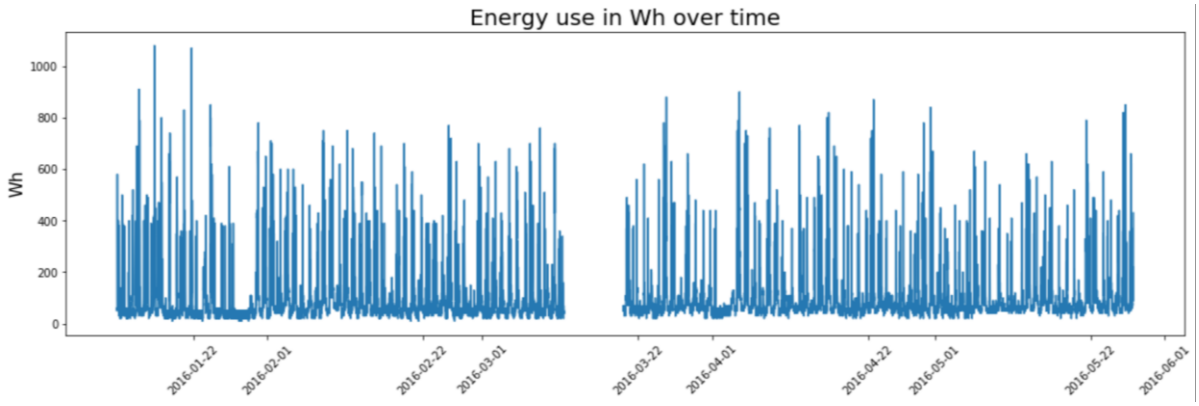
REPORT REGARDING ELECTRICITY USAGE OF A HOUSEHOLD

EXPLORATION

The dataset consists of 19735 observations and 29 variables including the target variable "Appliances". On average the energy use is 97 Wh with a minimum of 10 Wh and a maximum of 1080Wh. The temperatures from the different rooms in the house tend to be relatively similar with an average maximum of the temperature in the laundry room area at 22.32 degres C and an average minimum of the temperature in the parents room at 19.48 degres C.

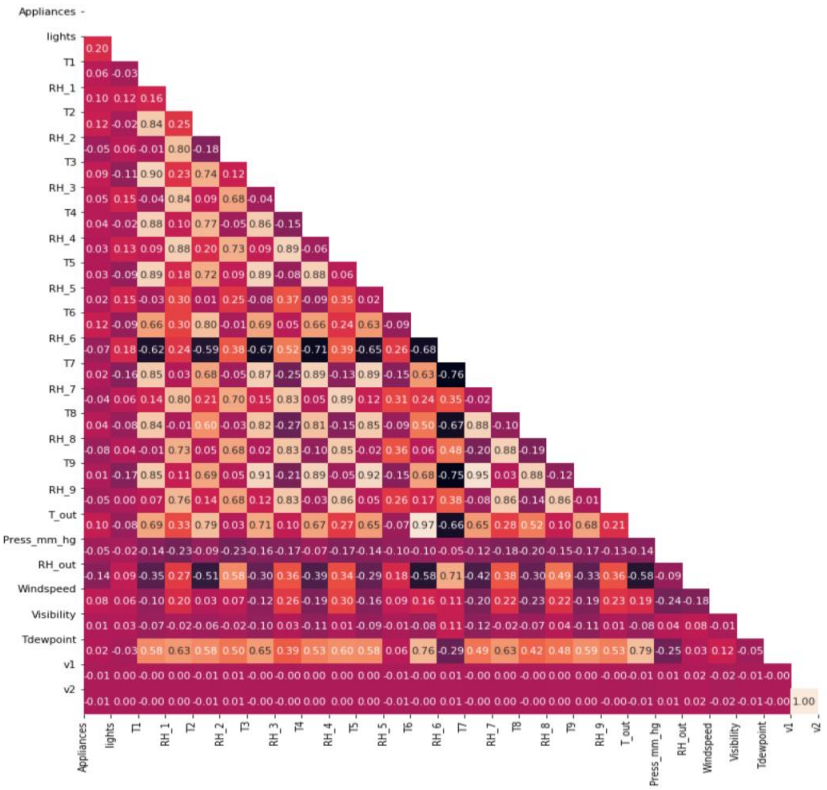
Below is a graph that shows the evolution of the energy use over time.



From this graph we can see that the maximum energy use occurs in January and more precisely the 1080Wh energy use occurs on the 16th of January at 6:50PM. Furthermore, from this graph we can see that a couple of missing observations are present in March. Finally, there are no clear energy use trends that we can identify from the above graph. Other analysis such as weekday energy versus weekend energy use did not provide any clear trend either (see in the Python script).

The correlation matrix (right) enables us to see interesting information:

- All the variables except v1, v2, RH-out, Press_mm_hg, RH-6 to RH-9 and RH-2 are positively correlated with Appliances, even though the coefficients are small.
- The largest correlation coefficient with Appliances comes from the feature lights
- Features v1 and v2 are the same features and have a correlation equal to 1. Furthermore, they are not correlated with any other feature in the dataset. These features will be removed in the preprocessing part.
- Except for v1 and v2, the highest correlation is between T6 and T-out which is logical since the sensor for T6 is located outside the building.
- Features such as visibility, windspeed and Pressure_mm_hg have low correlation with other variables of the dataset



MODELING

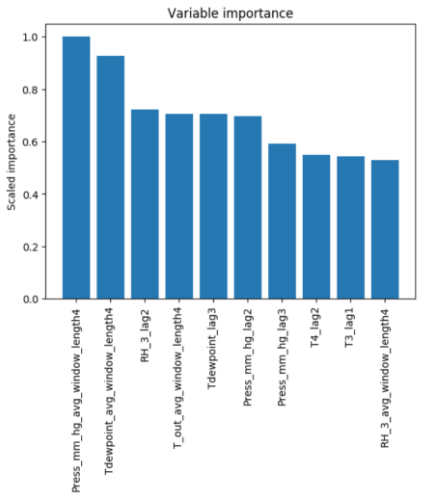
Data cleaning and preprocessing

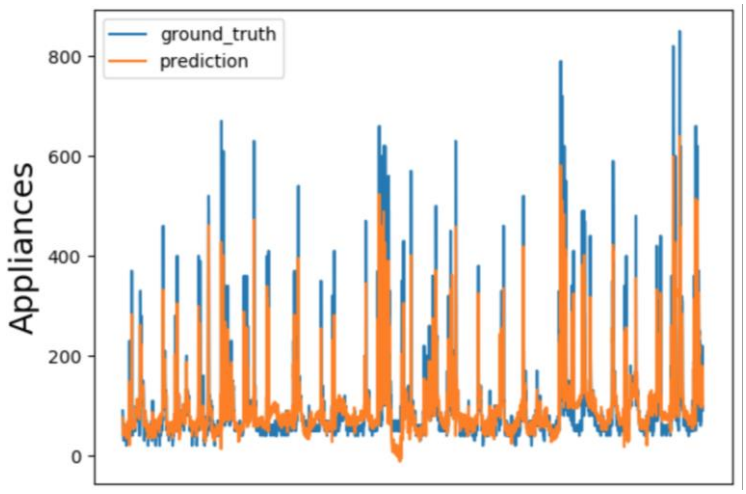
- The observations corresponding to the missing values of the target variable were removed while the missing values for the independent features were imputed by interpolation between the last observation and the next observation.
- In order to apply machine learning to this problem we must first transform the time series into a supervised problem. Independent variables at time t-3, t-2 and t-1 are used to predict the target variable at time t. The independent variables at lagged times are further aggregated by mean with windows of number of lags.
- The dataset was then split 80% training and 20% testing.
- H2O[1] is the machine learning library that was used to do the modeling.
- Since more than 150 variables were created to transform the time series problem into a supervised problem, feature selection with LASSO was achieved to keep only the features that have an importance > 0.

From the table below, the best model is a Generalized Linear Model (GLM) with an **RMSE of 57.13** on the test data.

model_id	mean_residual_deviance	rmse	mse	mae	rmsle
GLM_grid_1_AutoML_20191021_165522_model_1	3263.97	57.1312	3263.97	29.741	nan
GBM_4_AutoML_20191021_165522	4180.99	64.6606	4180.99	39.2769	0.445075
StackedEnsemble_BestOfFamily_AutoML_20191021_165522	4508.38	67.1445	4508.38	48.1209	0.522532
GBM_2_AutoML_20191021_165522	5019.72	70.85	5019.72	47.9066	0.519196

Below on the right is a chart with the top 10 most important variables. On the left, is a chart of predictions versus ground truth.





The risks of applying machine learning methods to time series data is that usually machine learning methods in a supervised problem do not take into account the notion of time and how features evolve over time which is very important. Therefore, results may be wrong or strongly biased if the problem is not converted into a supervised problem appropriately.

BENEFITS

I believe that for energy providers it is important to predict a household's electricity usage because it facilitates electricity demand management and utilities load planning.

I believe it can also enable home owners to keep track of their power usage and therefore adapt their consumption according to periods of the year and therefore potentially reduce their electricity bill. Devices such as smart electricity metering technologies gather a huge amount of consumption data on daily and hourly basis and may enable a house owner to keep track of its consumption and potentially reduce ineffective energy usage.

[1] http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html