

Capstone Project-2 Appliances Energy Prediction

Team Members

Navin Kodam Meet Delvadiya Jyoti Chiluka Muta Ravali



Content:

- 1. Defining Problem Statement
- 2. EDA and Feature Engineering
- 3. Feature Selection
- 4. Preparing Dataset for Modeling
- 5 . Apply Model
- 6. Model Validation and Selection
- 7. Conclusion



The Dilemma:

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. Each wireless node transmitted the temperature and humidity conditions around 3.3 min. Then, the wireless data was averaged for 10 minutes periods. The energy data was logged every 10 minutes with mbus 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 Pipeline:

<u>Data processing-1</u>: In this first part we have removed unnecessary features.

<u>Data processing-2</u>: In this part, we manually go through each features selected from part 1, and encoded with numerical features.

EDA: In this part, we do some exploratory data analysis (EDA) on the features selected in part-1 and 2 to see the trend.

<u>Create a model</u>: Finally, in this last but not last part, we create models. Creating a model is also not an easy task. It is an iterative process. We show how to start a simple model, and slowly add complexity for better performance.



DATA SUMMARY:

lights: Energy use of light fixtures in the house.

T1: Temperature in kitchen area

RH_1: Humidity in kitchen area

T2: Temperature in living room

RH_2: Humidity in living room

T3: Temperature in laundry room area

RH_3: Humidity in laundry room area

T4: Temperature in Office room

RH_4: Humidity in Office room

T5: Temperature in Bathroom

RH_5: Humidity in Bathroom



DATA SUMMARY (Contd...):

T6: Temperature outside the building (northside)

RH_6: Humidity temperature outside the building (northside)

T7: Temperature in ironing room

RH_7: Humidity in ironing room

T8: Temperature in teenager's room

RH_8: Humidity in teenager's room

T9: Temperature in parent's room

RH_9: Humidity in parent's room

T_out: Temperature outside (from Chievres weather station)

Press_mm_hg: Pressure (from Chievres weather station)

RH_out: Humidity outside (from Chievres weather station)



DATA SUMMARY (Contd...):

Windspeed: Windspeed (from Chievres weather station)

Visibility: Visibility (from Chievres weather station)

Tdewpoint: Tdewpoint (from Chievres weather station)

rv1: Random variable 1

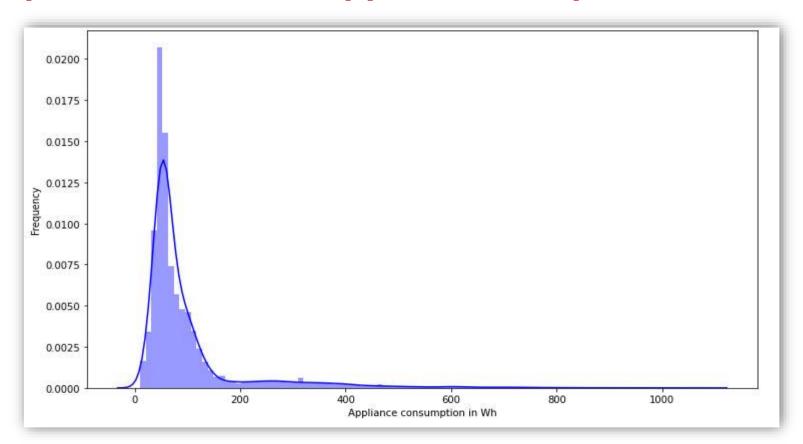
rv2: Random variable 2

Date: Date and time format

Appliances: Energy used by appliances

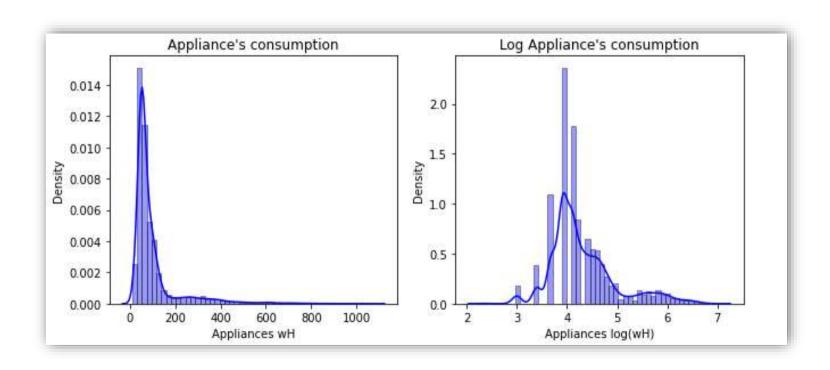


Dependent variable Appliances Graph:



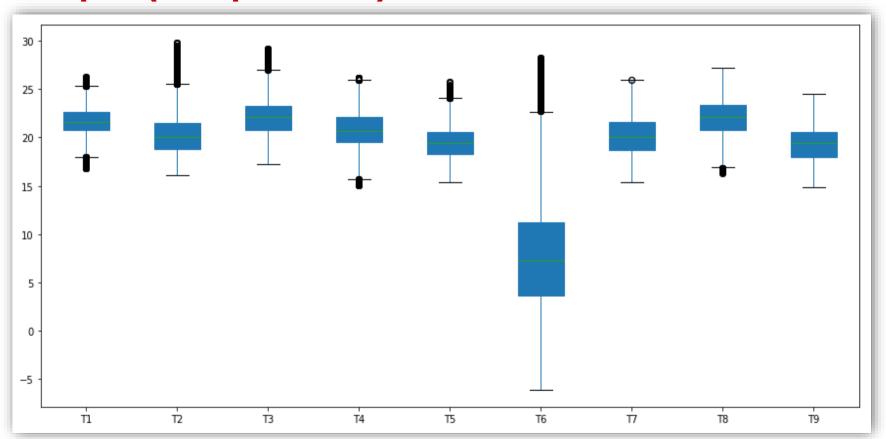


Displot (Appliances wH vs Appliances log(wH)):



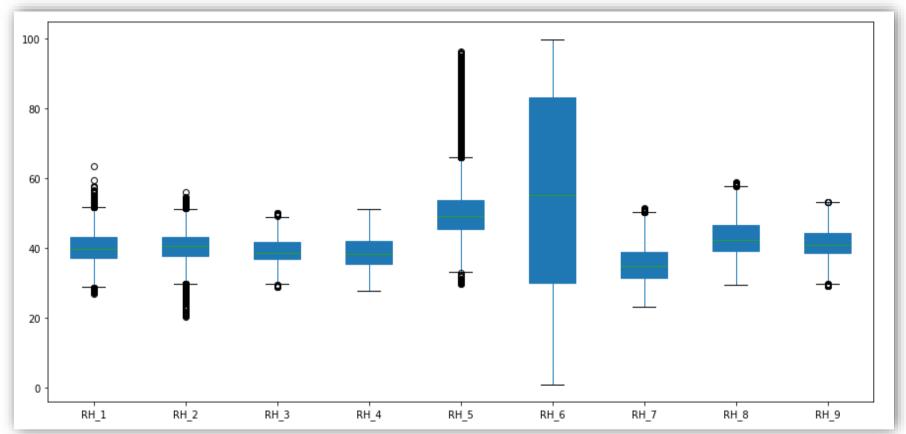


Box plot(Temperature):



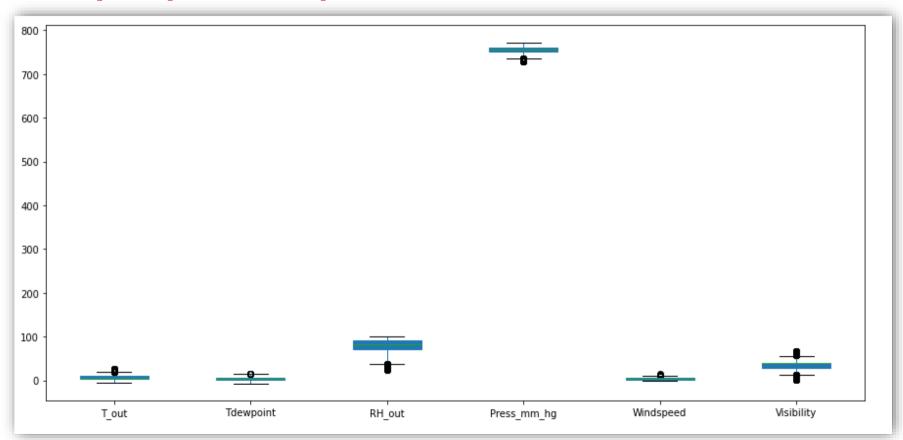


Box plot(Humidity):



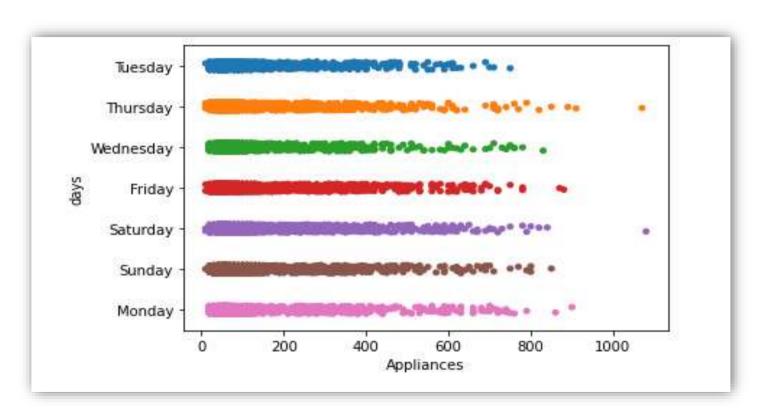


Box plot(Weather):





Dependent variable count w.r.t. Days:



Αl

Correlation:

Appliances	- 1	0.055	0.086	0.12	0.06	0.085	0.036	0.04	0.017	0.02	0.007	0.12	0.083	0.026	0.056	0.04	0.094	0.01	0.051	0.099	0.035	0.15	0.087	0.00023	0.015	
71	0.055	1		0.84	0.0025	0.89	0.029	0.98	0.098	0.89	0.015	0.65	0.62	0.84	0.14	9.83	0.0064	0.84	0.072	10.10			0.088	0.076	0.57	
RH_1	0.006	0.16	1	027	0.9	0.25	0.84	0.11	0.88	0.21	0.3	0.32	0.25	0.021	0.0	0.03	18000	0.12	18770	0.94				0.021	1664	
72	0.12	0.84	0.27	1	9:17	-0.79	0:12	10.00	0.23	0.77	0.03	9.0		0.00	0.23		0.069	ERME I	0.16	0.79			0.052	0.07	(958)	
RH_2	0.06	0.0025	10.00	0.17	1.	014	0=0	8.047	0.77	0.11	0.25	0.0097	0.39	0.051	100	0.043	0.68	0.055	0.66	0.034			0.069	8 0054	0.5	
T3	coes	0.00	8.75	0.74	0.14	1	0.011	0.05	0.12	0.89	0.066	315	3.65	0.85	017	6.9	0.044	0.9	0.13	0.007		028			0.65	
RH 3	0.036	0.029	0.84	0.12	0.68	0.011	1	0.14	0.9	0.05	0.38	0.077	0.53	0.25	0.83	0.28	0.83	0.2	0.83	0.12			0.26	0.017	041	
14	0.04	0.88	011	0.76	0.047	0.85	0.14	1	0.049	0.87	0.076	0.65	9.7	0.80	0.044	0.0	0.095	0.89	0.026	0.66	0.075				0.52	
RH_4		0.098	o mi	0.23	0.72	0.12	0.9	0.049	1	0.092	0.35	0.26	0.59	0.13	0.89	0.17	28.0	0.045	DHE	0.29				0.0026	8.62	
15		0.00	0.21	:07/2	0.11	0.89	0.05	0.07	0.092	1	0.033	Oniz	0.63	0.07	0.15	0.112	0.016	0.91	0.072	tinis-			0.15	0.084	10.59	
-	0.007		0.3	0.03	0.25	0.066	Description of the last	0.076	0.35	0.033	1.	0.078		0.14		0.087		0.14	Name of Street	0.053	017	0.19	0.087	0.013	0.078	
76	10000	0.65		0.0	0.0097	-	0.077	10.05	0.76		8 078	1	0.477	8 67		0.48	0.074	9.67	0.18	097	014	3.57	0.17	0.083	Name and Post of the Owner, where the Owner, which is the Owner, which	
0.000	0.026	0.63	0.023	0.55	0.051	0.65	051	0.58	0.33	0.87	0.26		1	110	0.034	0.88	0.49	094	0.99		0.066	0.41	8.19	011	847	
	0.056	014	0.0	023	0.053	017	0.83	0.044	0.89	0.15	-	0.26	0.36	0.034	7	0.12	0.00	0.028	ORE	0.29	027	0.38	0.21	0 0072	0.64	
13	2000	0.83	0.03	058	0.041	0.0	0.26	9.8	0.17	0.52	0.087	0.46	207	0.88	0.12	0.12	0.21	0.87	016		016	U.S	022	0.06	039	
	0.094	I Committee	-0/4	0.069	2.610	0.044	0.89	0.095	0.95	0.016	0.36	0.074		021	0.88	021	1	0.11	0.00	0.12	0.23	7649	l'eze	0.046	0.5	
19	100000	0.04	0.12	066	0.055	0.9	0.2	0.09	0.045	0.91	014	0.07	10.74	0.94	0.028	0.87	0.11	1	0.0087	0.67	016		0.10	0.1	0.50	
ян э	0.051	0.072	Sn.	916	0.88	913	0.23	0.026	0.06	0.072	9.23	0.18	0.39	0.078	0.86	9.16	0.86	0.0087	1	0.22	0.18	0.36	024	0.0087	054	
	0.099	0.65	034	0.79	0.034	0.7	0.12	0.06	029	0.65	0.053	0.97	0.64	0.63	0.29	05	0.12	10.67	0.22	1	0.14			0.077	0.79	
ress mm hg	0.035	0.15		0.13	026	0.19	023	0.075	0.25	037		014	0.066	0.098	0.27	016		0.16	0.18	014	1	0.092		604	024	
RH_out	0.15			0.53	0.58	0.28		0.39				0.57	8.77	0.43			0.49		0.36	0.57	0.092	1	0.18	0.083	0.037	
Windspeed	0.067	0.068		0.052	0.069	0.1	0.26	0.19		015	0.082	0.17	0.098	0.19		0.22		0.16	0.24	0.19	0.24	0.18	1	0.0075	0.13	
Visibility	0.00023	0.076	0.021	0.07	0.0054	0.1	0.017		0.0026	0.084	0.013	0.061			0.0072	0.06	0.046	01	0.0087	0.077	0.04	0.083	0.0075	1	0.042	
Tdewpoint	0.015	9.57	0.04	0.56	0.5	0.65	0.41	0.52	0.62	0.59	0.078	27E	0.26	0.47	264	0.39	0.5	0.58	0.54	0.79	0.24	0.037	013	0.042	1	
	Applances -	Ė	H. 1	7	FH.2	E	五	产	#	ħ	BH.5	jë.	H.	11.	NH.7	E	菱	19.	6 18	TOK	ss mm bg	RH_out-	Mindspeed	Vaibility .	Tidestipaint	



Preparing dataset for modeling:

Task: Linear Regression

Train test split (75%-25%)

Train Set: (14801, 24)

Test Set: (4934, 24)

Dependent Variable : Appliances

Appliances	T1	RH_1	T2	RH_2	13	RH_3	T4	RH_4	15	RH_5	T6	RH_6
60	19,89	0.475967	19.2	0.447900	19.79	0.447300	19.000000	0.455667	17.166667	0.5520	7.026667	0.842567
60	19.89	0.466933	19.2	0.447225	19.79	0.447900	19.000000	0.459925	17.166667	0.5520	6.833333	0.840633
50	19.89	0.463000	19.2	0.446267	19.79	0.449333	18.926667	0.458900	17.166667	0.5509	6.560000	0.831567
50	19.89	0.460667	19.2	0.445900	19.79	0.450000	18.890000	0.457233	17.166667	0.5509	6.433333	0.834233
60	19.89	0.463333	19.2	0.445300	19.79	0.450000	18.890000	0.455300	17.200000	0.5509	6.366667	0.848933



We had reduced multicollinearity
 By removing irrelevant and less
 correlation features and considering

these new features

But these new features are not giving good score

 So we are considering the old features for Further implementation

1	considering
VIF	variables
33.886304	RH_5
49.249180	Т6
24.229162	RH_6
4.624325	Windspeed
19.462279	Tdewpoint
100.224776	RH_4_T7
55.452478	temp_
68.332990	RH_3_RH_out_RH_1
197.566361	RH_6_RH_7



	variables	VIF
0	T1	3804.104348
1	RH_1	1639.616095
2	T2	2490.017329
3	RH_2	2184.338515
4	Т3	1239.155390
5	RH_3	1587.782332
6	T4	932.716301
7	RH_4	1357.715241
8	T5	1187.885478
9	RH_5	45.091108
10	Т6	88.925465
11	RH_6	40.315447
12	T7	1613.381841
13	RH_7	518.846594
14	Т8	975.014239
15	RH_8	568.351963
16	Т9	2516.975132
17	RH_9	637.316129
18	T_out	399.738956
19	Press_mm_hg	2084.856382
20	RH_out	1297.930593
21	Windspeed	5.248122
22	Visibility	12.029393
23	Tdewpoint	132.494808
24	rv1	inf
25	rv2	inf



Applying Model (Polynomial Features):

Fitting on Polynomial features of degrees (1,2,3) Showing the R2 score and Root mean square error of Train and Test

For degree 1: Train_RMSE= 93.26197768632537

Train_R2_Score = 0.15188149648747196

For degree 2: Train_RMSE= 84.31139430405246
Train_R2_Score = 0.306861790816

Train_R2_Score = 0.3068617908169937

For degree 3: Train_RMSE= 55.55639468980285 Train R2 Score = 0.69903540691406 Test_RMSE= 98.69343235583071 Test R2 Score = 0.13598579386328224

Test_RMSE= 86.86620652884946 Test_R2_Score = 0.3306610635036924

Test_RMSE= 37.64486125940055 Test_R2_Score = 0.8742940661330362



Model Validation & Selection:

	Train_R2_Score	Test_R2_Score	Test_RMSE_Score	Train_RMSE_Score
Name				
polynomial(degree = 3)	0.699035	0.874294	37.644860	55.556390
RandomForest	0.939304	0.518390	73.684394	24.949095
KNeighborsRegressor:	0.688354	0.440688	79.406315	56.533630
Gradient Boosting Regressor:	0.340765	0.233555	92.954056	82.223627
XGBRegressor:	0.327814	0.224481	93.502689	83.027337
Lasso	0.151375	0.129445	99.066258	93.289770
Ridge:	0.151029	0.128090	99.143389	93.308828
SVR:	-0.003329	-0.017707	107.112267	101.437397



Model Validation & Selection (contd...):

Observation 1: Support vector regression (svr) is giving worst r score for these dataset

Observation 2: As seen in the above slides... Random forest is giving high train score but having less test r score, polynomial(degree = 3) is giving High test r score but having less train r score

Observation 3: From the above observation we have come to the conclusion that we would choose our model from Random forest or polynomial features(degree = 3)



Model Validation & Selection (contd...):

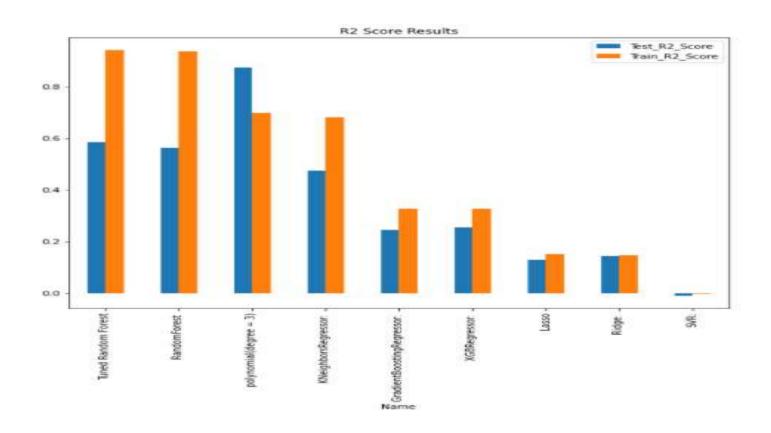
grid search.best params

- Tuning hyperparameters of Random Forest regressor and we got the best parameters and best estimators
- But the Rmse value of Random Forest is 24.26% and Rmse for Polynomial(degree = 3) is 37.64%
- So we concluded that
 Random forest is giving
 the best score than
 Polynomial for these dataset

```
{'max depth': 60, 'max features': 'sqrt', 'n estimators': 250}
grid search.best estimator
RandomForestRegressor(bootstrap=True, ccp alpha=0.0, criterion='mse'.
                      max depth=60, max features='sqrt', max leaf nodes=None,
                      max samples=None, min impurity decrease=0.0,
                      min impurity split=None, min samples leaf=1,
                      min samples split=2, min weight fraction leaf=0.0,
                      n estimators=250, n jobs=None, oob score=False,
                      random state=40, verbose=0, warm start=False)
grid search.best estimator .score(X train,Y train)
0.9434541975986162
grid search.best estimator .score(X test,Y test)
0.5864961299637497
np.sqrt(mean_squared_error(Y_test, grid_search.best_estimator .predict(X test)))
65.64558516057404
np.sqrt(mean squared error(Y train, grid search.best estimator .predict(X train)))
24.265303711829215
```



Comparisons of all models:





Conclusion:

- We are getting the good results when we selecting 22 features for the model implementation and dropping lights,rv1, rv2 and Visibility.
- > The best algorithm for this dataset is Random Forest Regressor as compared to the rest of the algorithms.
- After tuning the algorithm using Grid Search CV on Random Forest the score is not getting much difference compared to Polynomial Regression, because the correlation between the dependent and independent variables are very low in dataset.
- Feature reduction was not able to give much better accuracy.



Challenges:

- Mostly, features have very low correlation so feature selection was challengeable.
- > Most of algorithms doesn't give good score even after feature engineering.
- > Computation time in Polynomial Regression.



Thank you!