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, electricity prices have skyrocketed and consumers around the world are looking for options to reduce electricity costs. The invasion of Ukraine by Russia, in part, has put European electricity consumers, especially the Baltic nations like Estonia, in a peculiar situation with increased uncertainties about sustained availability and high energy costs.

Electricity usage depends on the hour of the day, the month of the year, electricity devices at home, and the weather conditions. High electricity usage also contributes to the environment's carbon footprint. The high electricity costs have led to consumers seeking better ways of reducing electricity costs by conserving energy. One way is by seeking a cheaper and alternative source of electricity generation.

As one of the electricity providers, how can we turn this into a win-win situation, how can we help our customers save electricity costs and as well preserve the environment? One way would be to optimise household energy usage by developing a predictive model capable of forecasting the electricity consumption per hour of the day. This will enable households to become aware of their hourly energy usage and in turn, be able to control the smart devices in the home to save costs. An accurate model is critical for more sustainable energy usage and will become a game changer in providing more disposable income 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, household electricity costs are saved and the carbon footprint in the environment is reduced.

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

- 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
- Tools/Packages to use include *Pandas* and *scikit-learn*.
- It is expected that this part will take 4 days to complete

Data Analysis and Visualisation

- Answer the data mining goals by presenting visuals of them and using various statistical tests to analyse them.
- *Scipy*, *matplotlib*, *seaborn*, and *pandas* libraries will be used.
- It is expected to take 3 days to complete.

Modelling

- Split historical data into train and validation sets, with validation consisting of data for the last 7 days,
- Use train data to develop models and test performance on the validation dataset.
- Develop an ARIMA using original features as predictors
- Random Forest and Gradient Boosting models using original and extracted features as predictors.
- Develop Random Forest and Gradient Boosting algorithms by creating lagged features of the dependent variable
- Perform hyperparameter tuning on the machine learning models using a 5-fold cross-validation using a time-series split fashion. This is to prevent data leakage of future data during cross-validation. For the ARIMA model, optimal p,d, and q values of the model will be determined.

- Optimal parameters will be used to fit train data and performance evaluated on validation data.
- Compare model performances based on MAE. MAPE and RMSE will be used also.
- Final fitting of optimal parameters on whole historical data and then forecasting on the test dataset.
- Feature selection method will be applied to the machine learning models to select features with high predictive power. This depends on the available time
- *Statsmodels* and *pmdarima* libraries will be used for ARIMA while *scikit-learn* and *lightgbm* libraries will be used for random forest and gradient boosting algorithms.
- It is expected that this part will take 10 days to complete.

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.