CAPSTONE PROJECT

Regression on Appliances Energy Prediction

CONTRIBUTOR:

- Yogesh K

Table of Contents

Machine Learning in Energy Prediction

- Regression models for energy use can help to understand the relationships betweendifferent variables and to quantify their impact.

- Prediction models of electrical energy consumption in buildings can be useful for anumber of applications:
  - to determine adequate sizing of photovoltaics and energy storage to diminishpower flow into the grid .
  - to detect abnormal energy use patterns
  - to be part of an energy management system for load control
  - to model predictive control applications where the loads are needed
  - for demand side management (DSM) and demand side response (DSR)and as aninput for building performance simulation analysis.

Problem Statement

Using different data sources and environmental parameters (indoor and outdoor conditions),specifically, data from a nearby airport weather station,temperature and humidity in different rooms in the house from a wireless sensor network and one sub-metered electrical energy consumption (lights) have been calculated. Our goal is to predict the energy use by appliances.

Understanding our data

- The data set is at 10 min for about 4.5 months. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network.
- The energy data was logged every 10 minutes with m-bus energy meters.
- Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru) and merged together with the experimental data sets using the date and time column.
- Two random variables have been included in the data set for testing the regression models and to filter out non-predictive attributes (parameters)
- Appliances, energy use in Wh (Dependent variable) - This is our Target Variable

Understanding the Data – (contd)

Temperature:

- T1, Temperature in kitchen area, in Celsius
- T2, Temperature in living room area, in Celsius
- T3, Temperature in laundry room area
- T4, Temperature in office room, in Celsius
- T5, Temperature in bathroom, in Celsius
- T6, Temperature outside the building (north side), in Celsius
- T7, Temperature in ironing room , in Celsius
- T8, Temperature in teenager room 2, in Celsius
- T9, Temperature in parents room, in Celsius
- Tout, Temperature outside (from Chievres weather station), in Celsius

Relative Humidity:

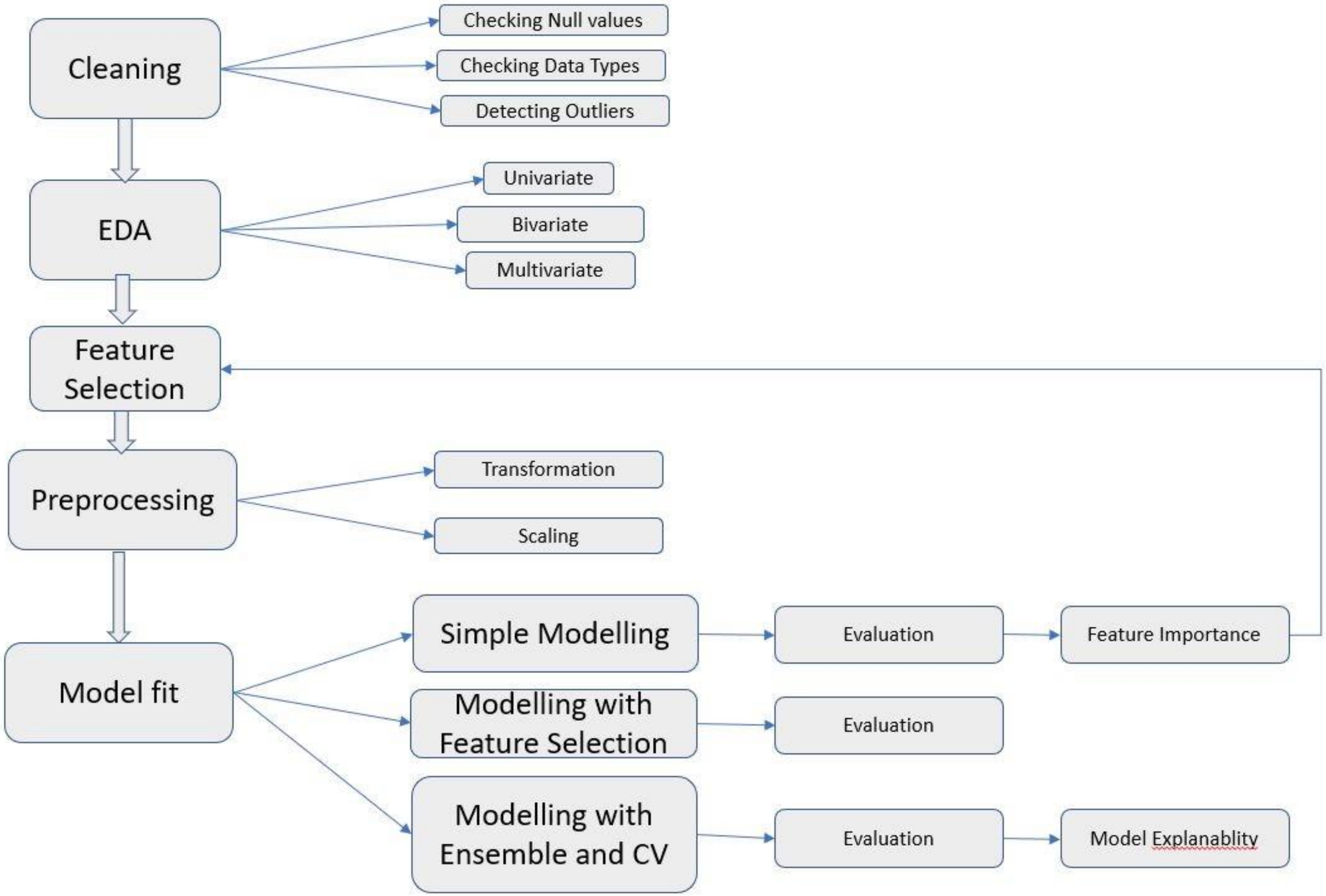
- RH_1, Relative Humidity in kitchen area, in %
- RH_2, Relative Humidity in living room area, in %
- RH_3, Relative Humidity in laundry room area , in %
- RH_4, Relative Humidity in office room, in %
- RH_5, Relative Humidity in bathroom, in %
- RH_6, Relative Humidity outside the building (north side), in %
- RH_7, Relative Humidity in ironing room , in %
- RH_8, Relative Humidity in teenager room 2, in %
- RH_9, Relative Humidity in parents room, in %
- RH_out, Relative Humidity outside (from Chievres weather station), in %

Understanding the Data – (contd)

Other Variables:

- lights, energy use of light fixtures in the house in Wh (Drop this column)
- Pressure (from Chievres weather station), in mmHg RHout
- Wind speed (from Chievres weather station), in m/s
- Visibility (from Chievres weather station), in km
- Tdewpoint (from Chievres weather station), Â°C
- rv1, Random variable 1, nondimensional
- rv2, Random variable 2, nondimensional
- date time year-month-day hour:minute:second

Process Outline

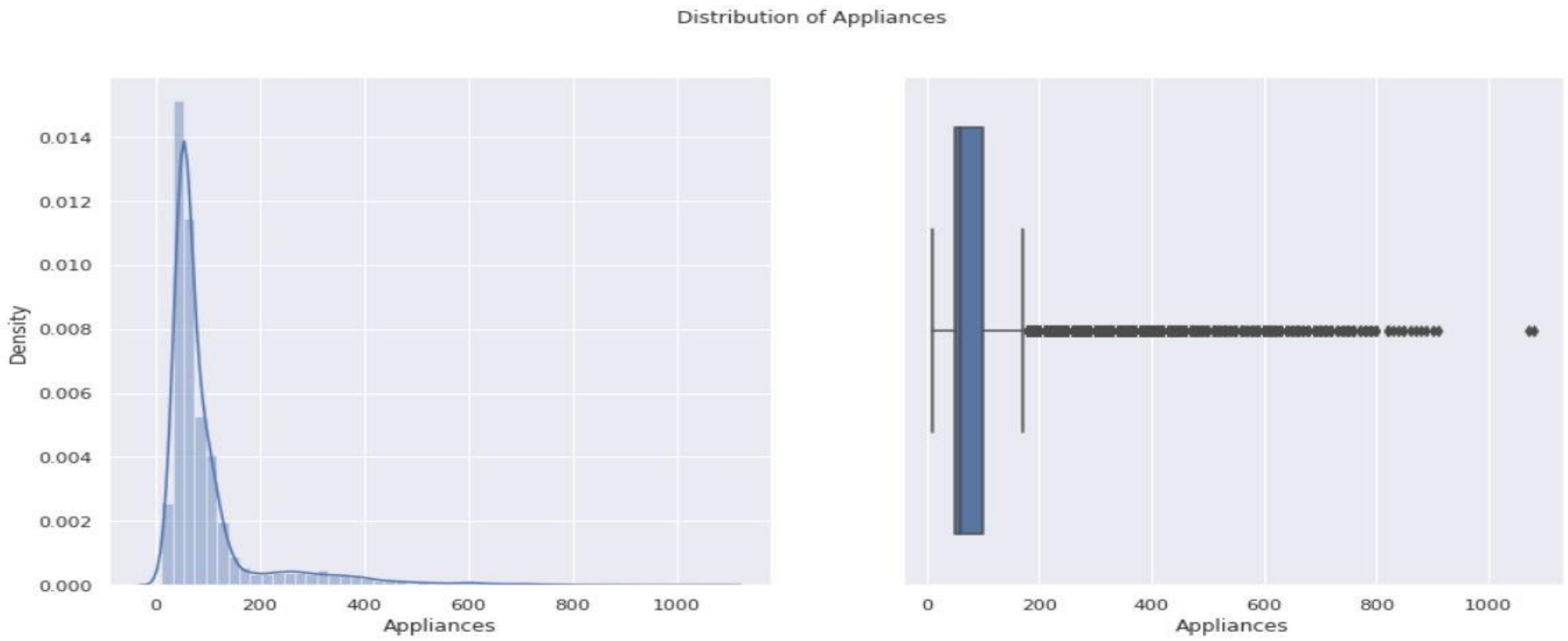


Exploratory Data Analysis

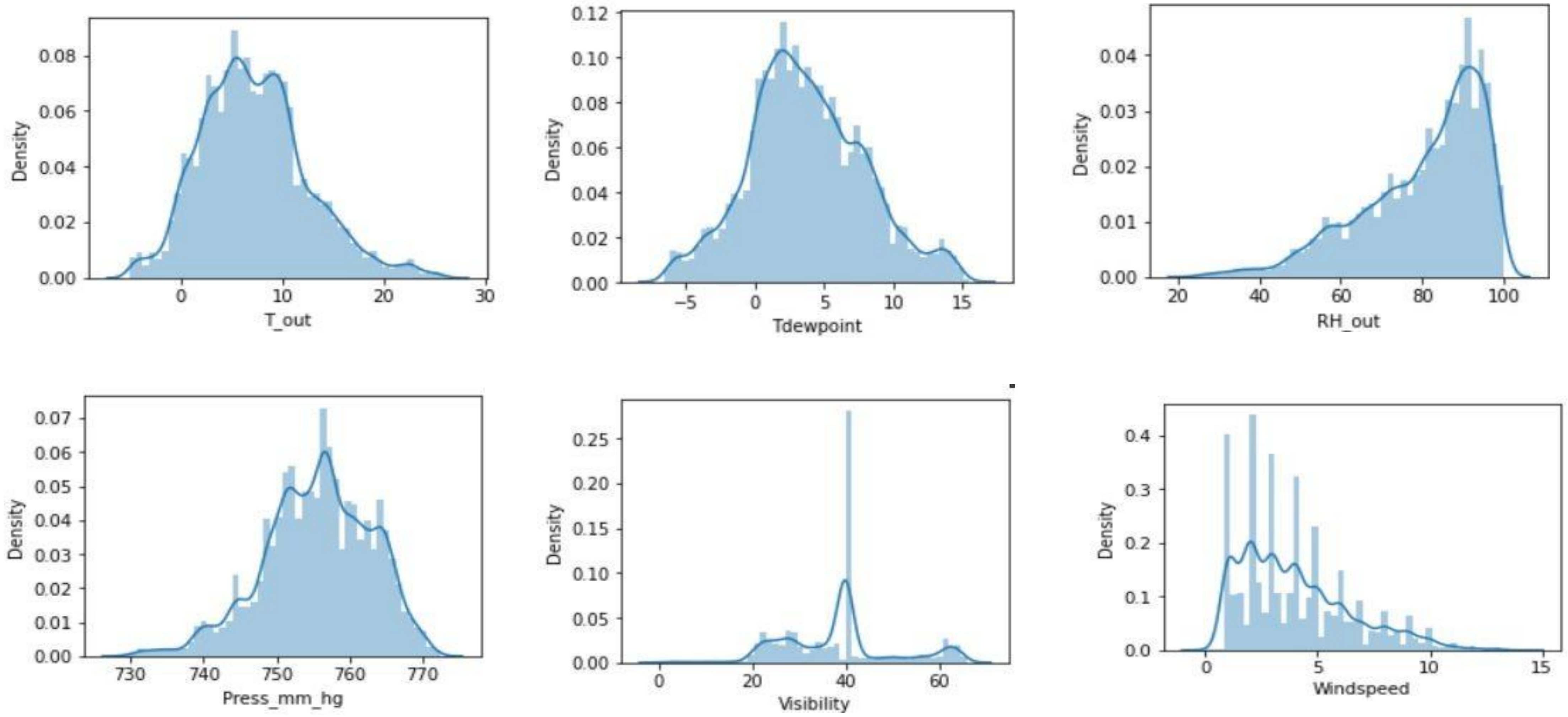
Target Variable Distribution

- We Observed that our Target Variable has a Right skewness
- Also we can observe that there are outliers in our target variable
- The target variable has most values less than 200Wh, showing that high energy consumption cases are very low.
- We have removed the outliers based on the Inter Quartile Values

	Appliances
count	19735.000000
mean	97.694958
std	102.524891
min	10.000000
25%	50.000000
50%	60.000000
75%	100.000000
max	1080.000000



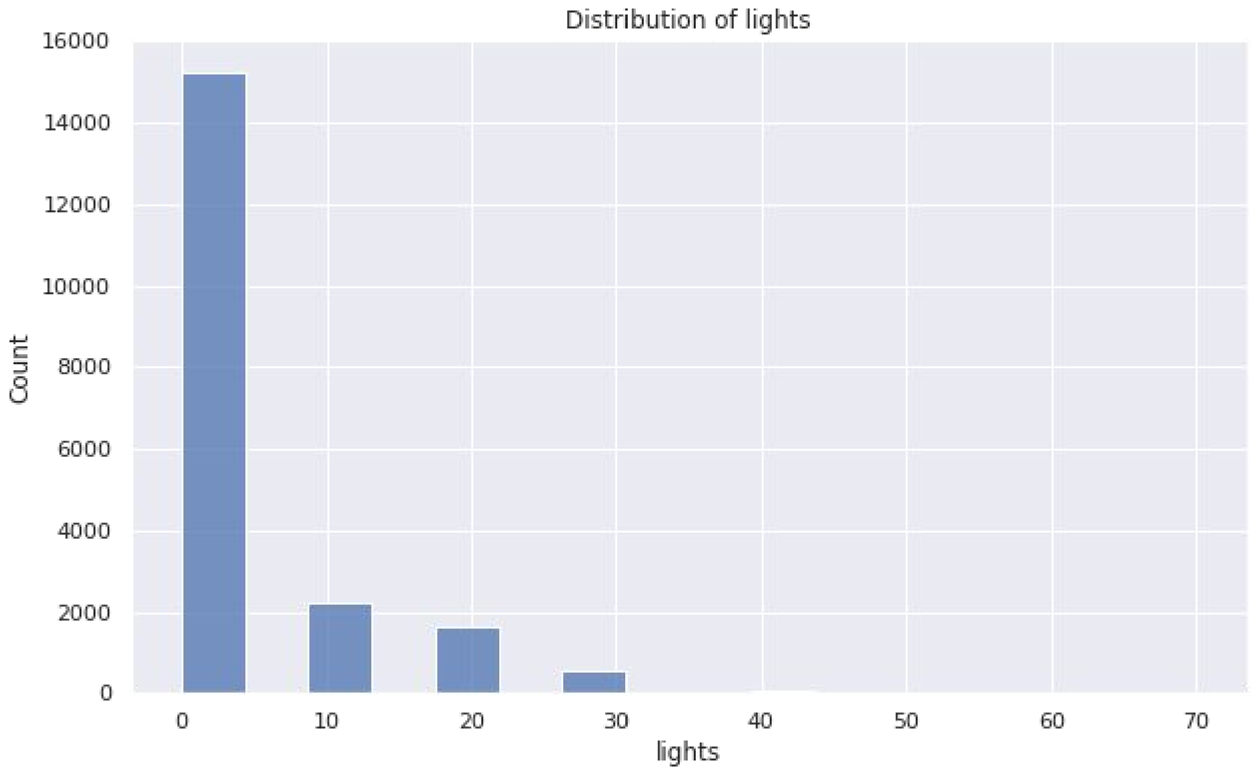
Univariate Analysis



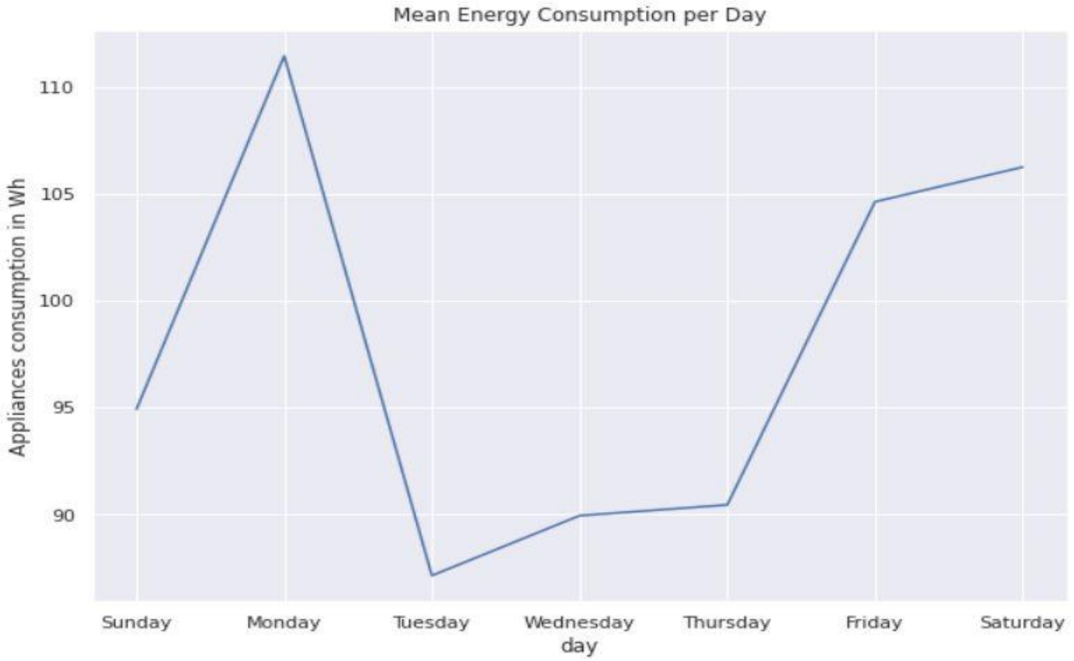
- We can see that Visibility, Wind Speed and RH_out are skewed.
- The other variables - Temperature and Humidity followed mostly normal distributions

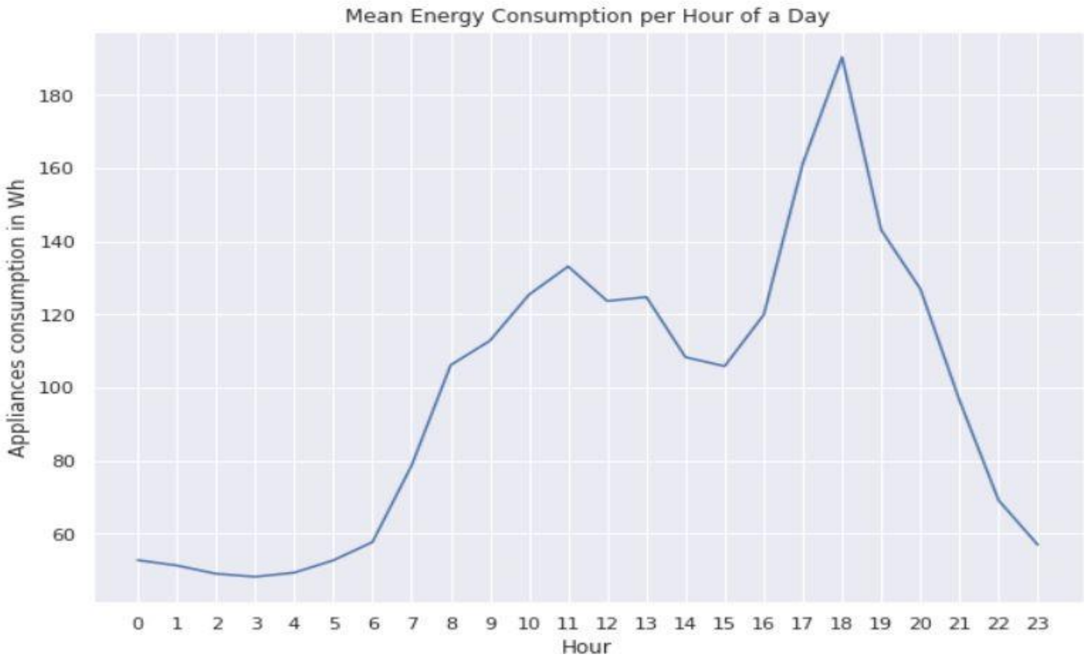
Lights variable

- Light column has 15252 entries with value = 0.
- It could mean there is no human presence in that room at that time.
- Or it could be during the day where lights are not turned on; or it could beduring the night when lights are turnedoff.



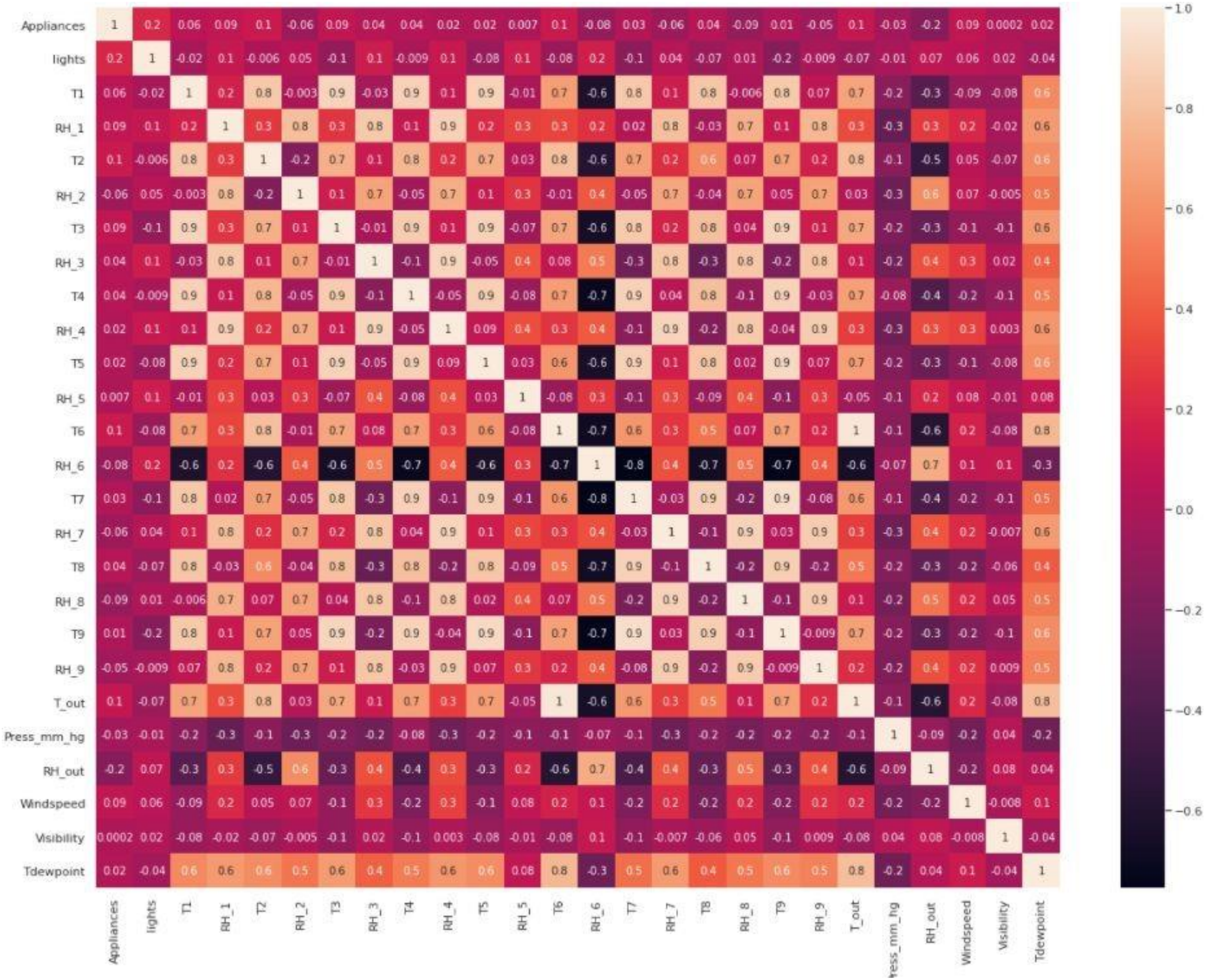
Bivariate Analysis





- During night time appliance usage is low.
- High consumption during morning hours.
- And it peaks during the evening.
- Energy consumption is high on weekends and low during the weekdays.

Multivariate Analysis - Correlation



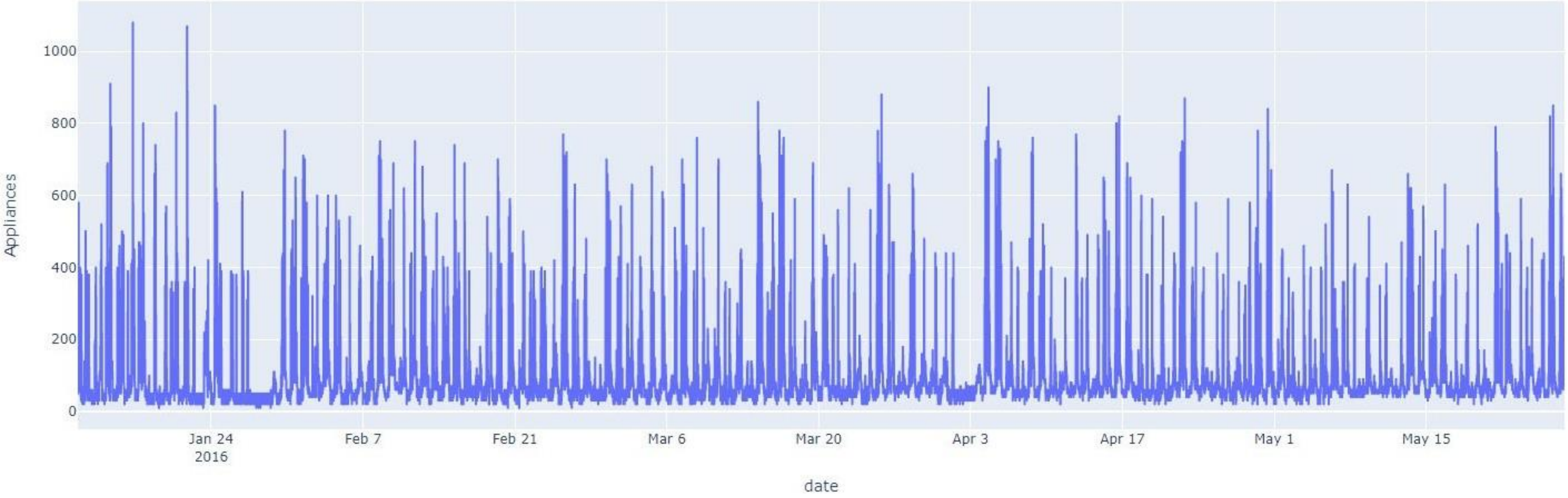
- None of the variables are highly correlated with the target variable.
- Correlations between indoor temperature and humidity is high as expected.
- T_out and T6 have a correlation of1 - both are the outside temperatures.
- Similarly RH_out and RH_6 are outside humidity. They have a high positive correlation of 0.7.
- RH_6 has a negative correlation with the indoor temperatures and also outdoor temperature. This is expected as temperature and relative humidity are expected to beinversely proportional.

Checking Correlation for Target Variable

feature	correlation
Appliances	1.000000
hour	0.416503
lights	0.291109
T8	0.268293
T2	0.264739
T1	0.248221
avg_temp	0.242225
T6	0.223875
T_out	0.213651
T4	0.195689
T5	0.191782
T3	0.180061
T7	0.175519
T9	0.154471
Tdewpoint	0.081550
RH_5	0.072040

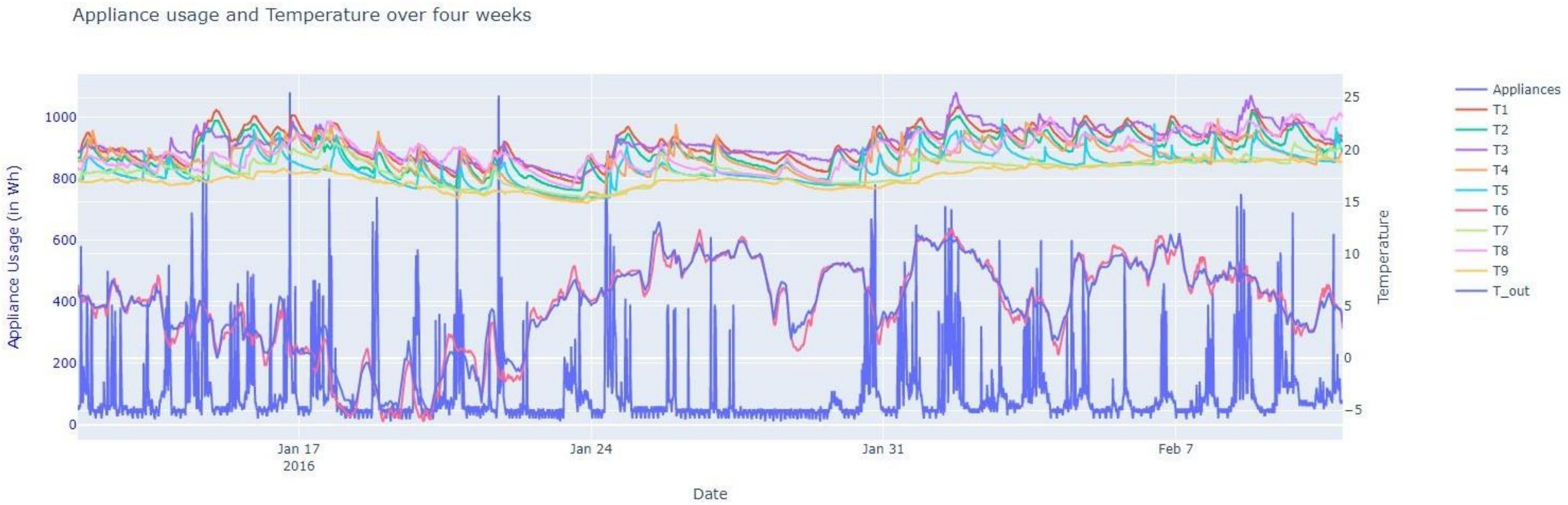
- Hour has been extracted from the date column. It is the number of hours after midnight
- T8 - Temperature in teenager room
- T2 - Temperature in living room
- T1 - Temperature in kitchen
- We do not see any significant correlation between any of our features and our target variable

Appliance usage



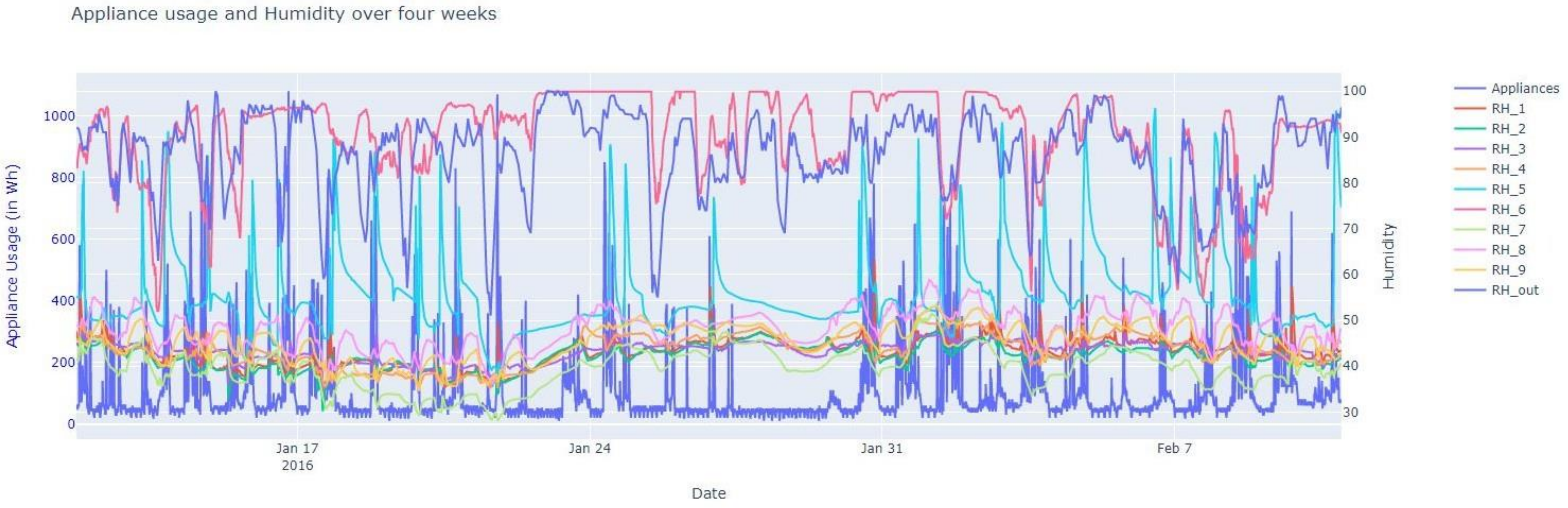
We can see that there peaks of high appliance usage and low appliance usage. This could indicates morning, noon and night

Appliance and Temperature



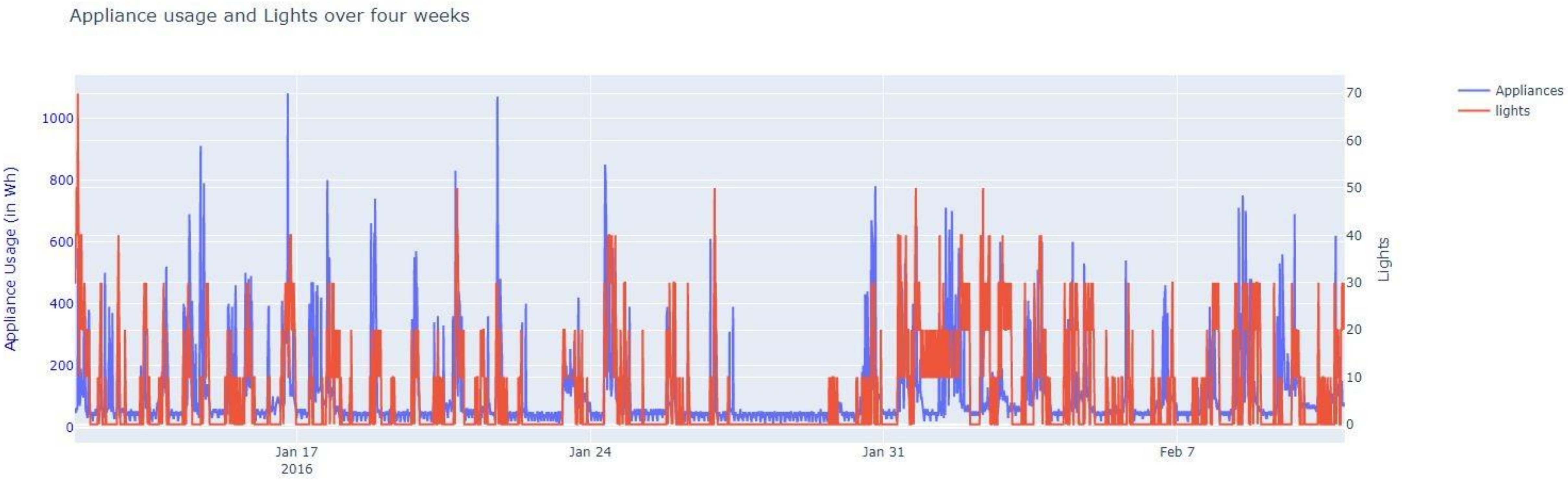
Excluding T6, T_out and T9, we can see that the temperature inside slightly goes up when the appliance usage is at itspeak.

Appliance and Humidity



RH5 which is the humidity in bathroom peaks when bathroom is in use - due to hot water usage during bathing.

Appliance and Lights



Light usage and appliances usage almost have the same peaks.

Evaluation Metrics

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \bar{y})^2}$$

- Root Mean Squared Error (RMSE) - is a standard way to measure the error of a model in predicting quantitative data
- R2 - compares our model with the baseline model. It is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

Modelling

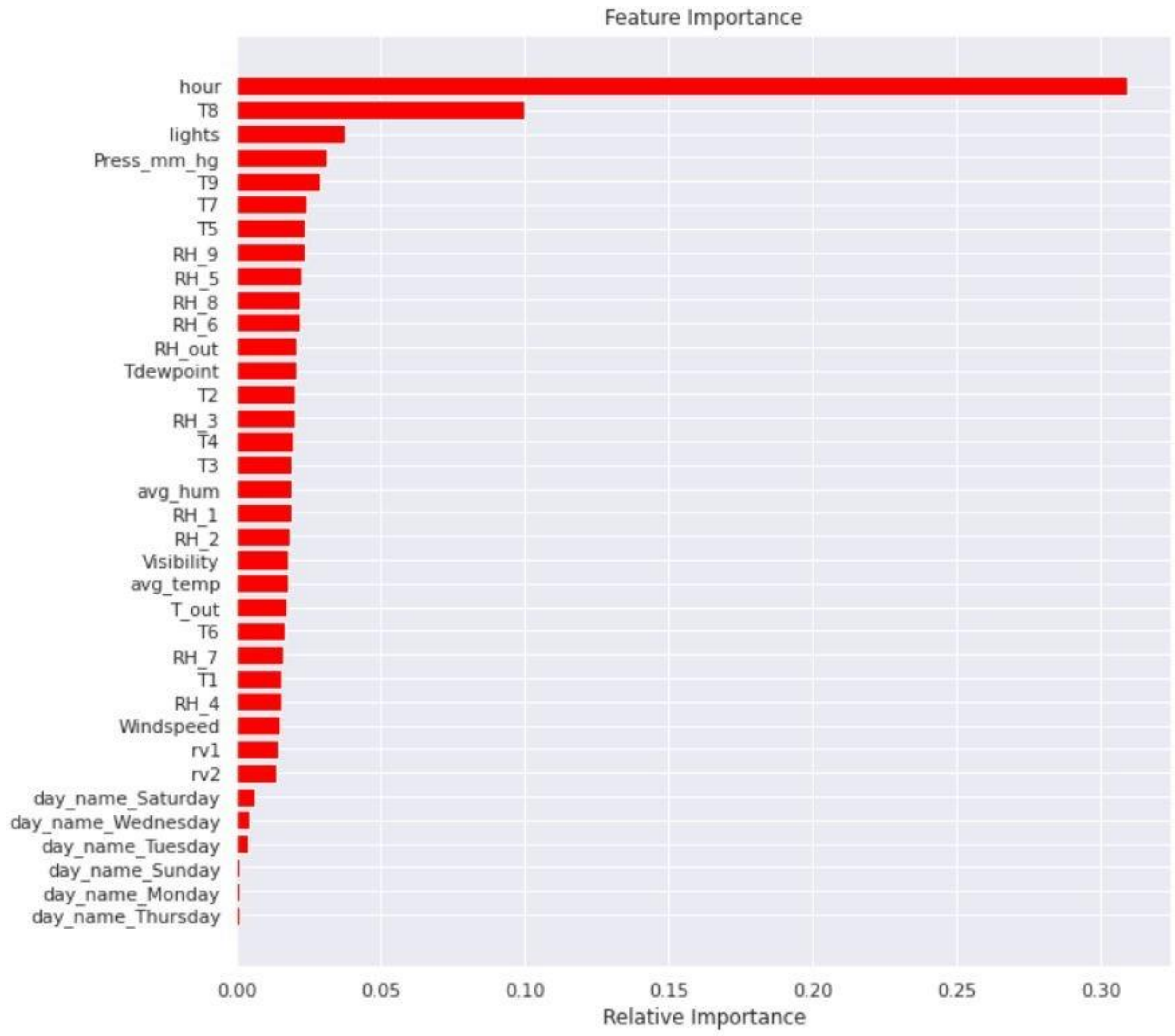
Modelling using Simple Models

Model_Name	Train RMSE	Test RMSE	Train R2	Test R2
LinearRegression	22.962356	23.492512	0.347200	0.326128
Ridge	22.962338	23.492431	0.347201	0.326132
SVR	16.220864	18.933349	0.674242	0.562303
RandomForestRegressor	6.494434	16.439150	0.947781	0.670028
GradientBoostingRegressor	18.811366	19.982732	0.561885	0.512440
XGBRegressor	18.816123	19.966047	0.561663	0.513254

- From here we see that Random Forest fits best on the data
- We can use this model to get the feature importance of variables

Feature Importance

- Based on this result , we can drop some columns
- T6 and T_out have a high positive correlation of 1. So we can drop T_out.
- We can drop Visibility and Windspeed based on low feature importance.
- We can see that the days of the week have very low importance.
- We are going to keep all other features as we want to see which rooms in the house are significant.



Modelling after dropping T_out, Visibility and Wind speed

Model_Name	Train RMSE	Test RMSE	Train R2	Test R2
LinearRegression	23.092101	23.628237	0.339802	0.318319
Ridge	23.092067	23.627664	0.339804	0.318352
SVR	16.889243	18.885483	0.646843	0.564513
RandomForestRegressor	6.482356	16.337665	0.947975	0.674089
GradientBoostingRegressor	18.929761	20.078029	0.556353	0.507778
XGBRegressor	18.916995	20.063821	0.556951	0.508475

- We observe that scores only slightly improve.
- Linear Regression and Ridge are the worst performing models as we didn't see any significant correlation between independent variables and the target variable
- We further perform hyperparameter tuning on these models to get the best scores

Modelling Using Ensemble and Cross Validation

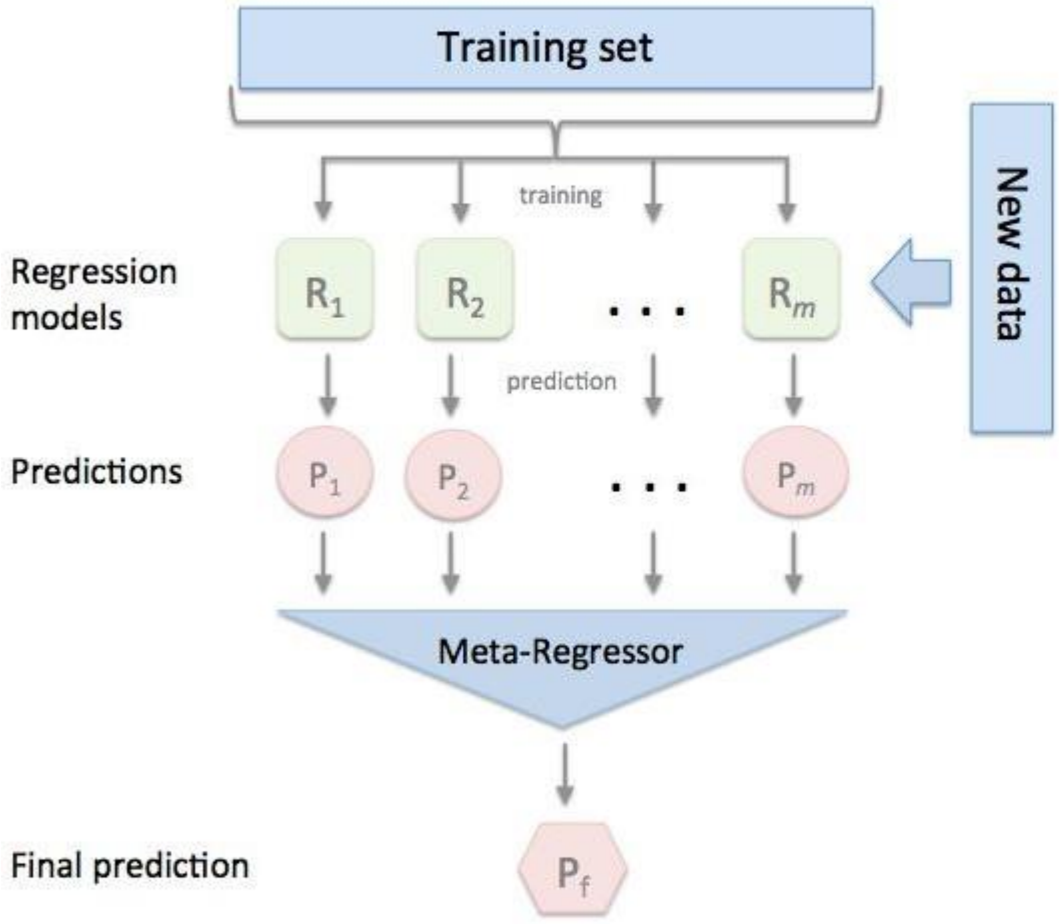
Model_Name	Train RMSE	Test RMSE	Train R2	Test R2
SVR	15.139850	18.374532	0.716214	0.587759
RandomForestRegressor	20.156087	21.014896	0.497009	0.460771
GradientBoostingRegressor	13.887269	17.873772	0.761229	0.609922
XGBRegressor	14.639507	18.042222	0.734661	0.602535
VotingRegressor	16.920032	18.944033	0.645554	0.561809
StackingRegressor	8.920476	16.328663	0.901480	0.674448

- After hyper parameter tuning, we can see that the overfitting has reduced.
- Voting Regressor (using weighted average) - did not give better results.
- We see best results with the Stacking Regressor.

Suggested Model

Stacking Regressor

```
stacking =
StackingRegressor(estimators=['SVR',
'Random Forest Regressor','Gradient
Boosting Regressor'],
final_estimator=xgbr, cv=5)
```



Summary

- As the first step, we understand the data & checked for null values, and outliers and performed EDA to get better understanding of variables .

- As part of data pre-processing, we performed feature scaling and outlier removal

- As so we have a Timestamp in our data, we needed to see the periodicity and trend of our dependent and independent variables.

- We tried multiple simple models and multiple advanced models with performed hyper parameter tuning and cross validation.

- Models Built: Linear Regression, SVR, RandomForest, Gradient Boosting XGBoost

- Advanced Models: Stacking Regressor, Voting Regressor, Average Ensemble

- Based on our targeted evaluation metric - RMSE and R2 scorel, we chose Stacking Regressor as the suggested model.

Limitations

- We saw that temperatures and humidity in different rooms are highly correlated. It might be sufficient tomeasure the temperature and humidity from the most representative rooms.
- The data consists measurements for only one house. If data was collected for multiple house, we could have captured more variance in our data.
- We do not know the number of occupants or the kind of appliances in use. This information might give us better insights on energy consumption.
- The data is only for 4.5 months. Different consumption patterns can be found depending on the different seasons in a year.
- We have considered this as a regression problem. As a further exploration, we can also perform time series analysis.

Q / A