KAGGLE-JOURNEY TO ZERO
Predict electricity consumption

TEAM:
Karolina Samasev
Madis Uljam Teuli

Link of the repository: https://github.com/karolinasamasev/DS_project

Task 2. Business understanding

Background

Our client, Enefit, one of the largest energy companies in Baltic countries, seeks to help their consumers as much as possible on their journey to zero. Both electricity cost and the environmental footprint could be drastically reduced by forecasting the consumption of the household and optimising its energy usage. Client believes that controlling smart energy devices in such a way that minimises the cost and environmental footprint of the consumption. As we are living in a rainy and foggy country, it could be noticeable that there is a strong connection between electricity usage and natural disasters. This poses a problem: the best time to act is before electricity costs increase to crazy prices, hovewer will the predictions have an impact on the local government to take the responsibility to immideately solve those issues.

Business goals

Our goals are to create an energy consumption prediction model for a single household and to test models until the best one is found.

Business success criteria

Our team’s business success criteria is to get our prediction model done.

Inventory of resources

Main resources for the project are two datasets described below

- **Dataset 1 (736.89 kB):** the training set includes the weather, electricity price and the electricity consumption for the period 2021-09-01 00:00 - 2022-08-24 23:00 for an individual household in Estonia

- **Dataset 2 (13.96 kB):** the test set includes the weather and the electricity price but not the consumption for the period 2022-08-25 00:00 - 2022-08-31 23:00 (the next seven days after the last timestep in the training data)

and, moreover, our team will use Python and all useful libraries and extensions, which are needed for successful data-mining and training.

Requirements, assumptions, and constraints

This competition is all about openness and our team should publish our code as a Kaggle notebook. Team may select up to 2 final submissions for judging.

The goodness of our model will be measured in how close our predictions were to the actual electricity consumption of that time, thus our main requirement is to minimise our errors in prediction.

We are assuming that the:

- consumption of electricity in the same weather conditions continues to be the same or similar for the time our model may be used (For example there will not be a massive renovation spree in Estonia the coming year.)
- the weather does not drastically change from what it was in the training period of 2021 – 2022.
- the main element in electricity consumption are elements related to weather, like heating and light consumption, or elements derived from weather, like people staying more indoors and taking part in activities that they would not in a good weather

We are constrained to only weather data and only 8592 rows of data with limited knowledge of prior approaches.

Risks and contingencies

As our team has a public dataset, no risk connected to leaking data. However, accurate household-level predictions are a critical prerequisite for a more sustainable energy usage, and if they are not, the risk connected to unfeasible energy usage in the future is extremely high. It may have a bad influence on the experience of those who will make use of our predictions.

Moreover, main risks are, that our assumptions are wrong:

- that the consumption of electricity will stay the same as in the aforementioned period; 2021 still had some lockdowns due to Covid, thus there might be a larger consumption of electricity in most of the data we have then there will be this year
- we are living in the age of climate change, thus the data we have collected might not reflect the weather we might see this year
- the main element of electricity consumption might have changed, for example everyone is buying the coolest newest VR headset that uses a lot of power OR that last year a lot

of households might have been mining cryptocurrencies and after this summer it is not as popular

We might want to weigh up newer data, as it might help to predict these risks, other than that we do not have good contingencies.

Terminology

MAE - mean absolute error is a measure of errors between paired observations expressing the same phenomenon

Accuracy - is the number of correctly predicted data points out of all the data points

Numeric data - numeric data types are numbers stored in database columns

Learning model - a file that has been trained to recognize certain types of patterns

Pandas – dataframe and manipulation package in Python

SKlearn – machine learning package in Python

Seaborn – data visualisation package in Python

Costs and benefits

This point is not relevant in the context of our project. However, we can make some predictions if our project would be taken for use:

For our team the benefit is getting a better grip on how to use data mining in business applications.

For Enerfit there is presumably a human labour cost for gathering, cleaning the data and creating this event.

For Enerfit they might find a model or an idea for a model that actually helps them in the future or new workforce to employ.

Data-mining goals

As our model output is numeric, our team will use regression. Here are data-mining steps

- Read in all the data
- Cleaning the data from unnecessary columns, rows and values; rename variables and make dataset look more readable for an eye
- Make an analysis using evaluation metric called Mean Absolute Error

- Modelling

Therefore, data-mining goals are to process the dataset from the provided one, which is clear from extracted features, and to visualise the dataset and results of modelling.

Data-mining success criteria

Main success criteria here is to get the accuracy of our training model above 80%.

Task 3. Data understanding

Gathering and Exploring data

Data requirements

Data is required to have accurate actionable data about weather and energy consumption.

Verify data availability

Data is available (8592 observations of 13 data points.)
with additional test data, that has 168 observations of 12 data points (The actual consumption is missing.)

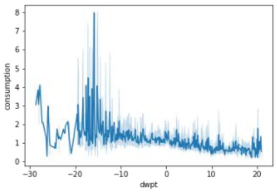
Define selection criteria

As we are given mostly weather data this is what we are working with. We are assuming, that this is the most important data for predicting energy consumption.

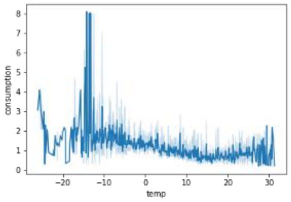
Exploring

Describing data

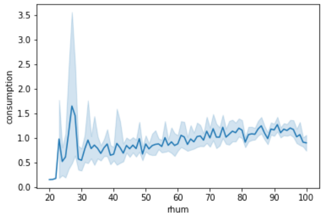
We are given 13 datapoints:
Time – accounts for time series and allows us to connect collected data to a specific season (hopefully allowing us to get even better predictions.



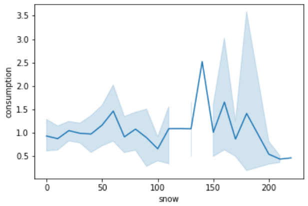
Dew point in °C – A combination of temperature, pressure and relative humidity that also might play an important role in determining the consumption. In this case we must be weary of multicollinearity.



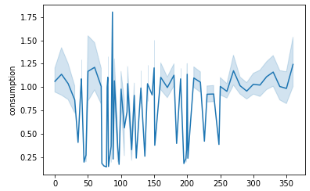
Temp – A normal numeric temperature value in °C that probably will be the main contributor to the energy consumption (Assuming, that other values are modifications of the temperature value.) connected to dew point. There seems to be a massive consumption increase whenever temperature falls below 0°C, above that the temperature and consumption do not seem to have a connection. Below -20°C there are very few observations



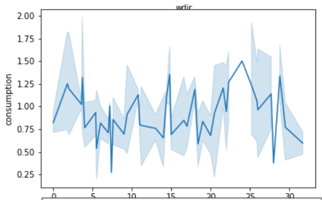
Relative humidity in percent – Can be handled as a numeric value. Connected to dew point. See to generally slightly increase energy consumption.
The one hour precipitation total in mm – how much snow, rain or other weather elements came down in an hour. Connected to dew point and seems to have little of significant data.



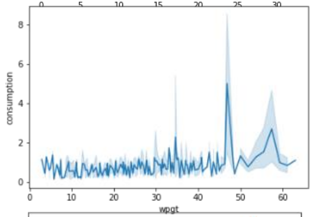
Snow depth – depending on how the data was collected it might have a positive effect (snow covering the building from the cold wind) or no effect at all. Most likely not a really important data point.)



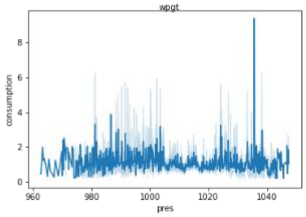
The wind direction in degrees (°) – a numeric value that might be more useful turned into a categorical one, as it is nonlinear (a circle). The graph right now shows that eastern wind is more cold and western more warm.



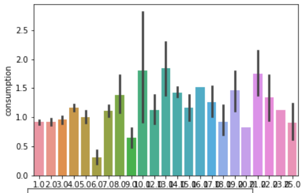
Average wind speed in km/h – most likely to be another important data point that might give great results if combined with the wind direction and the temperature. On its own it hardly shows anything.



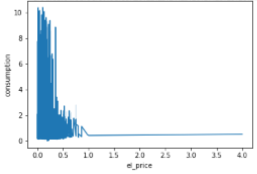
The peak wind gust in km/h – seems like on really high wind speeds the consumption increases a lot.



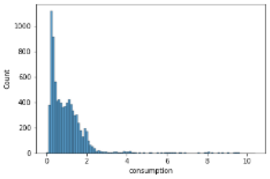
The sea-level air pressure in hPa – might also be important as it helps to detect if we are in a high- or low-pressure front. Might be turned into a categorical value. Might be multicollinear with dew point.



Coco – the weather conditional code, giving 27 unique weather states. Generally seems to give the range of possible energy consumption quite well.



Electricity price in Estonia on that hour – we might be able to see what other companies expect or predict from this, however translating this expectation to our model will not be easy.



Consumption – the actual value we are predicting. Most values seem towards the low end.

Grouping data:

Market:

Y – Consumption
Electricity price

Main:

Temperature
Air pressure
Peak wind gust
Wind direction
Relative humidity
Coco

Extra:

Precipitation
Avg wind speed
The one hour precipitation
Snow depth
Dew point

	temp	dwpt	rhum	prcp	snow	wdir	wspd	wpgt	pres	coco	el_price	consumption
count	8592	8592	8592	2159	119	8592	8592	8592	8592	8396	8592	8590
mean	6.7	2.5	77.0	0.1	78.3	201.6	9.2	20.9	1013.2	4.9	0.2	1.0
std	9.3	8.2	17.5	0.4	63.1	87.8	4.8	10.0	12.6	5.0	0.1	1.1
min	-26.1	-28.7	20	0	0	0	0	2.9	962.6	1	7E-05	0
25%	0.4	-2.9	66	0	20	150	7.2	13	1007	2	0.093	0.363
50%	6.2	1.9	83	0	60	210	7.2	18.5	1015	3	0.136	0.811
75%	13.23	9	91	0	130	270	10.8	27.8	1021	5	0.2	1.366
max	31.4	20.9	100	7.9	220	360	31.7	63	1048	25	4	10.38

Verifying data quality

There are some missing values in Precipitation, snow and coco – We believe that this isn't missing data per say but just notion that there was no snow, precipitation or special weather to

note. Thus we believe, that there are no missing data values or strong problems with data quality. There however are bound to be outlies that we might want to eliminate.
We are missing data on temperatures below -20°C, as it happens rarely.

Task 4. Planning your project

Our team's plan is:

- To clean all the data from unnecessary fields, e.g columns
- To visualise data to make it more readable
- Data analysis will be performed by using Mean Absolute Error
- Then we should analyse all the subsequent results
- One of the last steps would be training models
- Finally, the codebase of all the project will be documented

Tools: We will use libraries such as pandas and numpy, presumably Seaborn and SKlearn