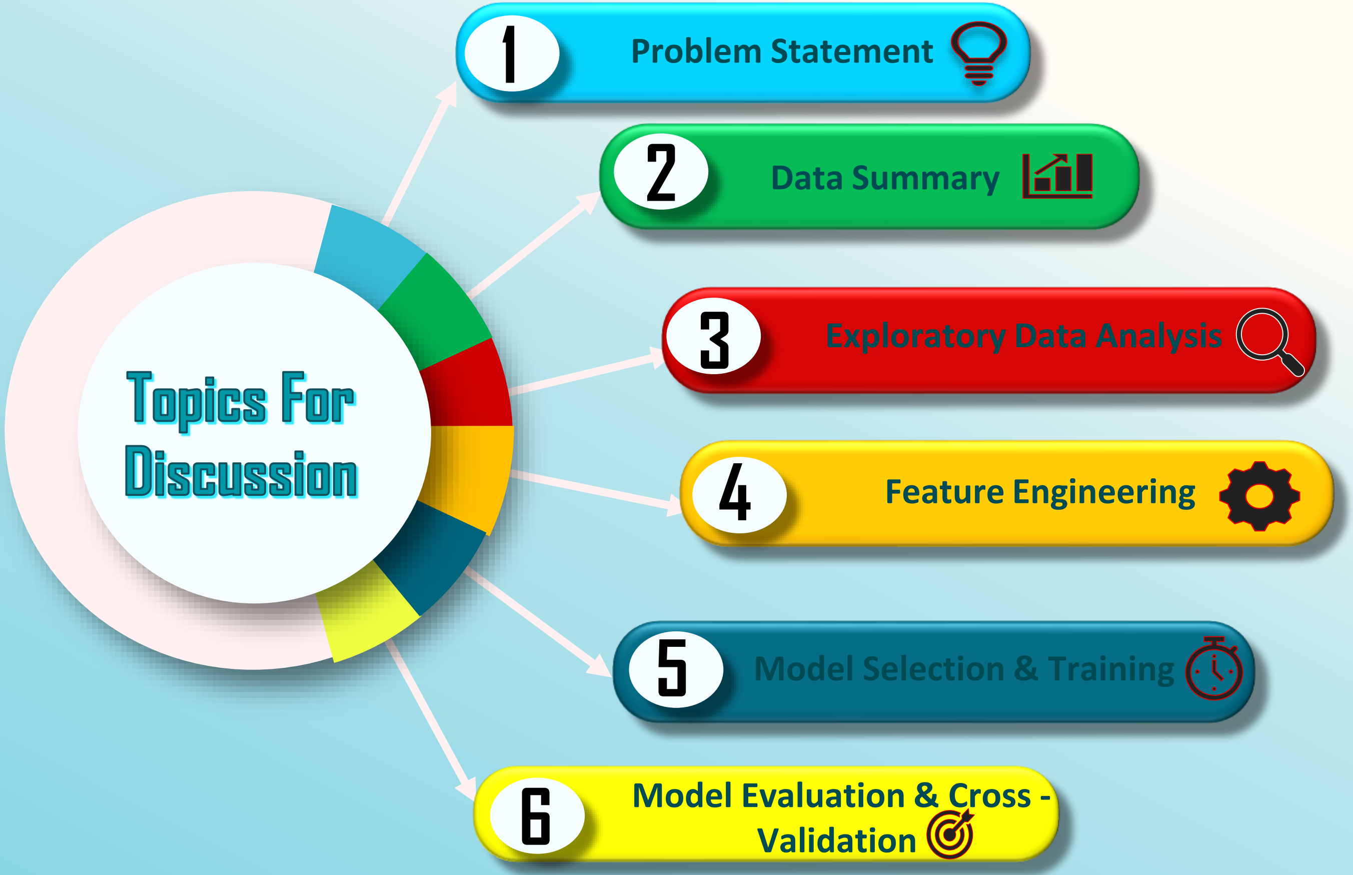
Capstone Project

Appliances Energy Prediction

By

Umesh, Ifraz, Shailendra, Akram

Project Outline



Topics For Discussion

1 Problem Statement

2 Data Summary

3 Exploratory Data Analysis

4 Feature Engineering

5 Model Selection & Training

6 Model Evaluation & Cross - Validation

Problem Statement

Objective of Project

- ❑ The goal is to predict energy consumption by appliances.
- ❑ Develop the perceiving of various parameters affecting energy consumption.
- ❑ In the age of smart homes, the ability to predict energy consumption will not only save the money of end-user but also can contribute to energy conservation motive.

Data Summary

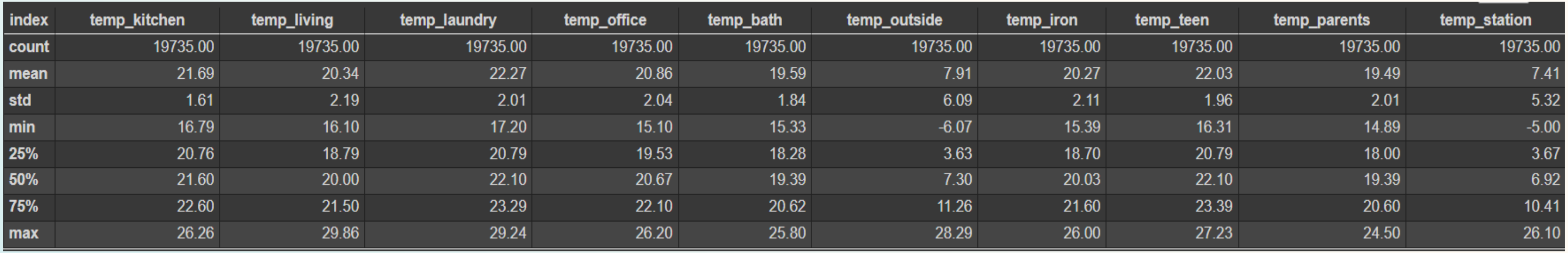
01 The dataset we have is a series of sensor data collected from a building in Belgium at an interval of 10 mins for a period of about 4.5 months.

02 The sensor data consists of temperature and humidity levels in different rooms in the building.

03 Among other features are weather reports on Pressure, Wind speed, Visibility, and T-dew point, which are recorded at weather station Chievres Airport, Belgium.

04 The target variable is the total energy consumption of the building in Wh. The dataset has no null values.

Descriptive Statistical Analysis (Temperatures)



index	temp_kitchen	temp_living	temp_laundry	temp_office	temp_bath	temp_outside	temp_iron	temp_teen	temp_parents	temp_station
count	19735.00	19735.00	19735.00	19735.00	19735.00	19735.00	19735.00	19735.00	19735.00	19735.00
mean	21.69	20.34	22.27	20.86	19.59	7.91	20.27	22.03	19.49	7.41
std	1.61	2.19	2.01	2.04	1.84	6.09	2.11	1.96	2.01	5.32
min	16.79	16.10	17.20	15.10	15.33	-6.07	15.39	16.31	14.89	-5.00
25%	20.76	18.79	20.79	19.53	18.28	3.63	18.70	20.79	18.00	3.67
50%	21.60	20.00	22.10	20.67	19.39	7.30	20.03	22.10	19.39	6.92
75%	22.60	21.50	23.29	22.10	20.62	11.26	21.60	23.39	20.60	10.41
max	26.26	29.86	29.24	26.20	25.80	28.29	26.00	27.23	24.50	26.10

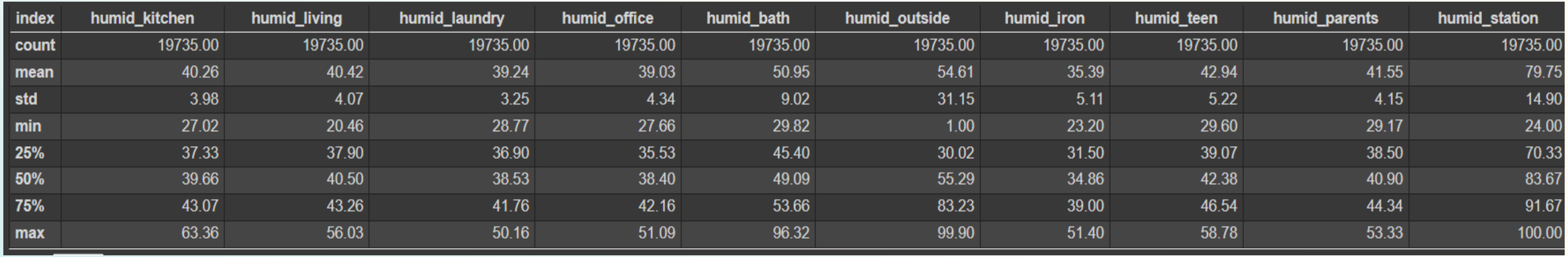
Fig 1: Temperature DataFrame

01 The average outside temperature over a period of 4.5 months is around 7.5 degrees. It ranges from -6 - 28 degrees.

02 While the average temperature inside the building has been around 20 degrees for all the rooms.

03 This implies Warming appliances have been used to keep the insides of the building warm.

Descriptive Statistical Analysis(Humidity)



index	humid_kitchen	humid_living	humid_laundry	humid_office	humid_bath	humid_outside	humid_iron	humid_teen	humid_parents	humid_station
count	19735.00	19735.00	19735.00	19735.00	19735.00	19735.00	19735.00	19735.00	19735.00	19735.00
mean	40.26	40.42	39.24	39.03	50.95	54.61	35.39	42.94	41.55	79.75
std	3.98	4.07	3.25	4.34	9.02	31.15	5.11	5.22	4.15	14.90
min	27.02	20.46	28.77	27.66	29.82	1.00	23.20	29.60	29.17	24.00
25%	37.33	37.90	36.90	35.53	45.40	30.02	31.50	39.07	38.50	70.33
50%	39.66	40.50	38.53	38.40	49.09	55.29	34.86	42.38	40.90	83.67
75%	43.07	43.26	41.76	42.16	53.66	83.23	39.00	46.54	44.34	91.67
max	63.36	56.03	50.16	51.09	96.32	99.90	51.40	58.78	53.33	100.00

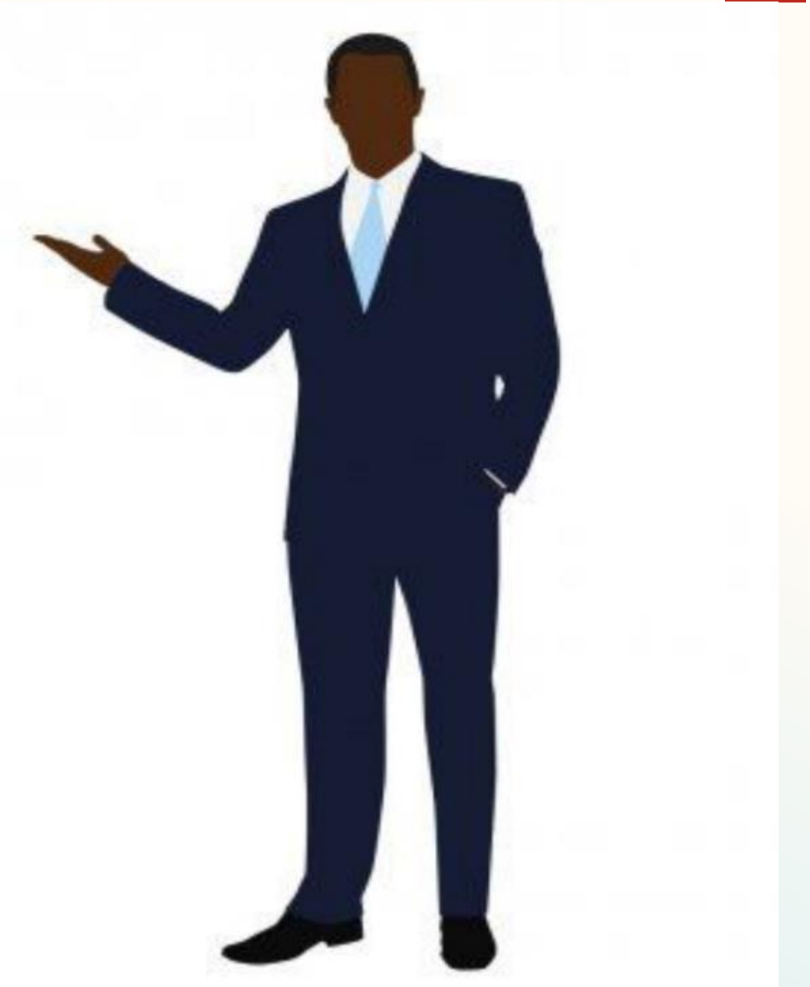
Fig 2: Humidity DataFrame

01 Average humidity at the weather station is significantly higher compared to outside humidity near the building

02 Kids and parent rooms show a comparatively higher average humidity as well signifying the fact that inhabitants of this building spend most of their time in these buildings.

03 Average humidity in the bathroom is significantly higher compared to other rooms due to obvious reasons.

Exploratory Data Analysis



Target Variable

01 Energy consumption of appliances ranges from 10 Wh to 1080 Wh.

02 About 75 % of energy values lie below 100 Wh, and about 93 % of them lie below 300 Wh.

03 Our target variable seems to be highly skewed, and our task is to predict the usual as well as the large surges in energy in the building.

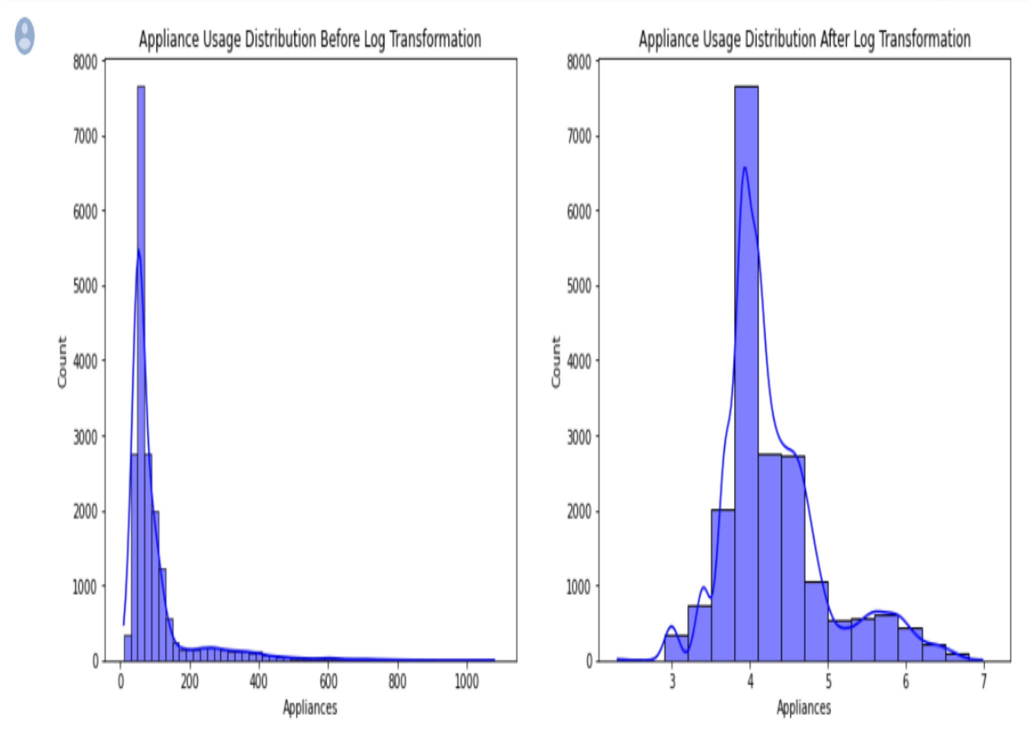


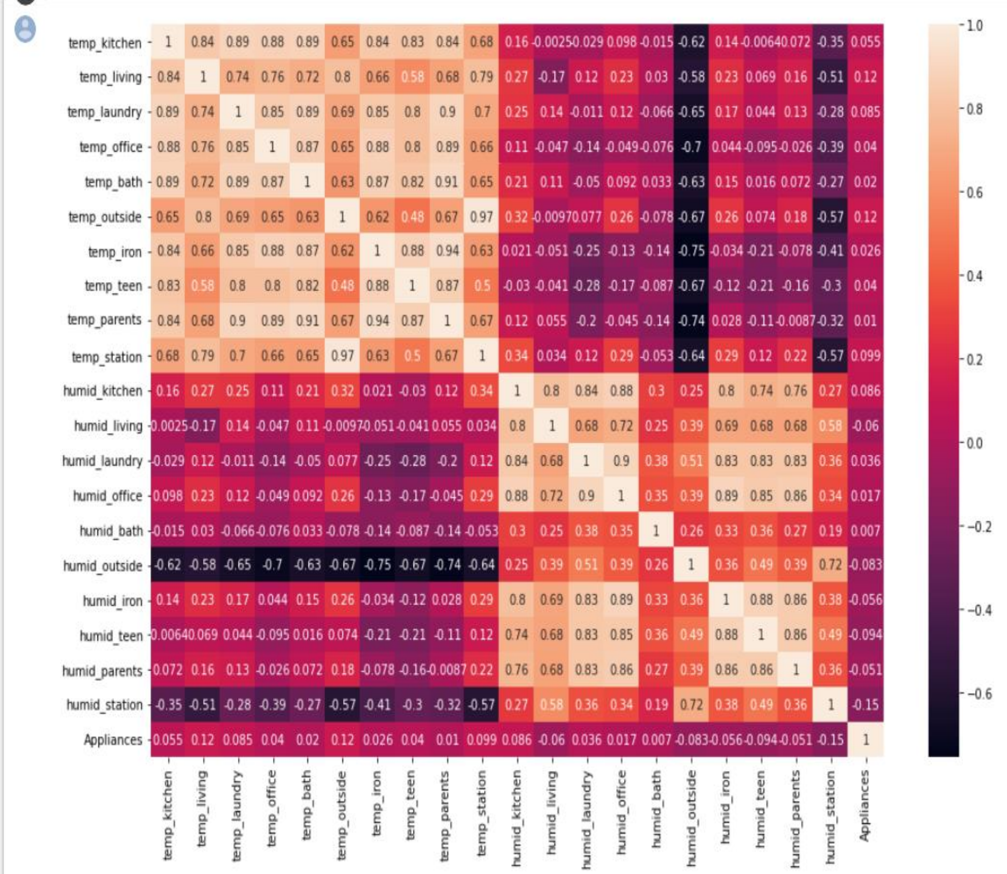
Fig 3: Target Variable Distribution

Pearson’s Correlation Matrix

01 The Temperature and humidity levels in each of the rooms are highly correlated among themselves.

02 There seems to be no relationship between humidity and temperature levels in the building. However, temperature and humidity levels outside the building are strongly negatively correlated.

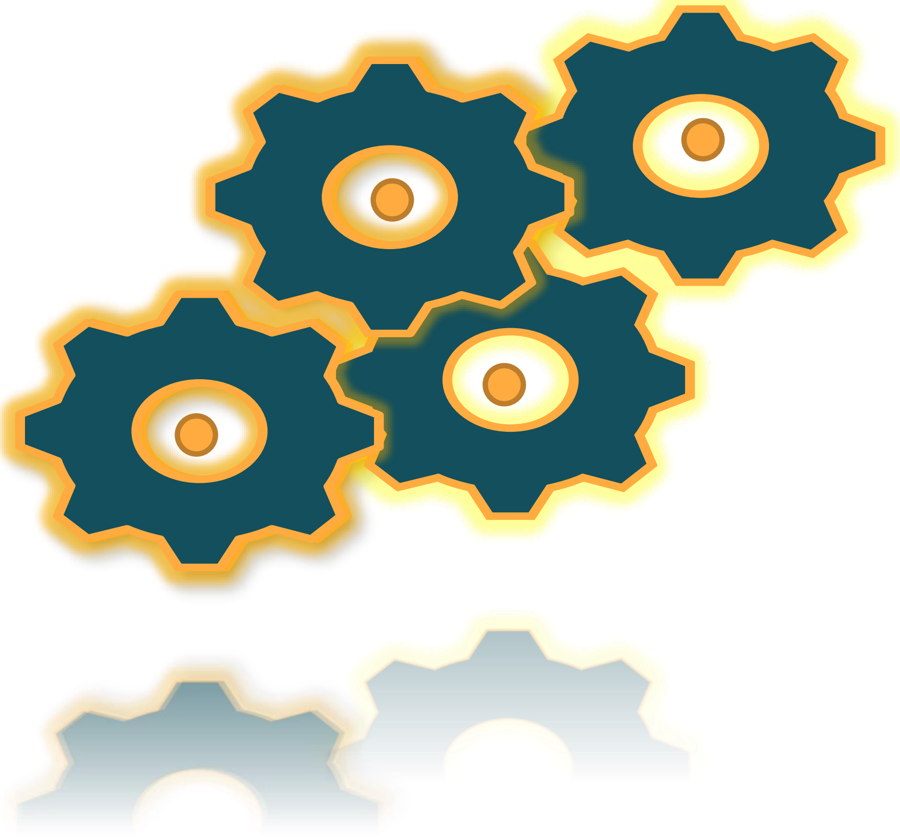
03 There is little to no correlation between these features and the target variable i.e., Appliance energy consumption variable.



	temp_kitchen	temp_living	temp_laundry	temp_office	temp_bath	temp_outside	temp_iron	temp_teen	temp_parents	temp_station	humid_kitchen	humid_living	humid_laundry	humid_office	humid_bath	humid_outside	humid_iron	humid_teen	humid_parents	humid_station	Appliances
temp_kitchen	1	0.84	0.89	0.88	0.89	0.65	0.84	0.83	0.84	0.68	0.16	-0.0025	0.029	0.098	-0.015	-0.62	0.14	-0.0064	0.072	-0.35	0.055
temp_living	0.84	1	0.74	0.76	0.72	0.8	0.66	0.58	0.68	0.79	0.27	-0.17	0.12	0.23	0.03	-0.58	0.23	0.069	0.16	-0.51	0.12
temp_laundry	0.89	0.74	1	0.85	0.89	0.69	0.85	0.8	0.9	0.7	0.25	0.14	-0.011	0.12	-0.066	-0.65	0.17	0.044	0.13	-0.28	0.085
temp_office	0.88	0.76	0.85	1	0.87	0.65	0.88	0.8	0.89	0.66	0.11	-0.047	-0.14	-0.049	-0.076	-0.7	0.044	-0.095	-0.026	-0.39	0.04
temp_bath	0.89	0.72	0.89	0.87	1	0.63	0.87	0.82	0.91	0.65	0.21	0.11	-0.05	0.092	0.033	-0.63	0.15	0.016	0.072	-0.27	0.02
temp_outside	0.65	0.8	0.69	0.65	0.63	1	0.62	0.48	0.67	0.97	0.32	-0.0097	0.077	0.26	-0.078	-0.67	0.26	0.074	0.18	-0.57	0.12
temp_iron	0.84	0.66	0.85	0.88	0.87	0.62	1	0.88	0.94	0.63	0.021	-0.051	-0.25	-0.13	-0.14	-0.75	-0.034	-0.21	-0.078	-0.41	0.026
temp_teen	0.83	0.58	0.8	0.8	0.82	0.48	0.88	1	0.87	0.5	-0.03	-0.041	-0.28	-0.17	-0.087	-0.67	-0.12	-0.21	-0.16	-0.3	0.04
temp_parents	0.84	0.68	0.9	0.89	0.91	0.67	0.94	0.87	1	0.67	0.12	0.055	-0.2	-0.045	-0.14	-0.74	0.028	-0.11	-0.0087	-0.32	0.01
temp_station	0.68	0.79	0.7	0.66	0.65	0.97	0.63	0.5	0.67	1	0.34	0.034	0.12	0.29	-0.053	-0.64	0.29	0.12	0.22	-0.57	0.099
humid_kitchen	0.16	0.27	0.25	0.11	0.21	0.32	0.021	-0.03	0.12	0.34	1	0.8	0.84	0.88	0.3	0.25	0.8	0.74	0.76	0.27	0.086
humid_living	-0.0025	-0.17	0.14	-0.047	0.11	-0.0097	-0.051	-0.041	0.055	0.034	0.8	1	0.68	0.72	0.25	0.39	0.69	0.68	0.68	0.58	-0.06
humid_laundry	-0.029	0.12	-0.011	-0.14	-0.05	0.077	-0.25	-0.28	-0.2	0.12	0.84	0.68	1	0.9	0.38	0.51	0.83	0.83	0.83	0.36	0.036
humid_office	0.098	0.23	0.12	-0.049	0.092	0.26	-0.13	-0.17	-0.045	0.29	0.88	0.72	0.9	1	0.35	0.39	0.89	0.85	0.86	0.34	0.017
humid_bath	-0.015	0.03	-0.066	-0.076	0.033	-0.078	-0.14	-0.087	-0.14	-0.053	0.3	0.25	0.38	0.35	1	0.26	0.33	0.36	0.27	0.19	0.007
humid_outside	-0.62	-0.58	-0.65	-0.7	-0.63	-0.67	-0.75	-0.67	-0.74	-0.64	0.25	0.39	0.51	0.39	0.26	1	0.36	0.49	0.39	0.72	-0.083
humid_iron	0.14	0.23	0.17	0.044	0.15	0.26	-0.034	-0.12	0.028	0.29	0.8	0.69	0.83	0.89	0.33	0.36	1	0.88	0.86	0.38	-0.056
humid_teen	-0.0064	0.069	0.044	-0.095	0.016	0.074	-0.21	-0.21	-0.11	0.12	0.74	0.68	0.83	0.85	0.36	0.49	0.88	1	0.86	0.49	-0.094
humid_parents	0.072	0.16	0.13	-0.026	0.072	0.18	-0.078	-0.16	-0.0087	0.22	0.76	0.68	0.83	0.86	0.27	0.39	0.86	0.86	1	0.36	-0.051
humid_station	-0.35	-0.51	-0.28	-0.39	-0.27	-0.57	-0.41	-0.3	-0.32	-0.57	0.27	0.58	0.36	0.34	0.19	0.72	0.38	0.49	0.36	1	-0.15
Appliances	0.055	0.12	0.085	0.04	0.02	0.12	0.026	0.04	0.01	0.099	0.086	-0.06	0.036	0.017	0.007	-0.083	-0.056	-0.094	-0.051	-0.15	1

Fig :Heatmap Correlation Plot

Feature Engineering



Principal component Analysis

01 Given, temperature levels in different rooms had a very low correlation with the target variable, and high correlation among themselves, we reduce the feature set into lower dimensions that could explain maximum variance.

02 PCA 1 and 2 explain more than 91 % variance in the temperature levels in different rooms in the building.

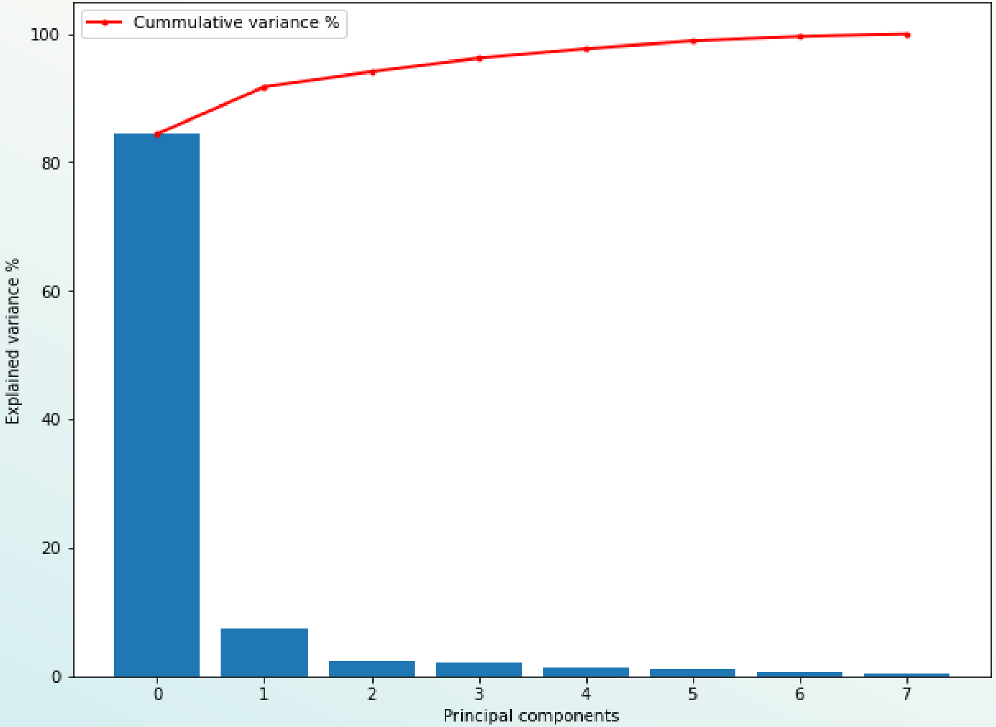


Fig : PCA on Temperature Features

Principal Component Analysis

01 Given, humidity levels in different rooms had a very low correlation with the target variable, and high correlation among themselves, we reduce the feature set into lower dimensions that could explain maximum variance.

02 PCA 1 and 2 explain more than 91 % variance in the humidity levels in different rooms in the building.

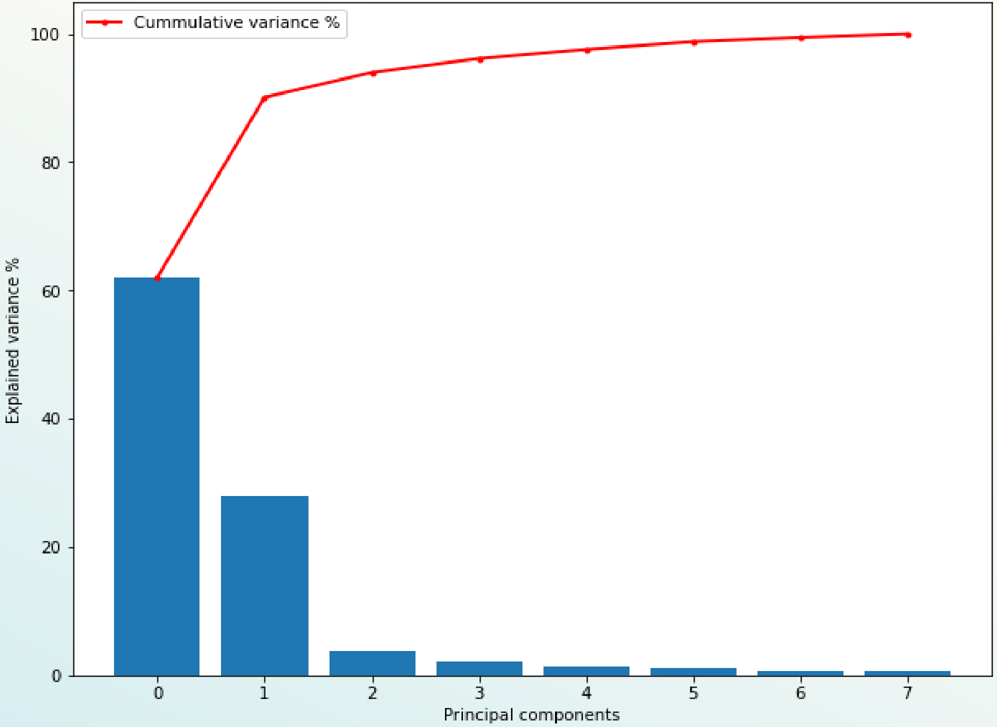


Fig: PCA Humidity Features

Date time (Hour)

01 We have a column called date which includes the timestamp corresponding to each of the sensor data samples.

02 There seems to be a pattern to the energy consumption of appliances at different times of the day.

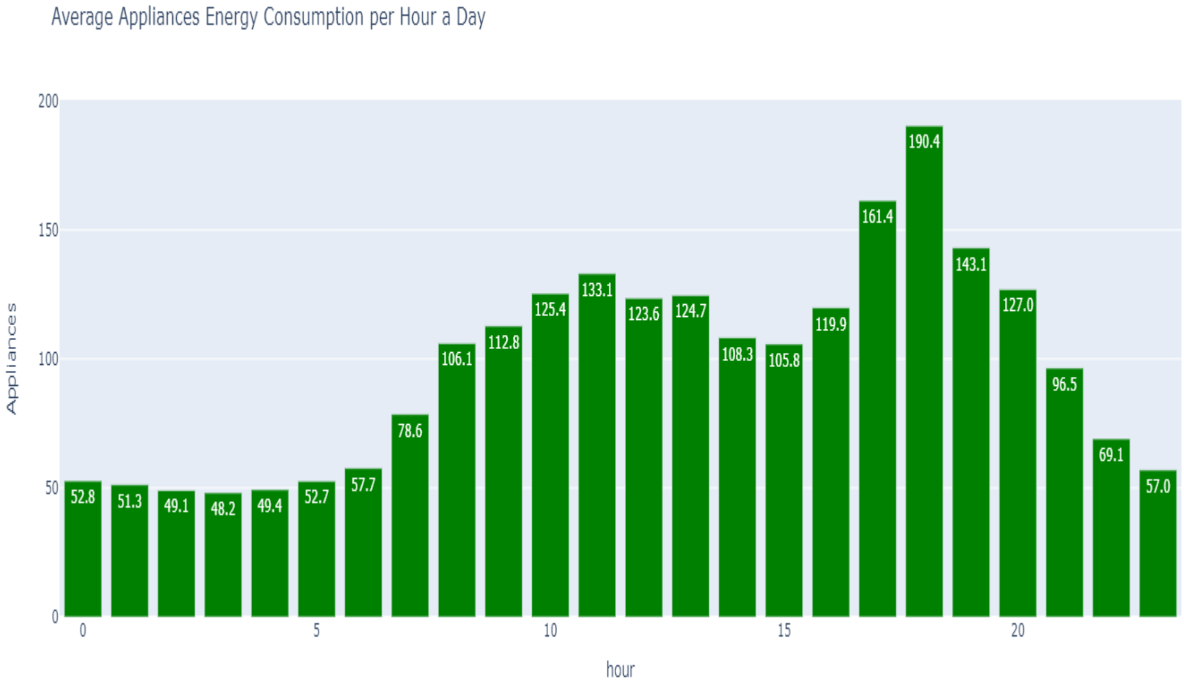


Fig: Average Energy Consumption Per Hour of a day

Date time (Month)

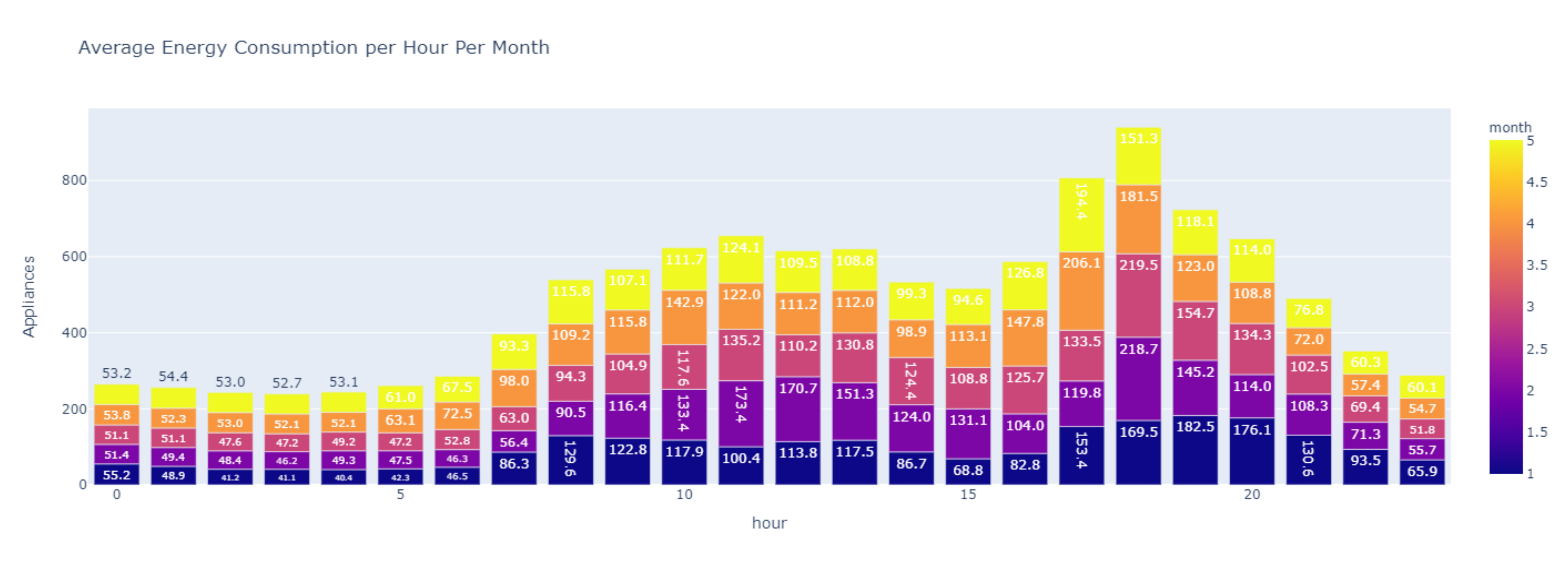


Fig: Average Energy Consumption Per Hour Per Month (for 4.5 Months)

Date time (Weekday & Weekend)

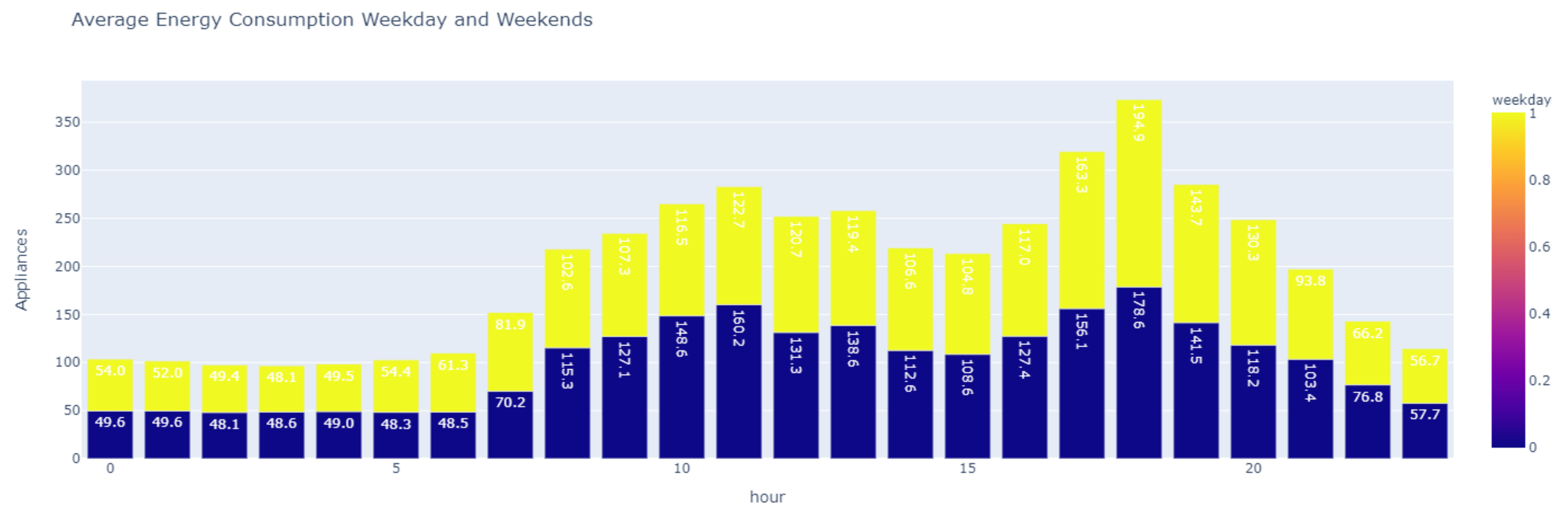


Fig: Weekday and Weekend Appliances Energy Consumption Comparison

Correlation (Newly Created Features)

01 There is no significant correlation between month and the appliance energy consumption, as we saw the pattern of consumption was almost similar overall months. The same is the case for weekdays and weekends.

02 However, there is a significant correlation between hours and the target variable.

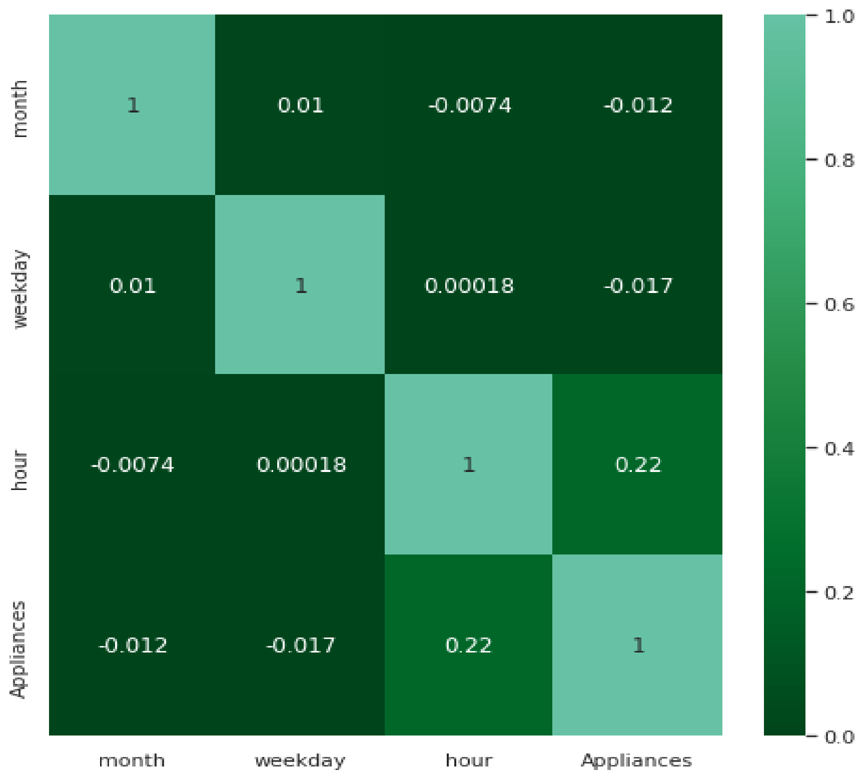


Fig: Heatmap Correlation Plot

Session

We divided the dataset into 3 sessions :

Class 1: 10 PM - 6 AM

Class 2: 6 AM - 3 PM

Class 3: 3 PM - 10 PM

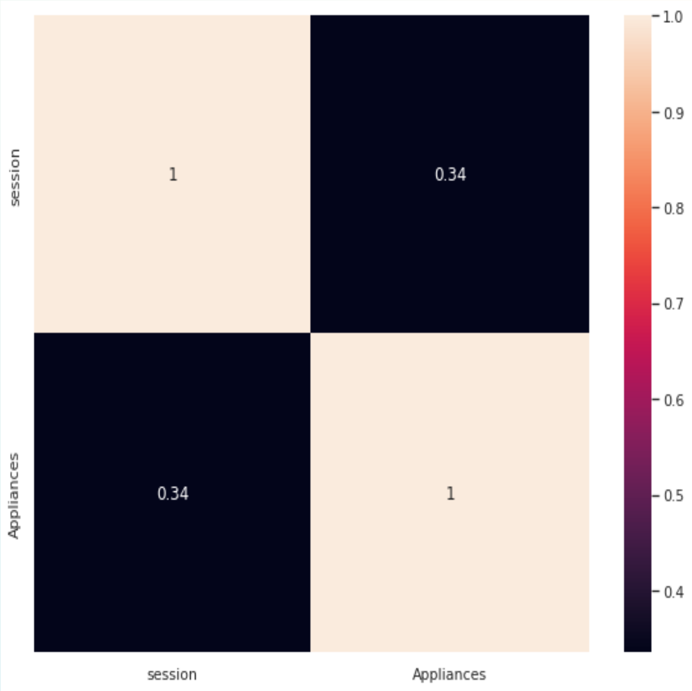


Fig: Heatmap Correlation Plot

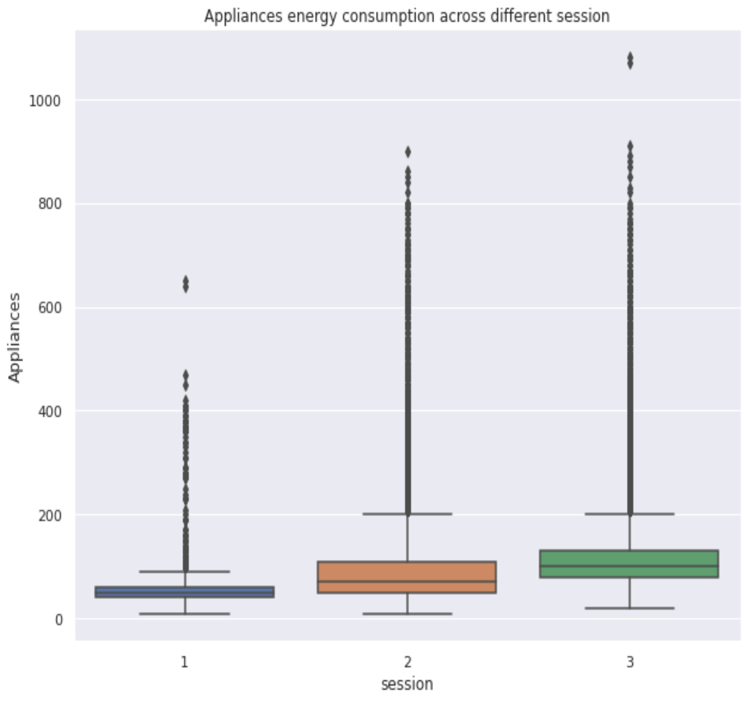


Fig: Appliances Usages for each Session

Final Features

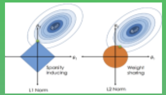
Feature Set I (With PCA)	Feature Set II (Without PCA)
Session	Session
Temperature - pca 1	Temperature of all rooms inside building
Temperature - pca 2	Humidity of all rooms inside building
Humidity - pca 1	Temperature outside
Humidity - pca 2	Humidity outside
Temperature outside	Pressure
Humidity outside	Wind speed
Wind speed	

Table: Final Feature for Modeling

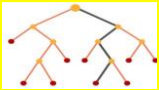
Model Training and Evaluation

Model Trained

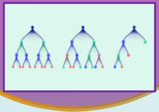
Lasso and Ridge regression



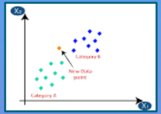
Random Forest & Extra trees Regressor



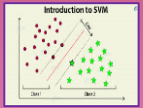
GBR & XG-Boost Regressor



K Neighbours Regressor



Support Vector Regressor



Performance Check (With PCA Features)

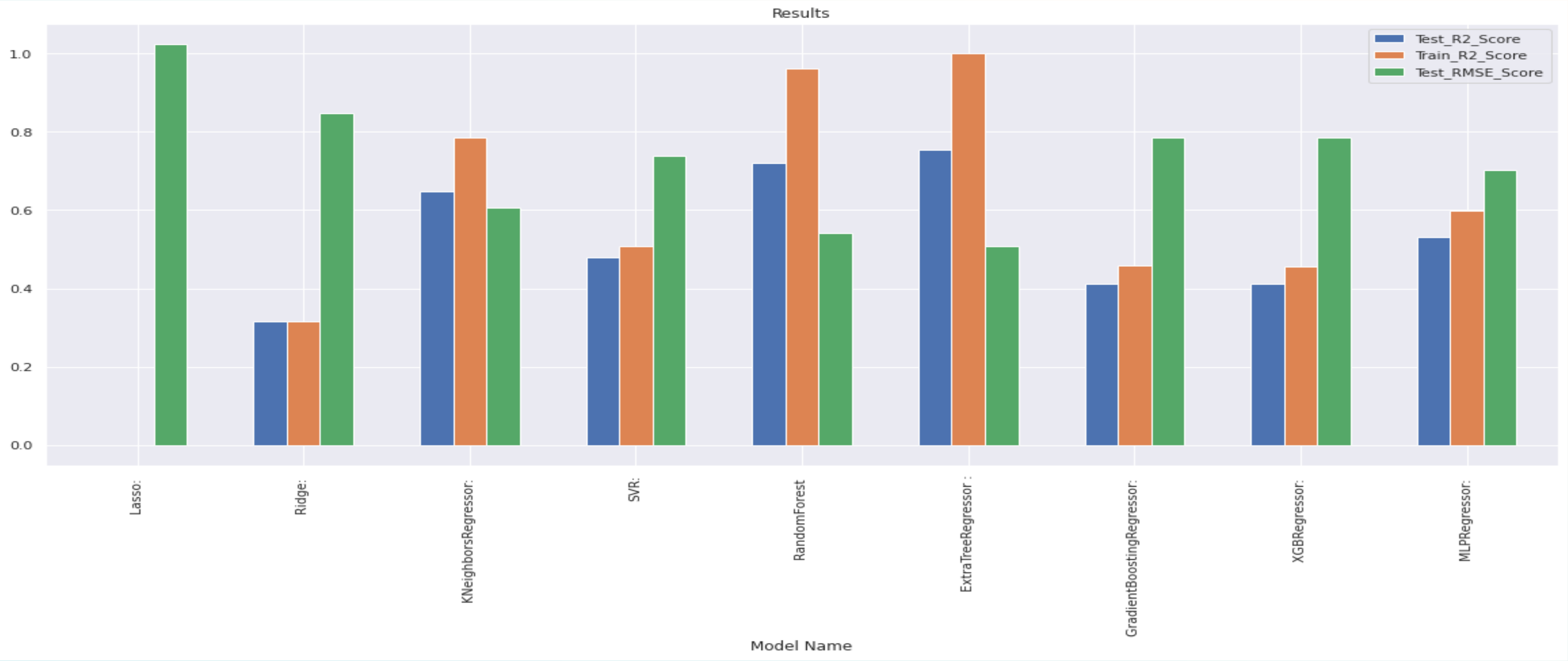


Fig: Performance Check

Performance Check (Without PCA Features)

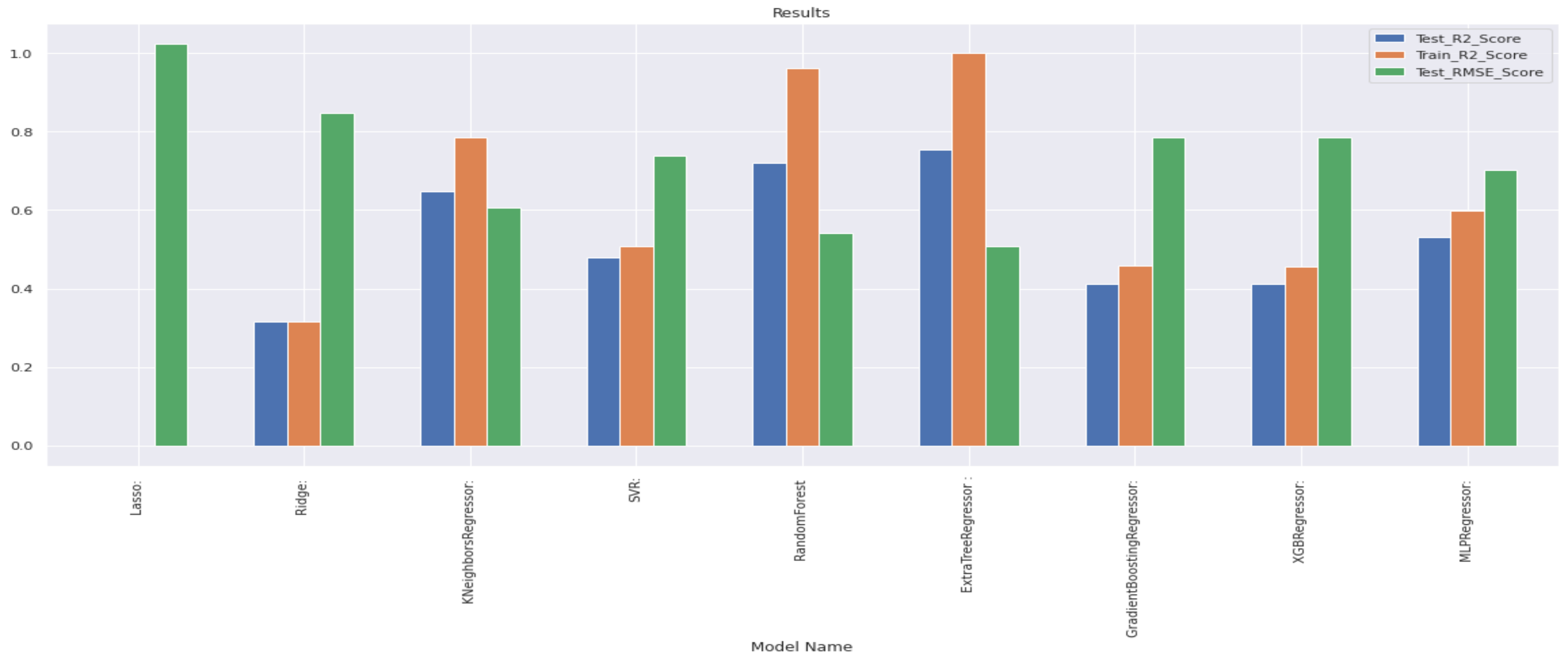


Fig: Performance Check

Best Performing Model

01 Both features set with or without PCA features perform equally well on our selected tree-based model.

02 Although Extra trees regressor overfits our training data with an R2 score of 1, it also gives, by far, the best performance on the test set as well compared to other models, with an R2 score of 0.7584.

03 The test RMSE is also comparatively too low compared to other models.

04 Hyperparameter tuning of our model doesn’t have any significant impact on our test results. We were able to improve the test R2 score results by less than 1%.

Analysis of our model

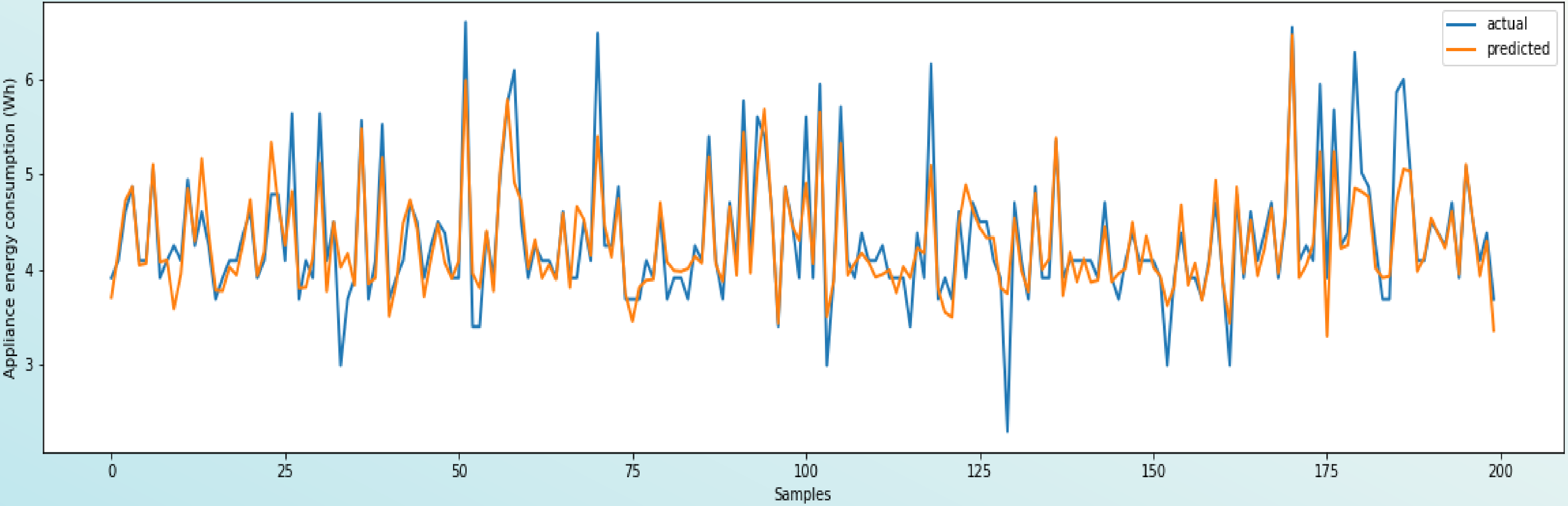


Fig: Model Analysis

Analysis of our model

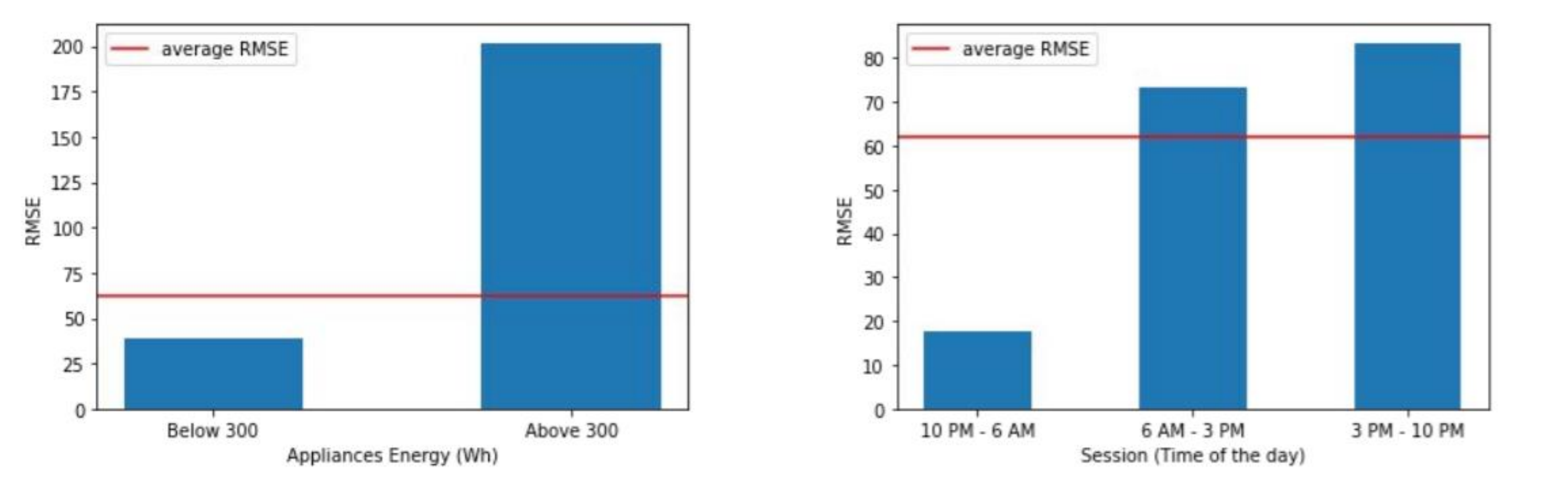
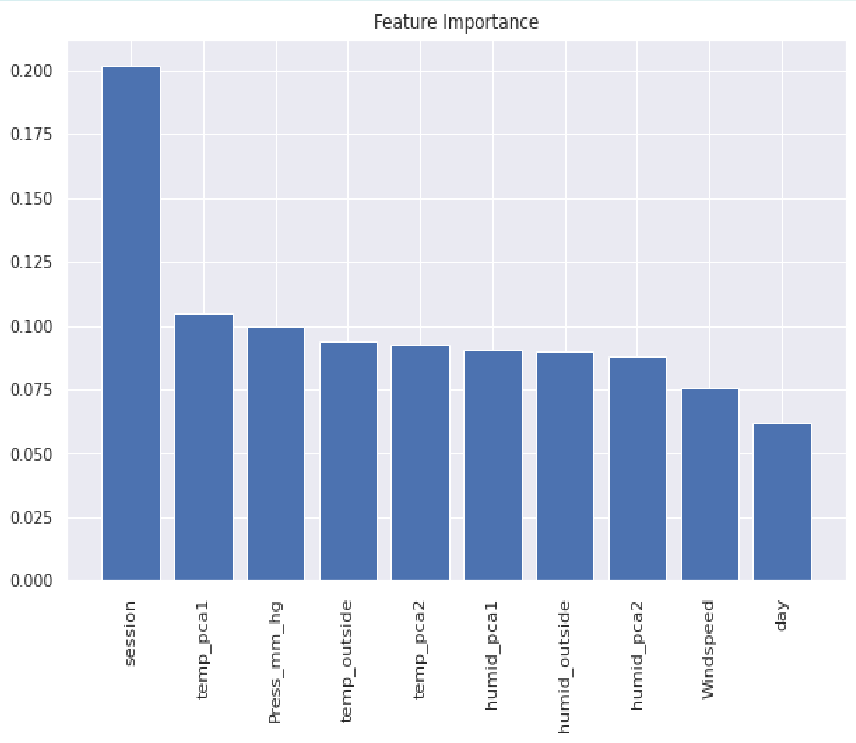
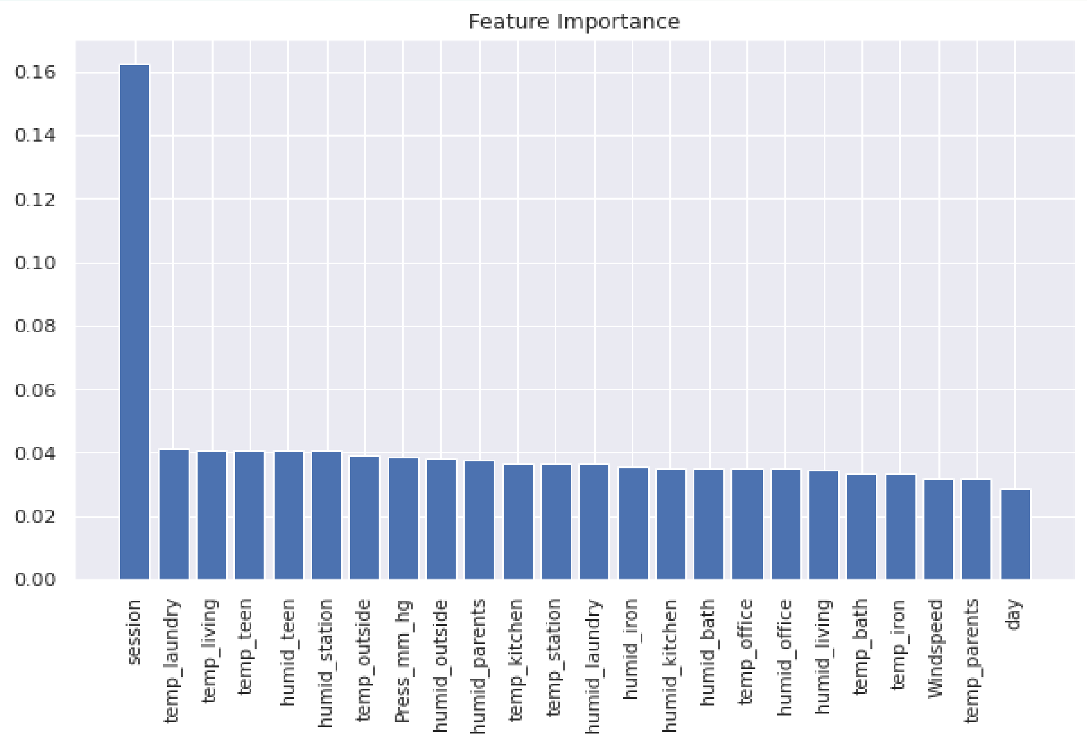


Fig: Model Analysis

Feature Importance



Feature Importance with PCA



Feature Importance without PCA

Conclusion

01 The humidity and temperature feature has little to no linear correlation with the target variable.

02 The time zone of the day plays an important role in deciding the power consumption of appliances.

03 The best Algorithm to use for this dataset is Extra Trees Regressor (tree-based algorithm).

04 PCA helped us to reduce our feature set dimension considerably without affecting the performance of our models significantly.

05 The untuned model was able to explain 75.50% of the variance (R2 score = 0.7550) on the test set, while the tuned model was able to explain 75.84% of the variance (R2 score = 0.7584).

06 The least RMSE score on the test data set is found to be around 0.5 by the Extra trees regressor model, which is considerably good compared to other models.

07 Tree-based models are by far the best model while dealing with a data set that has most of its features having no linear correlation with the target variable. For similar reasons, linear models such as linear regression, Ridge, and Lasso perform the worst.

THANK YOU!