Project C7:
KAGGLE-JOURNEY-TO-ZERO

Members: Uku Parts
https://github.com/UkuParts/journey-to-zero

1. **Business understanding**

This project is being done in service to Enefit, a subsidiary of Eesti Energia, whose primary goal is to proliferate the use of “green” energy in order to reduce the environmental footprint of electricity use.

Enefit claims that, “Both electricity cost and the environmental footprint could be drastically reduced by forecasting the consumption of the household and optimizing its energy usage (controlling smart energy devices in such a way that minimizes the cost and environmental footprint of the consumption).”

To that end, they have created a Kaggle competition where anyone can submit a model for predicting the electricity usage of a given household, based on data of that household's previous usage. Their hope is that they get an “accurate” model from this competition. What exactly constitutes “accurate” is not defined.

Since this is just a Kaggle competition and I am just 1 university student, my resources are rather limited. I am the only person working on this and I only have my own personal hardware and software to use. Because of this, the cost to Enefit is basically 0 while the potential benefit is considerable. In terms of data, Enefit has provided a dataset for the competition that consists of 1 year of electricity consumption and other relevant info for a single household.

In exchange for the low cost to Enefit, I also have no obligations to them. The deadline for the competition is 09.12.2022 but even if I don’t submit anything by then, there will be no consequences for me. There’s also no strict criteria that define what makes a submission “acceptable”.

Because of the simple nature of my setup, there aren’t many things that could go wrong in a way that stops me from finishing my work in time. If my computer breaks or something, I can just get another one. If the Estonian power grid fails, I can use a generator. Depending on the cause of that failure, predicting the electricity consumption of a given Estonian household may also become very easy for a period of time.

My goal is to create the model that the competition requests with the lowest possible Mean Absolute Error (MAE), the metric by which the contestants are scored. I have no particular MAE in mind, I find that kind of a goal hard to set since I have no experience dealing with this kind of a model. I’ll simply do my best.

2. **Data understanding**

Since this is a Kaggle competition, the data is provided by the hosts. It's a dataset containing data about a single household. To me, this seems rather strange, since this is supposedly for a model that should generalize to every household in Estonia. However, since no public dataset exists for this purpose and gathering a large amount of data myself would probably be very difficult / impossible (because of privacy laws etc.), I will stick to using only the provided dataset.

From the competition data page: "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." Specifically, the training set consists of 8592 elements that are all electricity consumption readings at 1 hour intervals with a bunch of other added data. This added data is: air temperature, dew point, relative humidity, precipitation, snow depth, wind direction, wind speed, peak wind gust, air pressure, weather condition code and electricity price.

Reading the data and running describe() on it produced the result you can see in the appendix. The most relevant metric for our purposes is consumption. It had a mean of about 1.05 (kWh), a standard deviation of about 1.1, a maximum of 10.4 and a minimum of 0. Worth noting is that it also had 2 missing values.

I also looked at how the various factors are correlated with consumption. The results of that can also be seen in the appendix. It's mostly what you would expect, temperature is inversely correlated with consumption while wind direction, for example, has little effect.

It seems to me that this data is fit for our purpose, for the most part. There is the detail I mentioned earlier about how this is data from a single household and we're ostensibly creating a model that can generalize but I don't see anything I could do about that. Otherwise, the data seems to be of reasonably high quality.

There are only 2 missing values that will have to be dealt with in the consumption column. Since I have no access to the consumption data, I will have to either drop these 2 rows or fill in the value with something that seems reasonable (the value for the previous time-step, perhaps).

The snow and precipitation columns seem to be missing a lot of data. I need to either get this data or just drop the columns entirely. The condition code column is also missing some values that will have to be dealt with. Unlike electricity consumption, weather data is public so I can probably fill in any blanks I need to. This may still prove to be difficult though, since I don't know the exact location of this household.

3. **Planning**

The tasks that need to be done are:

- Research - looking into how a model for this purpose should be made. Both broadly (how to deal with time-series data) and specifically (there's actually a masters thesis written on this exact topic).
- Data cleaning - dealing with the missing values.
- Feature engineering - adding any features that seem like they might be useful.
- Writing code for the model - actually writing the code for the prediction model.
- Preparing a poster - preparing various figures, making sure everything looks at least alright. Not a part of the competition but since it's 50% of my grade I feel justified in putting it here.

I am a team of 1 so I will be doing all of this myself and plan to spend around 60 hours in total. Estimating how many hours will be spent on each individual task is difficult, since I have little (no) experience with projects like this.

I think most of my time will be spent on research, data cleaning and feature engineering. Writing code for the model shouldn't take more than a couple of hours, since I've been doing that for half the course already. Data cleaning may prove to be a problem depending on how easily I can get my hands on some weather data, it may take several hours. Research and trying out different models and feature engineering based on said research is where I expect to spend the bulk of my time.

I'm probably gonna spend a good few hours on the poster as well. Why is it 50% of the grade?

4. **Appendix**

Result of the pandas describe() function on the provided dataset.

	temp	dwpt	rhum	prcp	snow	wdir	wspd	wpgt	pres	coco	el_price	consumpt
count	8592.000000	8592.000000	8592.000000	2159.000000	119.000000	8592.000000	8592.000000	8592.000000	8592.000000	8396.000000	8592.000000	8590.000
mean	6.744204	2.486767	77.013617	0.056647	78.319328	201.564246	9.156355	20.869681	1013.229423	4.902930	0.160844	1.046
std	9.257806	8.184391	17.520566	0.384586	63.129130	87.792064	4.826976	9.956558	12.592944	4.958744	0.120034	1.095
min	-26.100000	-28.700000	20.000000	0.000000	0.000000	0.000000	0.000000	2.900000	962.600000	1.000000	0.000070	0.000
25%	0.400000	-2.900000	66.000000	0.000000	20.000000	150.000000	7.200000	13.000000	1006.500000	2.000000	0.092820	0.363
50%	6.200000	1.900000	83.000000	0.000000	60.000000	210.000000	7.200000	18.500000	1014.700000	3.000000	0.136440	0.811
75%	13.225000	9.000000	91.000000	0.000000	130.000000	270.000000	10.800000	27.800000	1020.700000	5.000000	0.199845	1.366
max	31.400000	20.900000	100.000000	7.900000	220.000000	360.000000	31.700000	63.000000	1047.500000	25.000000	4.000000	10.381

Result of the pandas corr() function run between consumption and all other columns.

```
temp: -0.2678644480354112
dwpt: -0.2504839596413663
rhum: 0.09440778205097211
prcp: -0.010667922845358507
snow: 0.00416288397956103
wdir: -0.005008931638109928
wspd: 0.048839999143518
wpgt: 0.07220583134057182
pres: -0.06894004928348695
coco: 0.1335385829080077
el_price: -0.12474038219665333
```