Capstone Project

Appliance Energy Prediction

by

Mahin Arvind Chanthira Sekaran

PROBLEM STATEMENT

The Appliance Energy Prediction data provides temperature, pressure and relevant weather data for every 10 min for about 4.5 months for a specific household.

The objective is to build a predictive model which could help in predicting the total appliance energy consumption in the house proactively with the intention of offering some intuition in curbing power consumption.

Contents

Methodology

- Exploratory Data Analysis
- Feature Engineering and Selection
- Feature Scaling and Standardization
- Choosing a Metric
- Spot Checking Regression Algorithm
- HyperParameter Tuning
- Model Explainability

Data Description

- The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. Each wireless node transmitted the temperature and humidity conditions around 3.3 min.

- This wireless data was averaged for 10 minutes periods. 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).

Data Description

- **The Appliance Energy Prediction dataset was made available with 28 features. The names and descriptions of the 28 features have been listed below.**

Attribute	dtype	Nature	Description
date	Object	Categorical	time year-month-day of recording
Appliances	int	numerical continuous	energy use in Wh
T_out	float	numerical continuous	Temperature outside (from Chievres weather station), in Celsius
T1,T2,T3,T4,T5,T6,T7,T8,T9	float	numerical continuous	Temperature in kitchen,living room, laundry, office,bathroom, outside building, ironing room, teen room and parents room
RHout	float	numerical continuous	Humidity outside (from Chievres weather station), in %
RH_1,RH_2,RH_3,RH_4,RH_5, RH_6,RH_7,RH_8,RH_9	float	numerical continuous	Humidity in kitchen,living room, laundry, office,bathroom, outside building, ironing room, teen room and parents room

Data Description

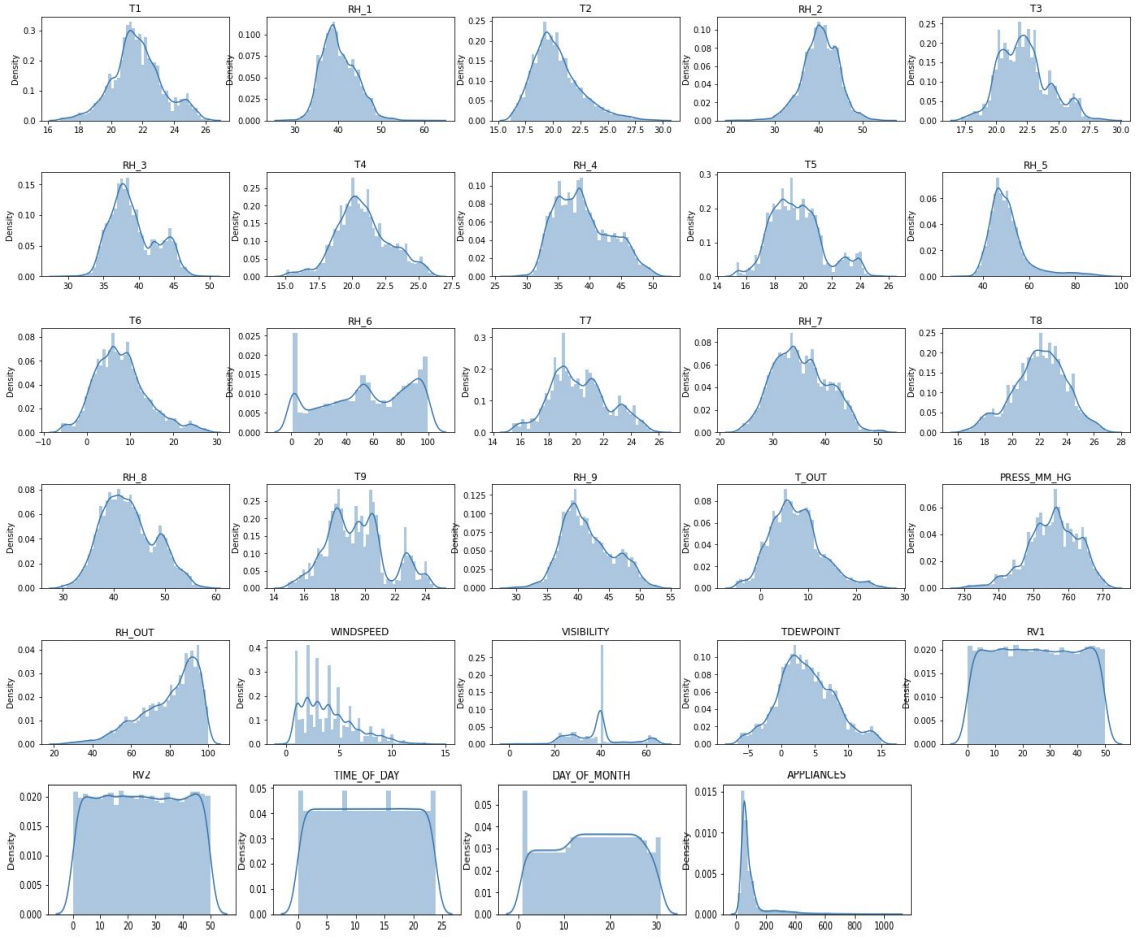
Attribute	dtype	Nature	
Tdewpoint	float	numerical continuous	(from Chievres weather station), Â°C
Windspeed	float	numerical continuous	windspeed (from Chievres weather station), in m/s
Visibility	float	numerical continuous	Visibility in km through Fog (from Chievres weather station)
rv1	float	numerical continuous	Random variable 1, nondimensional
rv2	float	numerical continuous	Random variable 2, nondimensional
Press_mm_hg	float	numerical continuous	Pressure (from Chievres weather station), in mmHg

Exploratory Data Analysis

- Univariate Analysis
  - Distribution of Numerical Features
  - Outliers in Numerical Features
  - Distribution of Categorical Features
- Bivariate Analysis
  - Correlation Heatmap
  - Appliance Energy Consumption throughout the timeline of the dataset
  - Hourly Appliance Energy Consumption Trends
  - Hourly Trends in Temperature
  - Hourly Trends in Humidity

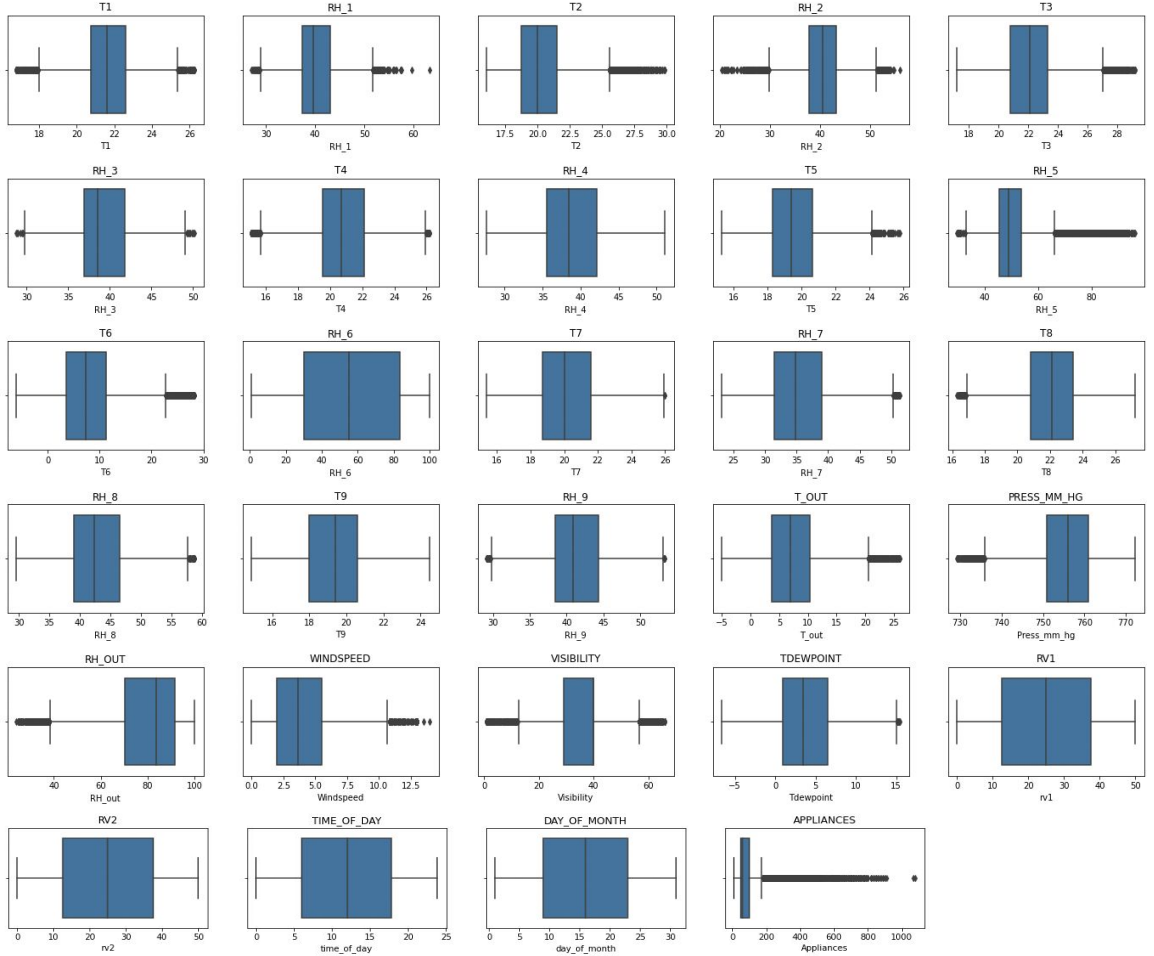
Distribution of Numerical Features

- Temperature and Humidity attributes have a gaussian-like distribution.
- The target variable ‘Appliances’ has a skewed gaussian distribution indicating a wide range of outliers over the 3rd quartile



Outliers in Numerical Features

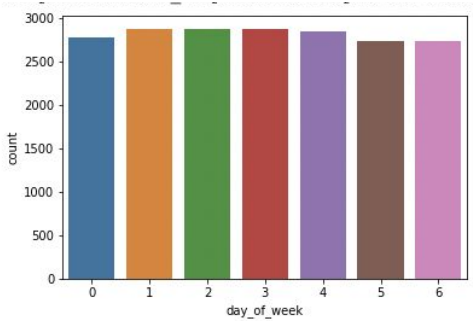
- T1, T2, T3,T4, T5,T6,T7,T8 and T_OUT have outliers.
- RH_1,RH_2,RH_3,RH_5, RH_7,RH_8,RH_9 and RH_OUT have outliers
- Windspeed, Tdewpoint, Visibility and the target variable Appliances also have outliers

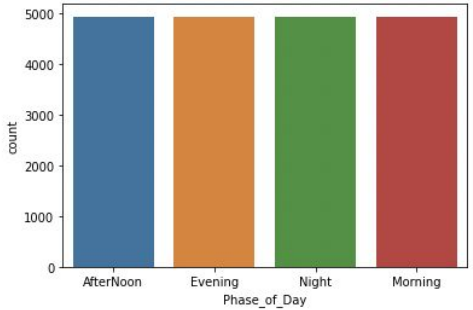


Distribution of Categorical Features

- From the date attribute, categorical features like day_of_week and Phase_of_Day were extracted as they are expected to have an effect on household appliance energy consumption

- As these attributes are equally recurring, we observe an even distribution for each category.



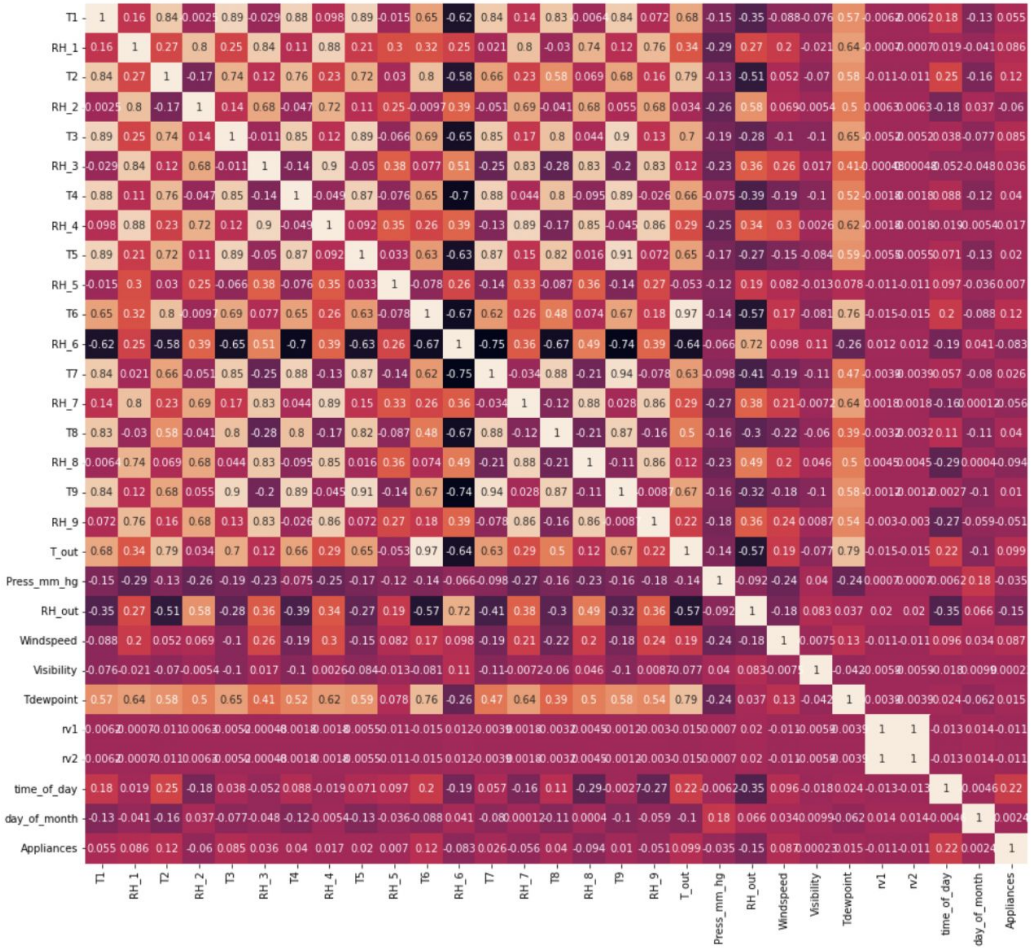


Correlation Heatmap

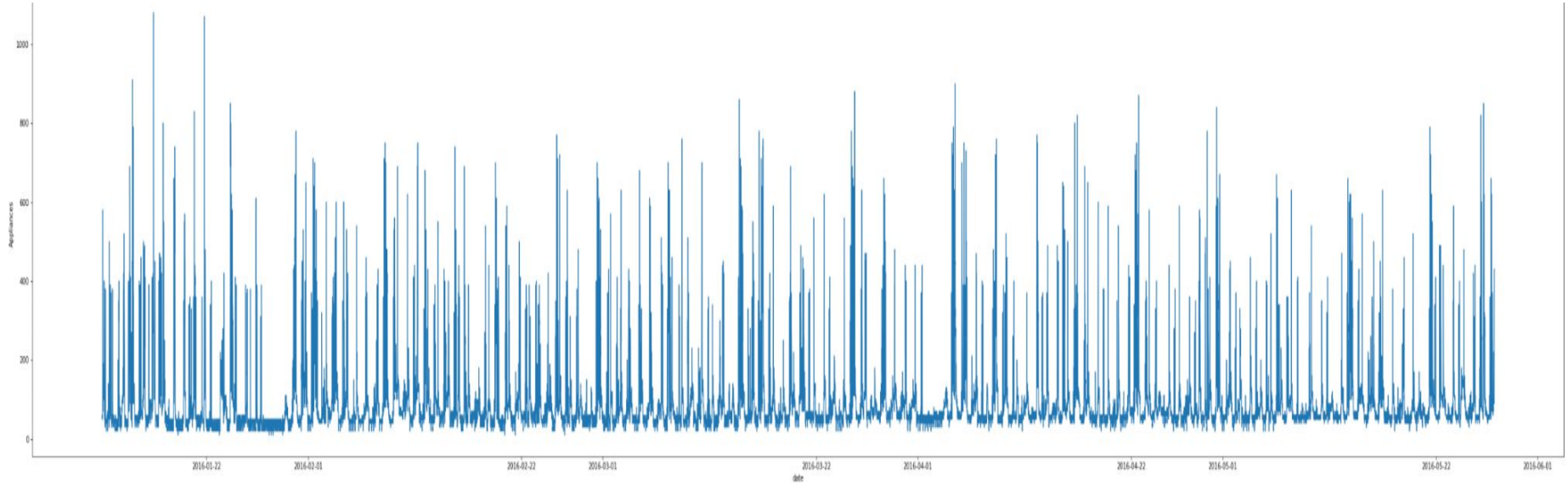
- We see strong correlation among temperature variables as change in outside heat can be experienced by all rooms except when changed only by human intervention such as use of thermostat, heaters etc.

- There is also a strong inverse correlation observed between RH_6 and all temperature features. This is because RH_6 is the outside humidity.

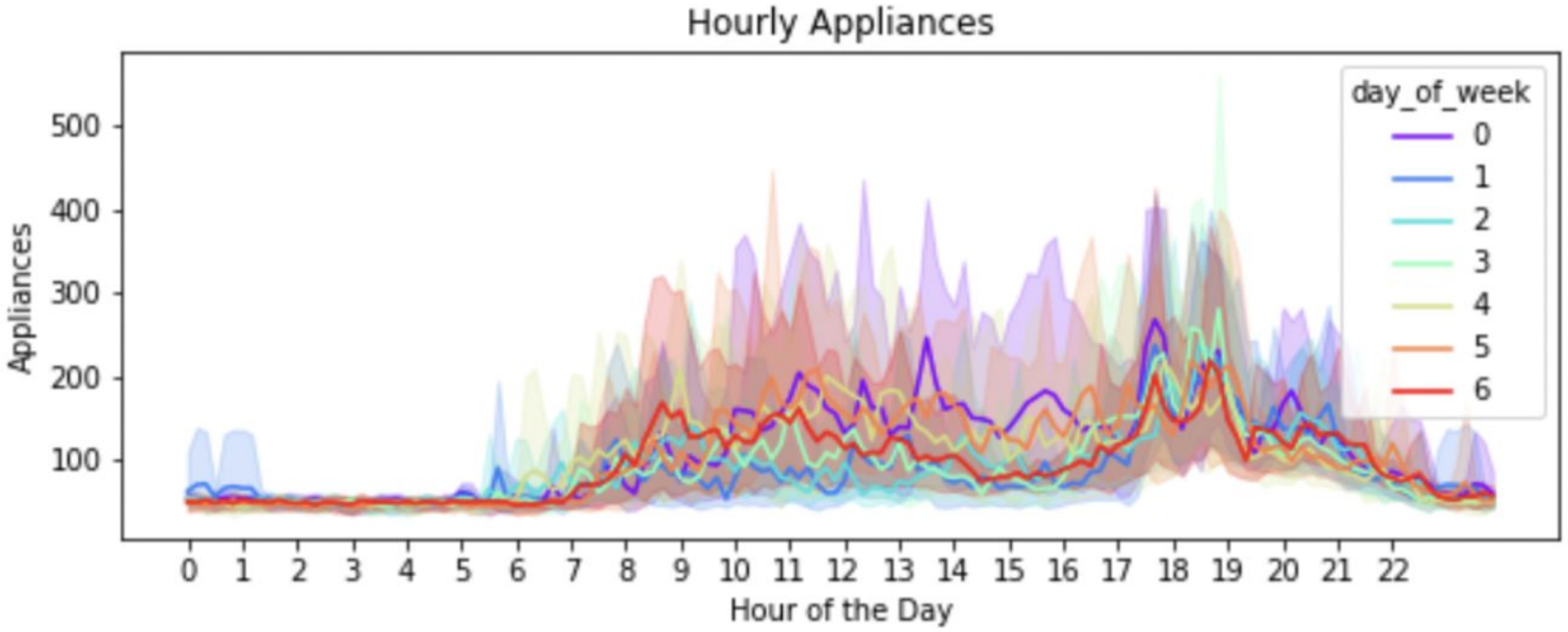
- As air temperature increases, air can hold more water molecules, and its relative humidity decreases



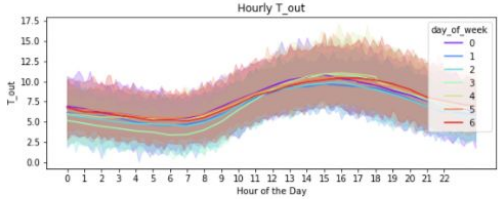
Appliance Energy Consumption

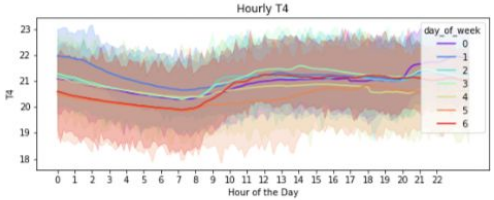


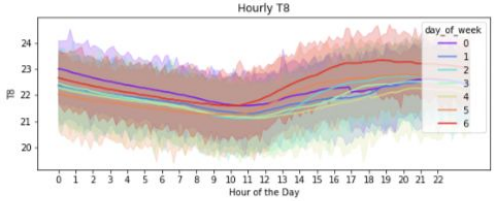
Hourly Appliance Energy Consumption Trends

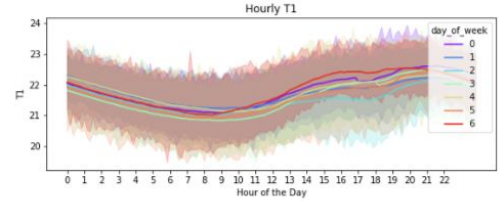


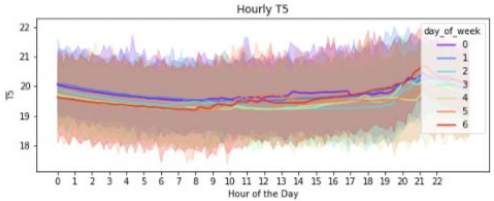
Hourly Trends in Temperature

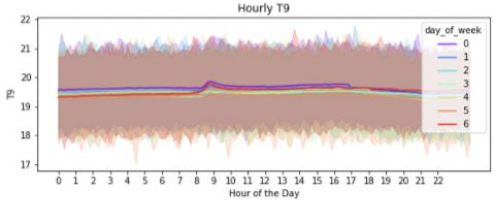


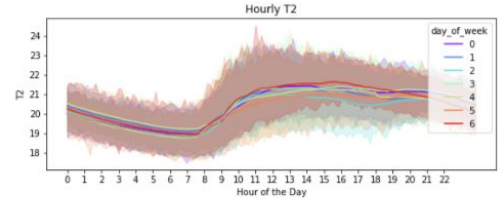


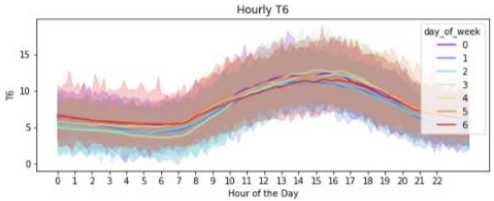


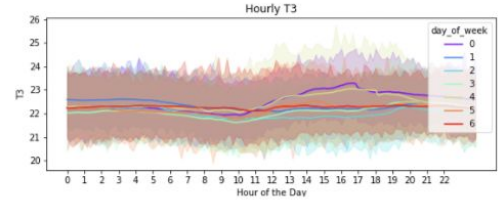


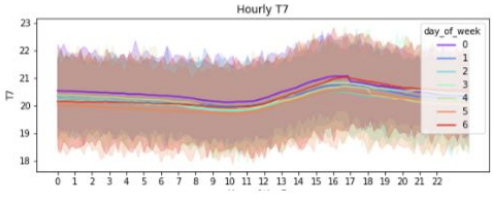




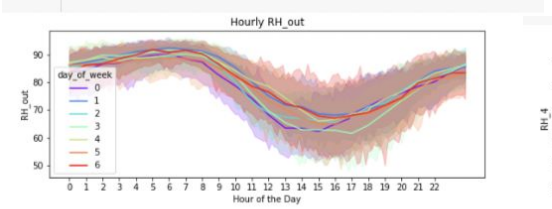


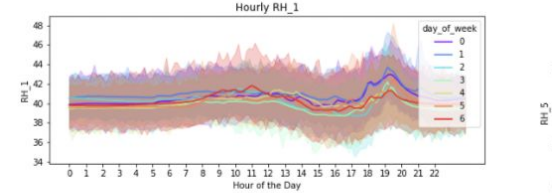


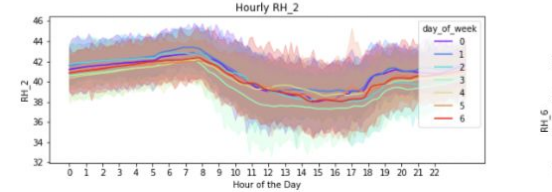


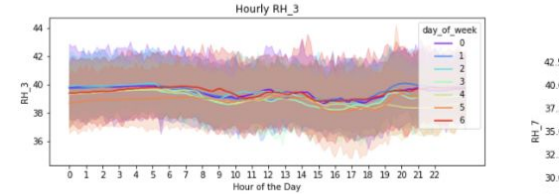


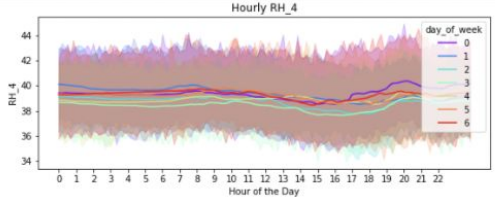
Hourly Trends in Humidity

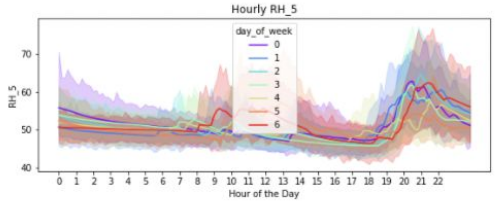


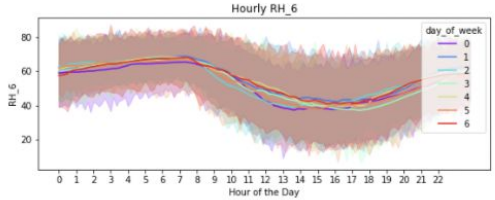


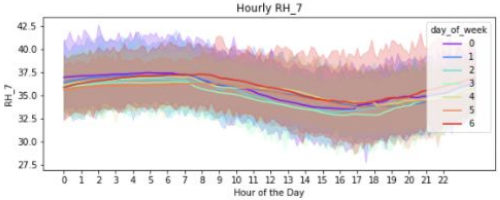


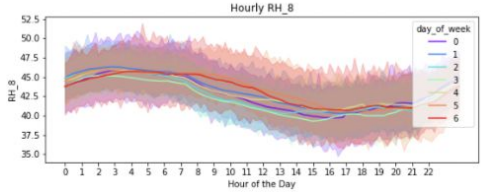


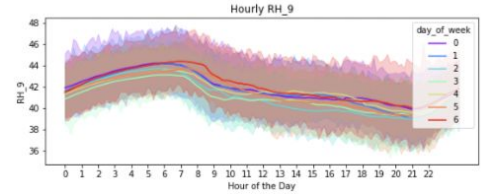












Feature Engineering and Selection

- Information gathered from the previous step is used to transform the features so that it could be learned efficiently by the deployed models in the later stages.
- In this step, the ‘date’ feature has been converted from object type to datetime and features like ‘phase of the day’ (morning, afternoon, evening, night), day of the week and hour of the day were extracted.
- As the time of the day is of cyclic nature, sin and cosine transformations were added to reflect this nature on the dataset.
- The outliers in continuous numerical features were handled by capping the data at the first and third quartile.
- Categorical features like ‘day of the week’ and ‘phase of the day’ were One Hot Encoded.

Feature Scaling and Standardization

- The attributes are rescaled using StandardScaler function from sklearn to have zero mean and unit variance.

- This is done in order to bring down all the features to a common scale without distorting the differences in the range of the values. This will ensure there is equal contribution to the analysis.

- This is done especially for training data in gradient descent based algorithms and distance based algorithms like K-NN.

Metrics

- **Root Mean Squared Error: RMSE is calculated as the square root of the mean of the squared differences between actual outcomes and predictions. Squaring each error forces the values to be positive, and the square root of the mean squared error returns the error metric back to the original units for comparison.**

- **Mean Absolute Error:MAE is calculated as the average of the absolute error of the actual values and the predicted values**

- **R-Squared: The R2 (or R Squared) metric provides an indication of the goodness of fit of a set of predictions to the actual values. In statistical literature this measure is called the coefficient of determination.**
  This is a value between 0 and 1, 0 for no-fit and 1 for perfect fit respectively.

- **Adjusted R-Squared: The adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a regression model. Since R2 always increases as you add more predictors to a model, adjusted R2 can serve as a metric that tells you how useful a model is, adjusted for the number of predictors in a model.**

Random Variable Evaluation

- Upon training the linear models, namely Linear Regression, Lasso Regression and Ridge Regression, the weights assigned to the arbitrarily added random variable, ‘rv1’ is checked.

S.No	Model	Weight for ‘rv1’
1	Linear Regression	0.014875
2	Lasso Regression	0.008851
3	Ridge Regression	-0.000000

- We can observe that Linear models have assigned near zero weights to the random variable, negating its influence in prediction of the target variable.

Spot Checking Regression Algorithm

Model Evaluation Report

S.No	Model Name	RMSE	MAE	R-Squared	Adjusted R2
1	Linear Regression	34.89304	25.731465	0.357475	0.268976
2	Lasso Regression	35.144995	25.915871	0.348162	0.251728
3	Ridge Regression	34.893478	25.731677	0.357459	0.268933
4	Decision Tree Regressor	27.866566	15.825944	0.590193	0.521114
5	Random Forest Regressor	21.003733	13.983722	0.767188	0.706933
6	Adaptive Boosting Regressor	37.818091	30.081382	0.245235	0.229025
7	Gradient Boosting Regressor	31.692642	22.800431	0.469935	0.356732
8	Bagging Regressor	22.361168	14.660882	0.736123	0.670872
9	K Neighbors Regressor	25.375333	16.581454	0.66019	0.579966
10	Linear SVM	36.679158	24.78428	0.290012	0.144261

HyperParameter Tuning

- In order to improve the performance of Random Forest Regressor, RandomSearchCV was used to optimize its parameters. The parameters that were in the search space were:
- The parameters that were in the search space were:

  'bootstrap':Whether bootstrap samples are used when building trees.

  'criterion': function to measure the quality of a split

  'max_depth': The maximum depth of the tree

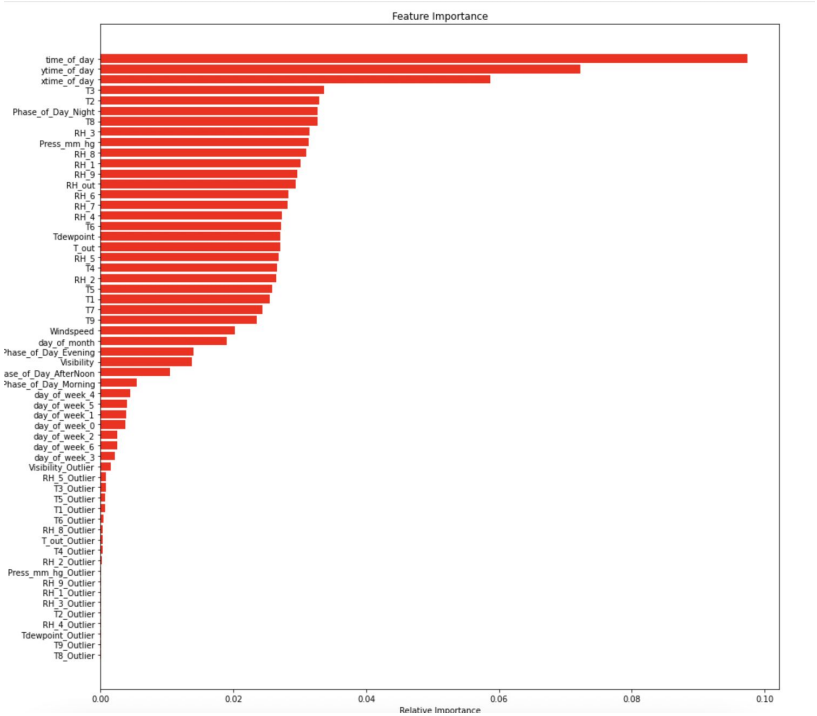
  'max_features':The number of features to consider when looking for the best split

  'n_estimator':The number of trees in the forest
- The best parameters were: 'bootstrap': True,'criterion': 'squared_error', 'max_depth': None, 'max_features':'log2', and 'n_estimators': 30
- After imputation of these values into the Random Forest Regressor model, the Tuned Random Forest Regressor R-squared score increased from 0.767 to 0.771.

Model Explainability

- Feature importance scores provide insight into the data and the deployed model.

- By looking at the Feature Importance graphs and the contribution chart from ELI5, we can gather that the appliance energy consumption largely depends on the time of the day and temperatures in rooms 3 and 2.



Contribution[?]	Feature	Value
+78.663	<BIAS>	1.000
+1.541	T2	0.368
+0.476	RH_5	0.256
+0.367	RH_3	0.875
+0.301	RH_4	0.952
+0.163	day_of_month	-0.005
+0.069	RH_2	-0.269
+0.028	Visibility_Outlier	-0.358
+0.021	T_out_Outlier	-0.157
+0.019	day_of_week_3	-0.413
+0.015	Tdewpoint	-0.878
+0.008	T6_Outlier	-0.165
+0.006	T1_Outlier	-0.157
+0.005	T4_Outlier	-0.111
+0.004	T5_Outlier	-0.100
-0.000	T3_Outlier	-0.098
-0.001	RH_8_Outlier	-0.042
-0.002	RH_2_Outlier	-0.115
-0.004	Press_mm_hg_Outlier	-0.029
-0.010	day_of_week_1	-0.416
-0.012	day_of_week_0	-0.404
-0.016	day_of_week_4	-0.413
-0.024	RH_5_Outlier	-0.142
-0.030	Windspeed	-0.249
-0.033	T1	0.027
-0.067	T4	-0.616
-0.067	day_of_week_6	-0.402
-0.089	T5	-0.176
-0.098	day_of_week_5	2.514
-0.112	day_of_week_2	-0.412
-0.214	RH_7	1.212
-0.235	RH_1	-0.010
-0.359	Press_mm_hg	0.956
-0.481	RH_9	1.340
-0.486	T9	-0.767
-0.722	T3	-0.877
-0.780	RH_8	1.146
-0.794	T7	-0.715
-0.837	Phase_of_Day_Morning	-0.573
-0.847	T8	-1.088
-0.862	xtime_of_day	0.303
-0.871	Visibility	-1.778
-0.895	T6	-1.531
-0.922	RH_out	1.293
-0.961	RH_6	1.222
-1.021	T_out	-1.344
-1.038	Phase_of_Day_AfterNoon	-0.578
-1.173	Phase_of_Day_Evening	-0.578
-3.578	ytime_of_day	1.381
-7.418	Phase_of_Day_Night	1.724
-8.762	time_of_day	-0.973

ELI5

Conclusion

- When evaluating the influence of Random Variable attribute the linear models have assigned near zero weights to the random variable, negating its influence in prediction of the target variable.
- Random Forest Regressor was found to be the best performing model with an R-squared score of 0.767.
- After optimizing the hyperparameters of the Random Forest Regressor, its R-squared score increased from 0.767 to 0.771.
- We find that this model's predictions are mainly contributed by the time of the day and temperature in rooms 3 and 2.
- As this dataset has a time component to it, we believe that better performances can be achieved by using Time Series Analysis concepts.

References

- https://blog.paperspace.com/implementing-gradient-boosting-regression-python/
- https://www.geeksforgeeks.org/ml-gradient-boosting/
- https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:~:text=Support%20Vector%20Regression%20is%20a,the%20maximum%20number%20of%20points.
- https://www.statology.org/adjusted-r-squared-in-python/
- https://machinelearningmastery.com/implement-machine-learning-algorithm-performance-metrics-scratch-python/#:~:text=RMSE%20is%20calculated%20as%20the,the%20original%20units%20for%20comparison.
- https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/
- https://machinelearningmastery.com/calculate-feature-importance-with-python/
- https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b
- “Machine Learning Mastery With Python, Understand Your Data, Create Accurate Models and Work Projects End-To-End”, Jason Brownlee
- Prediction model of household appliance energy consumption based on machine learning, Lei Xiang et al 2020 J. Phys.: Conf. Ser. 1453 012064

Thank You