CSV) files. All variables in the data will be relevant for forecasting. Other variables such as the hour of the day and the month of the year will also be extracted from the time variable.

Describing Data

The historical data contains 8,592 recorded hourly consumption rates, the electricity price for that hour and the weather data recorded. The test data contains 168 hourly measured weather data with the respective hourly electricity price. This test will be used to predict the consumption rate for that hour.

Exploring Data

During exploration, some data quality issues were found: There were variables with missing values and inappropriate data types. The hourly consumption, precipitation and electricity price distribution were found to have outliers: there's a possible skewness in them. The time variable

was converted to date type and features such as the month, hour, weekday and week of the year were extracted to be used for modelling.

Verifying Data Quality

The precipitation, snow depth and consumption variables were found to contain missing values. For the precipitation variable, we intend to fill them with zeros to indicate that there was no rainfall at that hour. Snow depth had so many missing values in both historical and test data, however, the missing values were agreed to be filled with zeros to indicate no snowfall. We assumed that the missing values occurred because snowfall is not experienced in those months of the year. Similarly, for consumption rates, we settled on backfilling the missing values (that is, using the previous data as the consumption rate for the next hour). We assume that the current hourly consumption rate is the same as the consumption rate of the hour preceding it.

Project Planning

Data Preparation

After loading historical and test datasets, both will be prepared based on the findings from the data exploration section. The steps to be taken will include

- Filling of missing values with suggestions from data exploration
- Variable type conversion to appropriate formats
- Feature Extraction and Selection
- Checking for outliers and treating them
- Data Transformation
- Saving preprocessed data

Pandas data cleaning package and the *scikit-learn's* preprocessing functions will be used in this section to preprocess data. It is expected that this part will take 4 days to complete

Data Analysis and Visualisation

Here, we will answer the data mining goals by presenting visuals of them and using various statistical tests to analyse them. Here, data visualisation and analyses packages such as *scipy*, *matplotlib*, *seaborn*, and *pandas* libraries will be used

This part is expected to take 3 days to complete.

Modelling

In this stage, we will develop a classical time series model and 1 or 2 machine learning models. We intend to use the ARIMA model as the traditional time series model and random forest and gradient boosting algorithms. For the machine learning models selected, we intend to use the original variables only as predictors and also create lagged features of the dependent variable and use them as predictors, alongside the original variables. After model development, performance will be evaluated using the mean absolute error.

Before this stage, the historical data will be split into two: a training and a validation set, where the last 7 days will be used as a validation set. This is because we are dealing with a time series model and are preventing data leakage of future data during training. For the machine learning models, hyperparameter tuning will be performed on the training data using a 5-cross validation method. Also, to ensure that future data is not leaked during cross-validation, each split will be done in a time-series split fashion. After the optimal parameters of each model are found, their performance will be evaluated on the validation set.

After hyperparameter tuning is completed, the optimal parameters will be used to fit the whole training data (training + validation sets) and then used to predict electricity consumption on the test set. Feature selection tasks may be taken to determine if the performance can be improved. This task will depend on the time availability. It is expected that this part will take 10 days to complete.

In this task, we will use the ARIMA model from the *statsmodels* and *pmdarima* library while the random forest model from *scikit-learn* and gradient boosting algorithm from *lightgbm* packages will be used.

Report Findings and Presentation

After analyses, a report of findings will be written and prepared for presentation to stakeholders in a poster presentation format. It is expected that this will take the last 4 days to complete.