Introduction to Data Science
Project D13
Taavi Künnapuu
Ahti Ankro

Journey to zero - Predict electricity consumption

Business understanding

Background, business goals and success criteria

Journey to zero is a movement that's goal is to reduce the amount of CO_2 and other greenhouse gases that warm up the planet. In addition, electricity prices have skyrocketed and therefore consumers around the world are looking for options on how to reduce their electricity costs and their environmental footprint. For this reason, Enefit, one of the largest energy companies in the Baltic is looking for ways to help their customers as much as possible on their journey to zero.

One way of drastically reducing both the electricity cost and the environmental footprint, is by forecasting the consumption of household electricity and optimizing its energy usage. By forecasting, we enable smart energy devices and the consumers the ability of choosing when to consume electricity. At times, when the electricity is at its cheapest, also means that the environmental footprint is as small as possible. This is thanks to the low load on the electricity grid which means electricity plants that produce bigger amounts of CO_2 don't have to go online.

A model that successfully forecasts a household's electricity consumption, helps to smooth out the spikes in the load on the electricity grid. This is a small part of the journey to zero goal.

Assessing the situation and resources

Requirements, risks, and contingencies

The requirement for this project, is to have completed the model by the 8th of December and have the presentation ready by the 12th. This brings us to one of the two risks we can face during our project. The first is the deadline. Additionally, we can submit our model for evaluation at most 5 times per day. This means we need to work on our model on an ongoing basis.

The second risk is about finding additional data to support our project. Although the given dataset is of high quality, to make our model competitive, we might need to introduce new data. Given the deadline, we need to find balance between having worked enough with the starting data and finding new data.

Terminology

Smart energy devices – These are devices that can take charge of when to activate, based on certain factors. For our project, it's important that the devices take into account the prediction of when the electricity consumption and price are low. Examples of these devices are smart thermostats and smart charging stations.

Costs and benefits

The costs for our project are currently only the planned person-hours for the data mining, model building and presenting the findings. This comes out to be at least 60 hours of work, as we are a 2 person team, with each person working 30 hours.

The benefits of the project relate directly to success of the model. In the year 2020, Estonia's households used a combined of almost 2TWh of electricity. This means, if our models help consumers and smart energy devices make better decisions, then even small improvements in consumption habits will have a big impact on the consumers electricity bill and their environmental footprint.

Data-mining goals & success criteria

Our main datasets that we're going to use for the models are already provided. We have weather data, electricity price and a household's electricity consumption given in a time-series of one-hour intervals.

To make our models as good as possible so that we can achieve our business goal of accurate forecasts, we need to find additional parameters that can affect a household's electricity consumption. We can find these relations in studies that have been conducted before about electricity consumption. In addition to finding the attributes that relate to our goal, we need to make sure that we can find the data about these attributes that suit our currently available data.

To measure the effectiveness of any new data we find, we're going to incorporate it into our model and see if the mean absolute error – MAE improves. If the MAE improvement is significant, for example 5%, we can count the new data findings successful, and it will be added to the training set.

Data understanding

Our data is already provided to us by the Kaggle competition and we do not need to gather data ourselves.
The data points provided are by hour from a period of one year (1. September 2021 - 24. August 2022). In total there are 8592 entries in the data. There are 11 different variables in addition to the consumption, which is being predicted and time. Fields in the provided data:

- **Time** - dates by hour
- **Air Temperature (°C)**
- **The dew point in °C**
- **The relative humidity in percent (%)**
- **The one hour precipitation total in mm**
- **The snow depth in mm**
- **The wind direction in degrees (°)**
- **The average wind speed in km/h**
- **The peak wind gust in km/h**
- **The sea-level air pressure in hPa**
- **The weather condition code -** shows the condition, for example clear, fair, cloudy and so on
- **The electricity price in Estonia on that hour (€/kWh)**
- **The electricity consumption (kWh) -** being predicted

The data has multiple different variables, which are helpful in predicting electricity consumption with the exception of snow depth, which may not be necessary since we are predicting consumption for a week at the end of august. With there being over 8000 entries, it should be enough to get a good result.

In general the data looks to be of good quality, although there are some missing data and problems. The precipitation values only exist starting from May 2022, before that there are only a couple values, we could try to find more data, but the only available data seems to be where the precipitation is the average of a month, which does not suit us, because our data is hourly.

With snow depth there are values only in winter and only at 8:00 in the morning. The missing values outside of winter months can be replaced with 0 and when there is snow in the morning, then it has to be taken as consistent for the whole day.

There are also few data points, where the weather condition code is missing. Also the weather condition code is the only variable that shows a class, which means that we need to apply one-hot encoding to it.

The date (y-m-d h:m:s) has to also be processed. The year can be totally removed, because it shouldn’t affect the result. The months could be put through one-hot encoding.

In addition, there needs to be some normalization, because there are variables like air pressure, which has values only around 1000 hPa.

Planning our project

Methods tools time

For our project we plan to use the following Python libraries:

- Pandas matplotlib, scikit-learn, numpy

Additional tools we're using are:

- Jupyter Notebook for the model building
- Github for version control
- Canva to create the presentation.

Our main tasks and the time estimates for the project are the following:

Task description	Est. time
Clean up and normalize the provided data.	4 hours
Feature engineering the data for our model. - Find meaningful feature engineering techniques for our model	12 hours
Build models and evaluate their accuracy. Also submit the models for evaluation on Kaggle. - Document used machine learning algorithms and their parameters	10 hours
Find additional data that might help us with our model. - Clean up and feature engineer any new data	10 hours
Mark down our important findings for the presentation.	4 hours
Create graphs that illustrate our data findings. - Graphs that help us with model building i.e they help us with data relations - Graphs that we will include in our final presentation	10 hours
Create a presentation that highlights our goals and findings.	10 hours