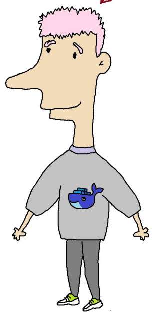
HOW MUCH DOES 1°C COST
YOU AND NATURE

ANNA PARFENOVA | SPICED ACADEMY
DATA SCIENCE BOOTCAMP | FINAL PROJECT

HOW MUCH DOES -1°C COST HIM ~~YOU~~ AND NATURE



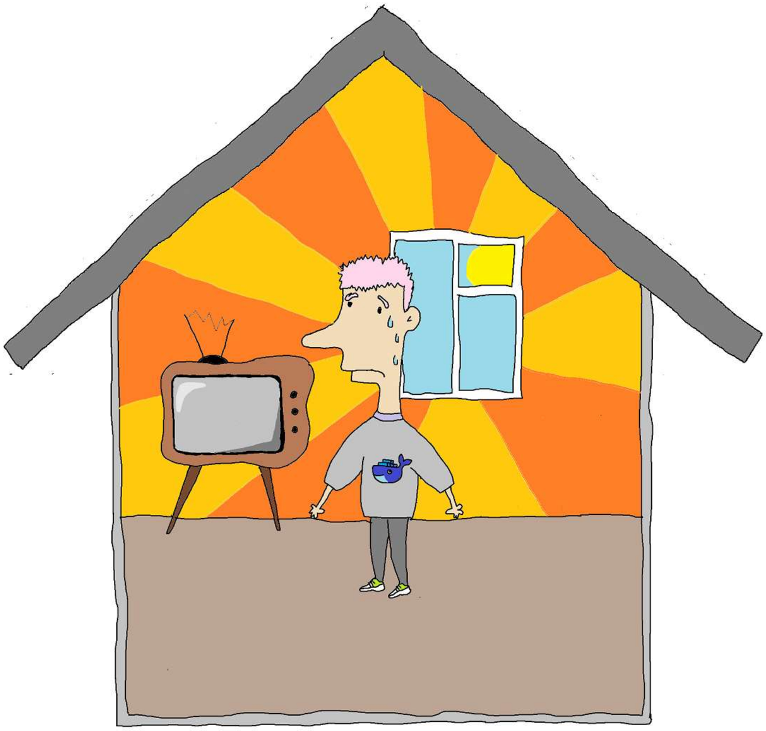
This project aims to calculate the effectiveness, cost and carbon footprint of using a portable air conditioner unit.
Its official title sounds pretty general, when actually it was calculating the effectiveness, cost and carbon footprint of using a very particular portable air conditioner unit in a very particular apartment that belonged to my friend Sasha.

Sasha is working as a Data Engineer in a cool Berlin startup and, more importantly, he has a very useful (for this project) hobby: Smart Home. It means that Sasha's house is literally stuffed with various sensors, constantly measuring temperature, humidity, CO_2-level and keeping track of important events like doors or windows openings, switching on/off lights and so on.

A little more about Sasha's background.
Sasha is living in Berlin from 2014. Originally, he is from Siberia.

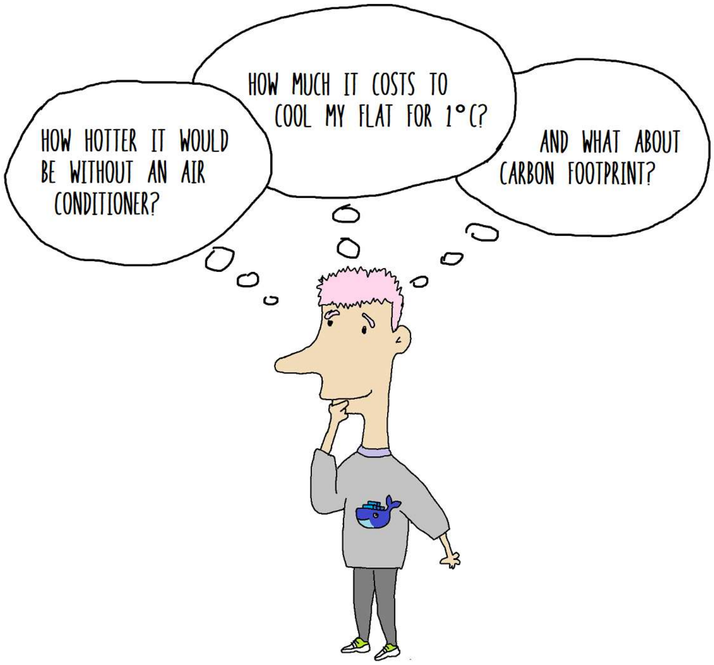
That is, probably, why Sasha really loves the winter in Berlin.
(And if you don't, you should visit Novosibirsk in February first, then reconsider!)

Sasha also enjoys Berlin in the summer…

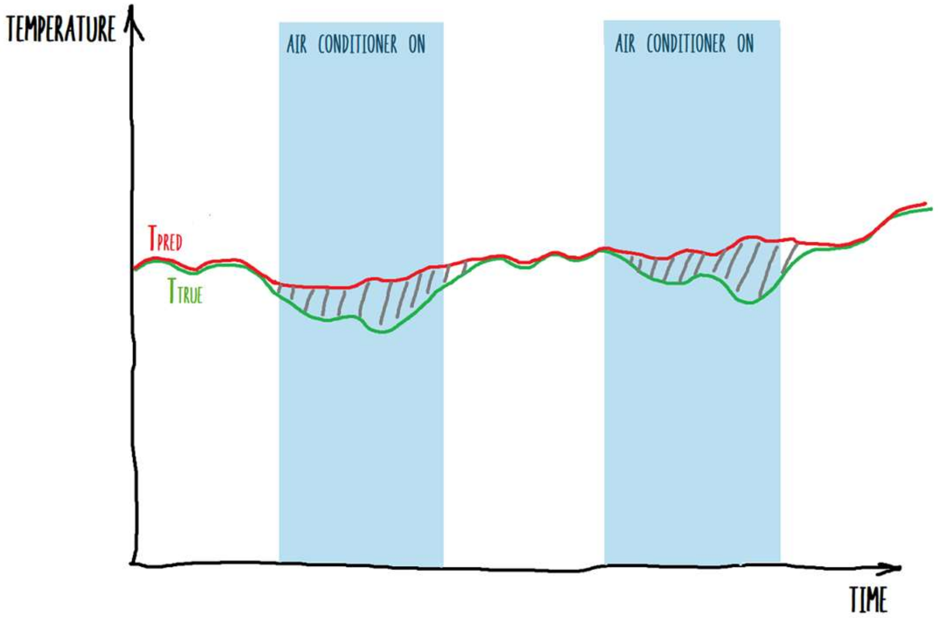


…but there is one problem.
While it might be pretty enjoyable outside, in his small “Neubau”-flat it gets really hot.
In 2020 it became a real problem, because Sasha, like many of us, works from home.

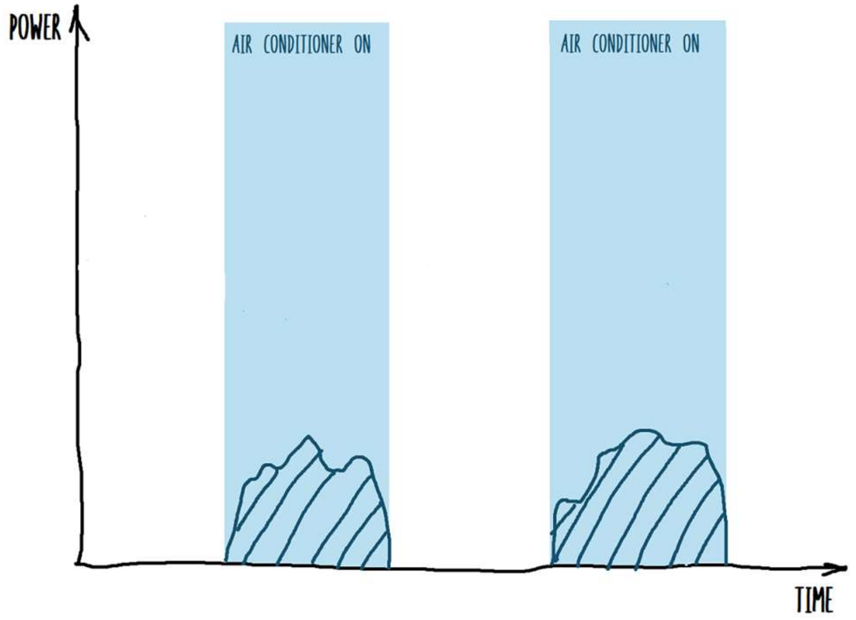
Finally, Sasha put his economical and ecological sustainability concerns to one side and decided to equip his flat with a portable air conditioner unit.
(And, let's be honest, we'd do the same. Sorry, Greta 😟)


HOW HOTTER IT WOULD BE WITHOUT AN AIR CONDITIONER?
HOW MUCH IT COSTS TO COOL MY FLAT FOR 1°C?
AND WHAT ABOUT CARBON FOOTPRINT?

However, Sasha could not completely clear his conscience. He was aware about discussions and concerns regarding effectiveness of this kind of portable cooling devices and despite the abundance of raw data, some questions remained unanswered.
Obviously, a detailed data analysis and even some machine learning techniques were needed to answer them. That is how an idea of this project was born.

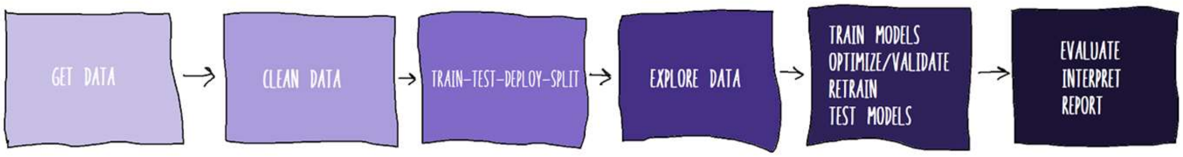


After careful consideration was given to the question, a project and solution was found: The task was to predict the temperature in his flat that would remain in the room without air conditioning in the span of time when actually the room was artificially cooled down. Then by integrating the difference between predicted and real time/temperature lines the total cooling value (units: °C×h) could be calculated.



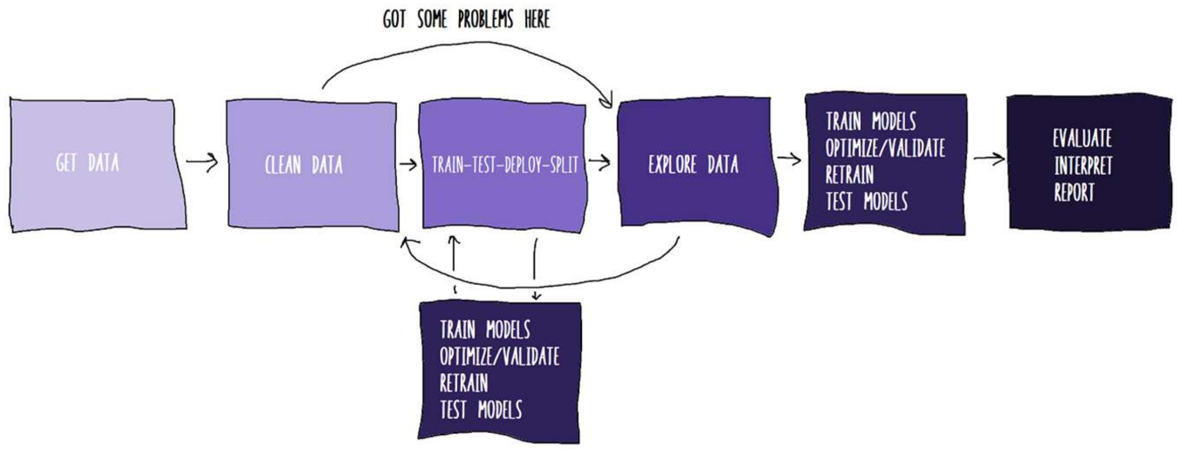
By knowing the running voltage of the air conditioner, the consumed energy (kW×h) can be easily calculated (also by integration).
Dividing the cooling by the energy will give us effectiveness of the air conditioner, and with some additional data about energy price and carbon footprint of energy production the cost and carbon footprint of cooling the flat for 1°C for 1 hour could be calculated.

INITIAL PROJECT PLAN

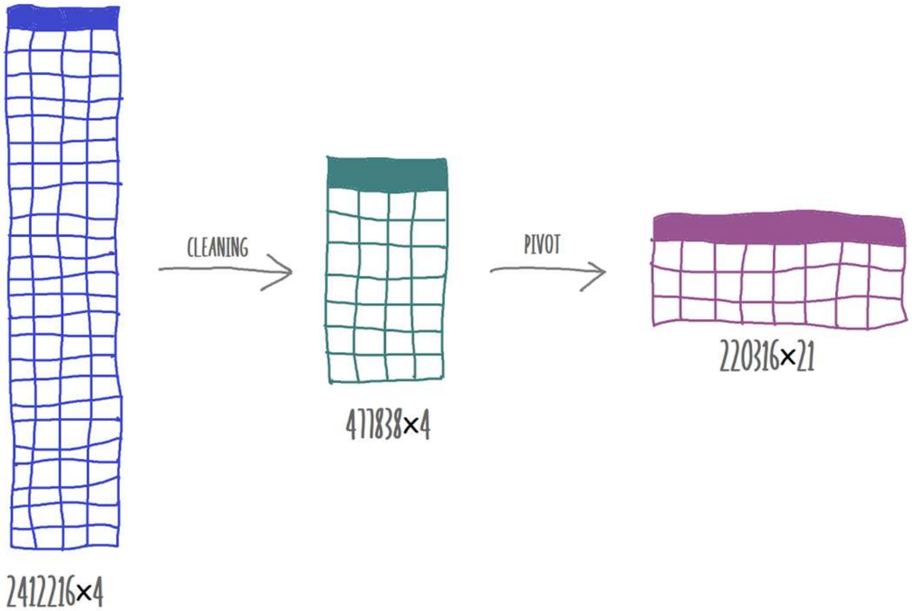


In this project a so-called business goal, or questions to answer was clear from the beginning, therefore, the initial project plan wasn’t any different from the classical machine learning project workflow, namely:
Gathering data (done by Sasha)/Data pre-processing/Researching the model that will be best for the type of data/Training and testing the model/Evaluation

FINAL PROJECT PLAN

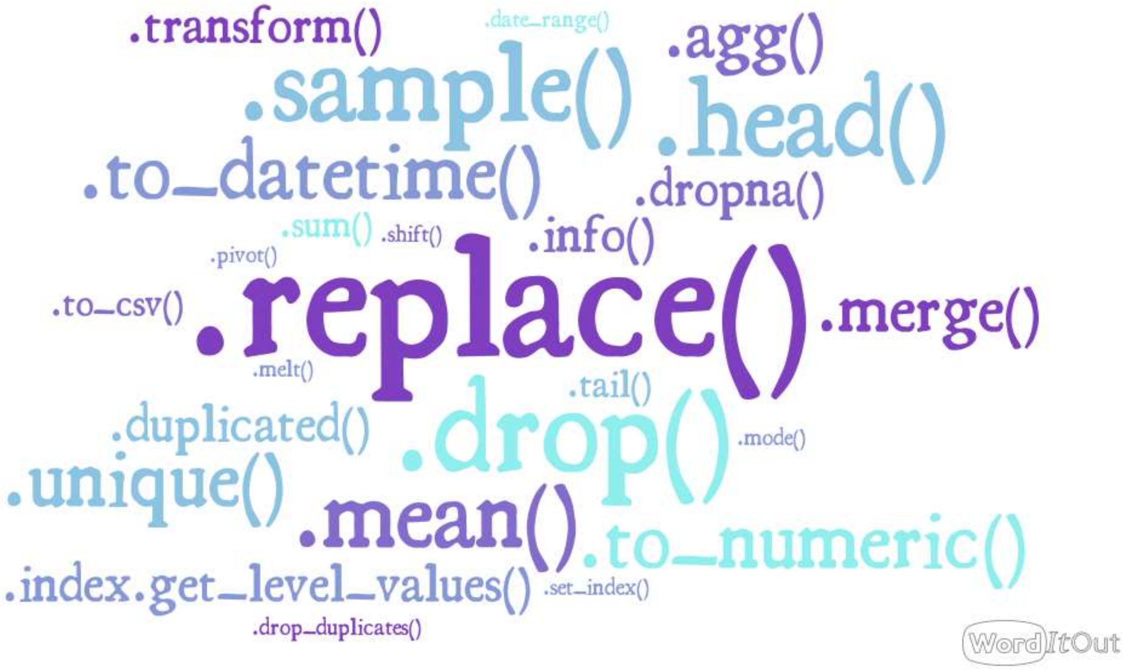


Unexpectedly (sarcasm 🙃) the plan had to be changed on the very early stage.
It happened to be a challenge to split the data for modelling (train+test) and evaluation parts. Which was, indeed, unexpected. More about that later.



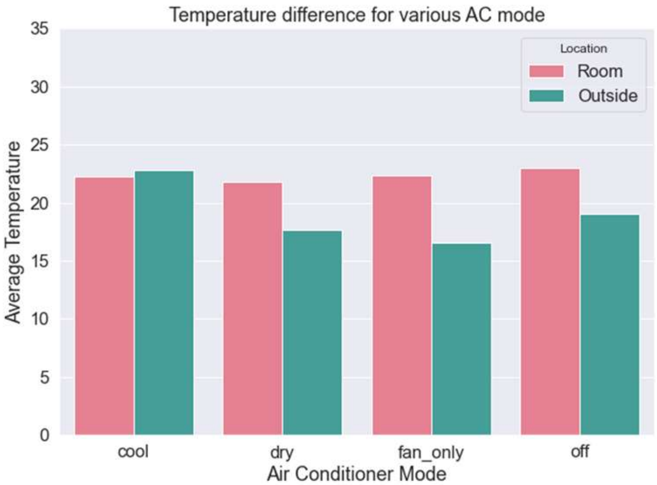
The row data had the form of one table with over two million rows, each of them contained Sensor type and name, various timestamps, values (numerical for measurements like temperature and categorical for event-sensors) and some other attributes.
The main challenge of data wrangling was to pivot this table into time-series or, more specifically, sequence of successive equally spaced points in time with values for each sensor.

WORD CLOUD OF MOST IMPORTANT PANDAS FUNCTIONS USED FOR DATA WRANGLING

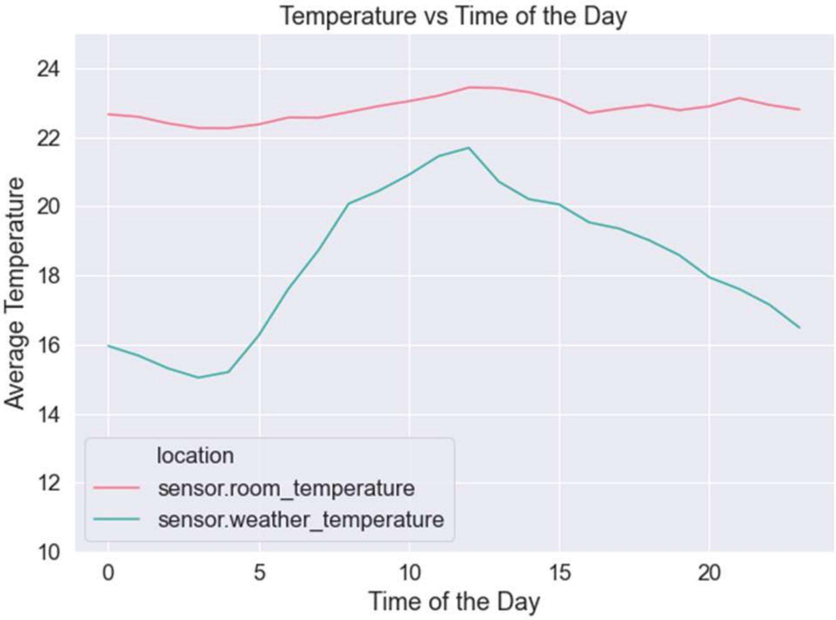


The main hitch during the data wrangling:
Some sensors were recording different values within one second (not always as a bug, often it was representing a real situation, when, for example, the window was opened and closed very quickly). For numerical data it was easily solved with combination of *pandas groupby()* and *mean()* functions, for categorical data the *mode()* function from *scipy* library was used.

In the original data observations from more than 70 sensors available, some of them were clearly irrelevant for the project (like humidity of the plants soil) and were excluded immediately.
Some data needed exploratory analysis to consider it's interdependency.
More information can be found in EDA Notebook.

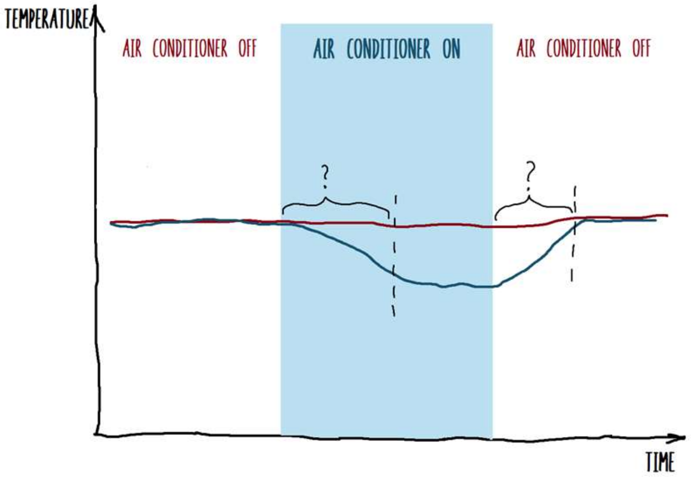


On the bar plot above average room temperatures and outside temperatures for the different mode of air conditioner unit can be seen. It's pretty clear, that "dry" and "fab only" mode are much closer to the "off" mode, than cooling. Based on this insight the values for the air conditioner mode were transformed into binary categorical values "on" (when "cooling") and "off" (when "off", "dry" or "fan only")

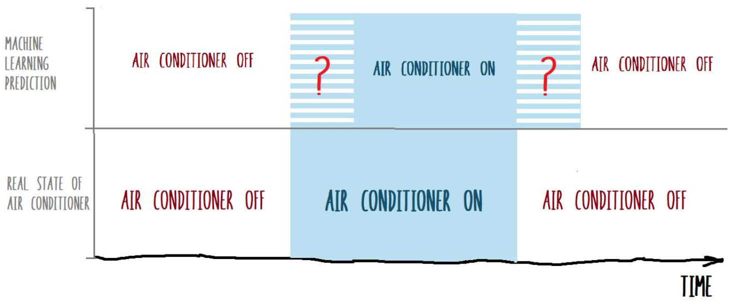


On the plot above dependencies of average outside and room temperatures on time of day can be seen. It is clear that the difference between these lines is strongly dependent on time, therefore important of our model. Neither day of month nor day of week showed any noticeable effect are were not included in the model.
For the logistic regression and linear regression models polynomial time features were used.

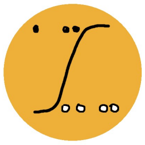
ANOTHER CHALLENGE EMERGED WHEN IT CAME TO SPLITTING THE DATA



The initial idea was to train the model on the temperature data for the time periods, when the air conditioner was "off" and then to re-construct the "uncooled" temperature, when it was on. But here comes the hysteresis problem: it takes time to cool down the room and, more importantly, the room stays cooler for some time after switching off the cooling and that time can not be used for training.

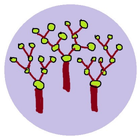


To estimate hysteresis lag regression machine learning was used, as the time after changing the state the air conditioner, when our model can not properly predict the right state or, in other worlds, to recognize that the state was changed.



LOGISTIC REGRESSION
ACCURACY = 91%

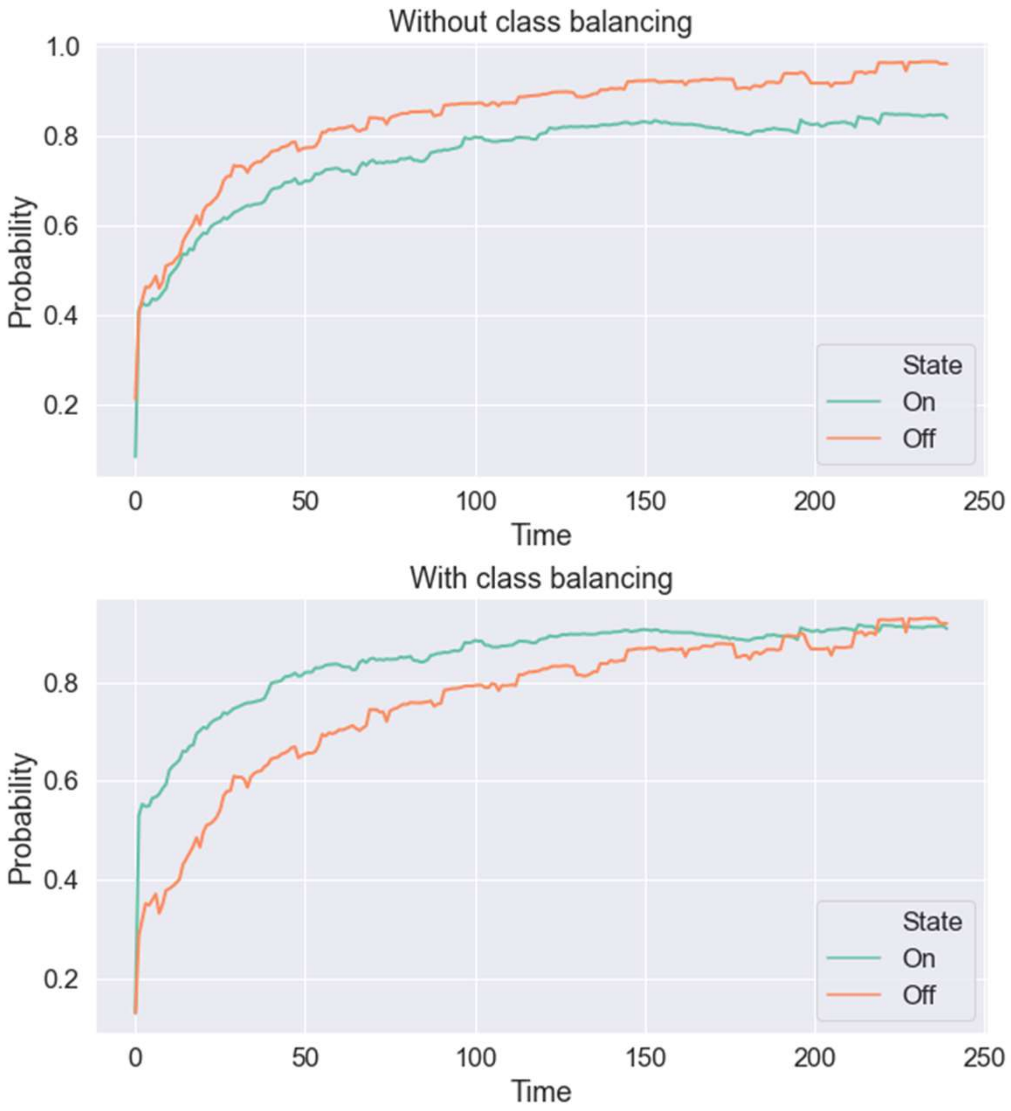
DECISION TREE
ACCURACY = 96%



RANDOM FOREST
ACCURACY = 99%

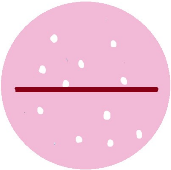
For this task three basic Classifiers from Scikit-Learn python library were tried. The best performing model was a random forest (max_depth=15, n_estimators=50).
However, for our purposes we used the results of Logistic Regression Classifier, because its output of the logistic function can be interpreted as a probability and will allow us to determine the hysteresis lag.

Initial dataset had noticeable class imbalance: air conditioner mode off has about 2.5 times more observations than “off” mode. Although this imbalance is quite moderate in comparison with what could be seen in classic credit card fraud detection or churn prediction problems, the effect of weight adjustment is clearly tangible.
Below you can see the logistic regression probability of the „correct“ state of air conditioner by the time after changing the state (for example time=10 means that the air conditioner was switched on/off 10 min ago).
Using the weight adjustment* does not affect the accuracy of the model (in both cases it was 91%), but in case of balanced classes probability lines approach the same value of over time, when for unbalanced classes the probability of underrepresented class never achieves 90%.

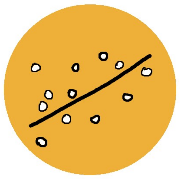


As to the hysteresis lag, every 100 minutes right after switching off the air conditioner were included into “deploy” dataset, when logistic regression model can predict the correct state with average probability 80% or more.

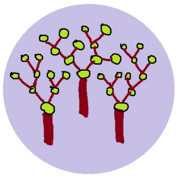
* Built-in *sklearn.linear_model.LogisticRegression* parameter *class_weight* was used.
The “balanced” mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as *n_samples / (n_classes * np.bincount(y)).*



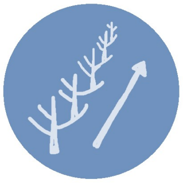
BASELINE MODEL: MEAN
RMSE = 1.65(°C)



LINEAR REGRESSION
RMSE = 1.23(°C)



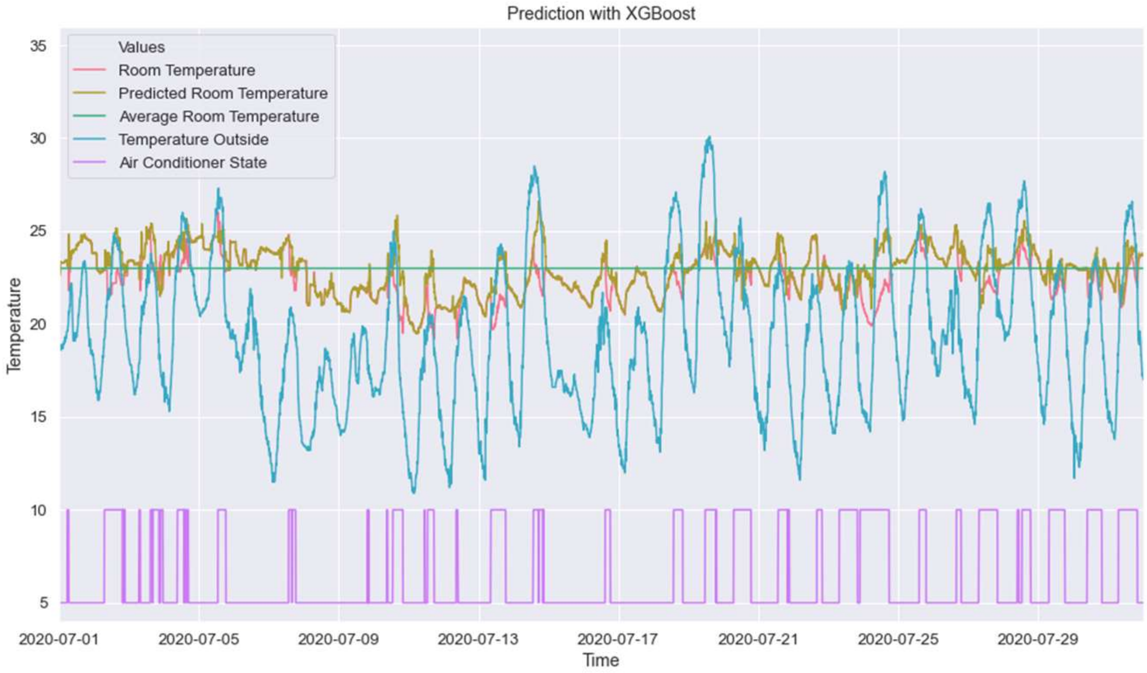
RANDOM FOREST
RMSE = 0.52(°C)



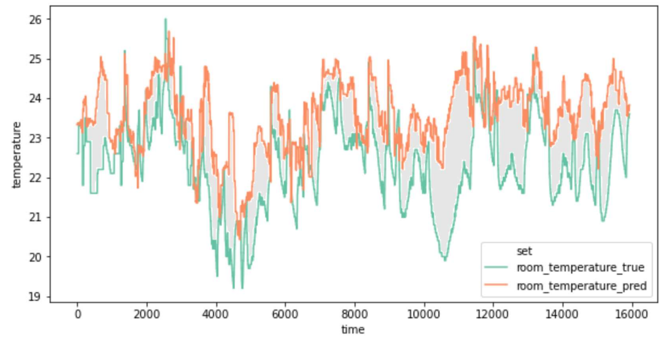
GRADIENT BOOSTING
RMSE = 0.43(°C)

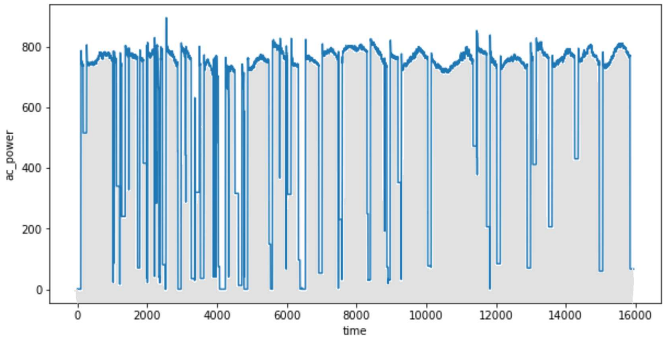
XGBOOST
RMSE = 0.09(°C)

For the temperature prediction four standard Regressors from Scikit-Learn python library were used. Random Forest and Gradient Boosting shown significant improvement in comparison to the Baseline Model (mean room temperature) and Linear Regression. But the best results were obtained with XGBoost model — with not much hyperparameters tuning test RMSE was 0,09°C, which is below thermometer sensitivity.



On the plot above observed room temperature, predicted room temperature as well as average room temperature and outside temperature for July 2020 sub-dataset can be seen. The binary valued line on the bottom indicates the state of air conditioner.

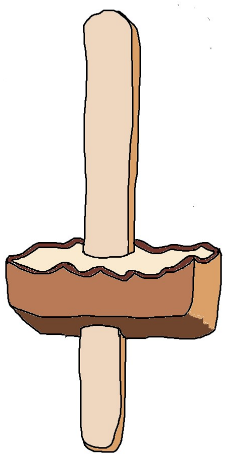


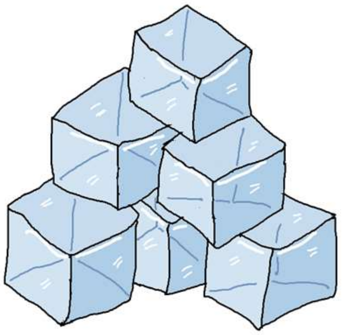


RESULTS:

- AVERAGE TEMPERATURE DIFFERENCE: 1.5°C
- 0.5 KWH NEEDED TO COOL THE ROOM FOR 1°C

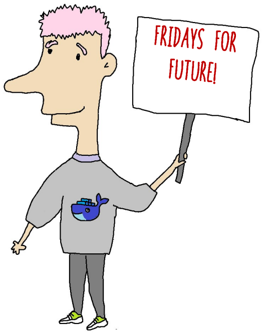
The mean difference between predicted and observed room temperatures for the time periods with air conditioning working in the cooling mode shows that the average artificial decrease in room temperature is about 1.5°C.
For cooling the room over night by 2°C (20 °C×h) ca. 10 kW×h would be used, which make up 10% of average single-person-household energy consumption in Germany.





13 EUROCENT OR 1/10 OF ICE CREAM OR 6 ICE CUBES FROM SUPERMARKET

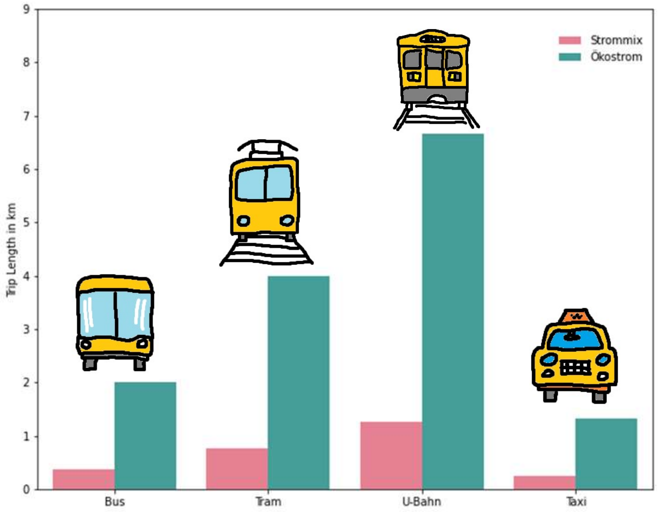
From the monetary point of view, however, the use of portable air conditioner doesn't look so chilling: cooling the room for 1°C×h cost ca. 13 Eurocent, which means that cooling the room over night by 2°C (20 °C×h) would be equivalent of 2 ice-creams or 1.5 package of ice.



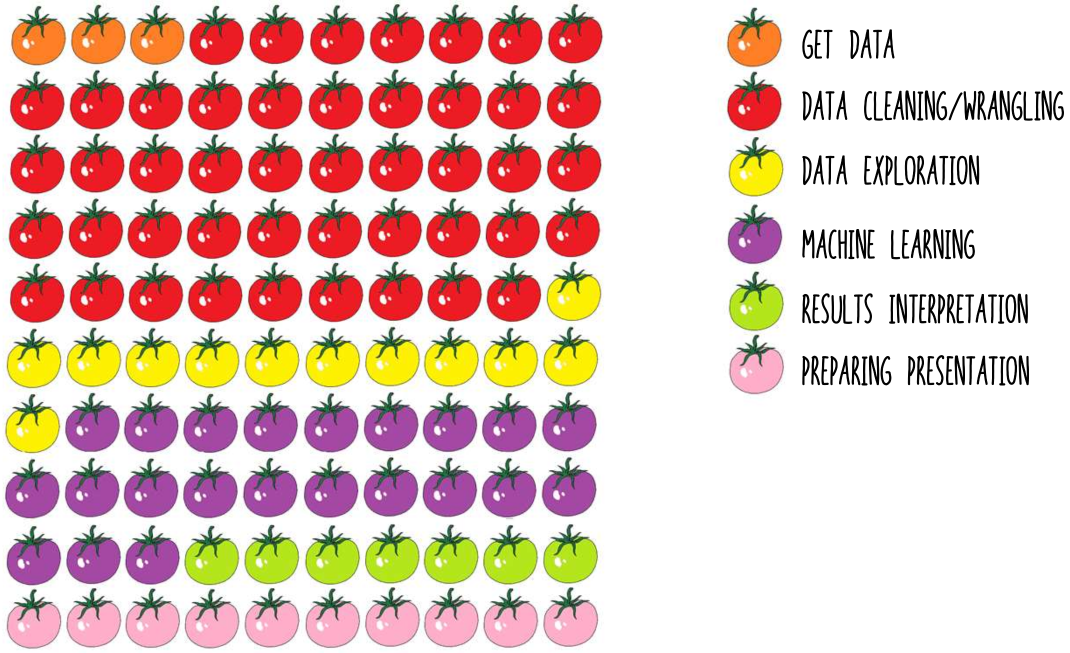
Calculating the carbon footprint of room cooling came to be the most tricky part of out results interpretation.
Sasha uses green energy provider, which declare their carbon footprint as 0.
They do, in fact, use only renewable energy sources and compensate the still caused by hydro-, wind- and solar- power plants greenhouse gas emission with planting trees in Canada.

CO$_2$ OF ECO-ELECTRICITY:
77 G/KWH | 2015
SOURCE: BUND DER ENERGIEVERBRAUCHER E.V.

CO$_2$ OF MIXED ELECTRICITY:
401 G/KWH | 2019
SOURCE: STATISTA GMBH

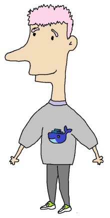


A deeper research brought us the “brutto” emission of electricity, provided by the company mentioned above. Unfortunately, the data is a bit outdated and considering they constant move towards “green” energy sources we can consider this number an upper-bound estimate. On the plot we can see travelling distance equivalent to cooling the room for 1°C×h for different type of transport and different energy providers (“eco” and “standard”).



It is often mentioned in data science or data analytics learning and training materials, that about 80% of time of data-specialist is normally spend on data cleaning.
I decided to conduct a little sub-project and tracked time spent on the different parts of this project. Although data cleaning and preparation stage took just a bit above 50% (and I should thank again my data engineer friend for that) it was still the longest part of the project.

AUTHOR IS GRATEFUL TO:



AND ALSO, PATRICK, PEBBLE AND ALL MY FRIENDS AND FAMILY, WHO SUPPORTED ME AND WILL CONTINUE TO DO SO