Capstone Project - 2

Appliances Energy Prediction

Kartika Sharma

How much wh?

- Problem Statement
- EDA and feature engineering
- Feature Selection
- Preparing dataset
  for modelling
- Model Fitting
- Evaluation
- Hyperparameter tuning

Problem statement

- **This Dataset is speculative models driven by power consumption data. The data used includes measurements of temperature and humidity from the wireless network, weather from the nearest airport and the recorded power consumption of the lighting equipment. Filtering data to remove unpredictable parameters and feature editing is a common function of this database. From a wireless network, data from the kitchen, laundry and living room is considered very important in power forecasting. Predictability models with weather data only, have selected atmospheric pressure (corresponding to wind speed) as the flexibility of the most accurate weather data in the** forecast. Therefore, atmospheric pressure may be necessary to incorporate it into power forecasting models and to create a performance model.

Data Summary

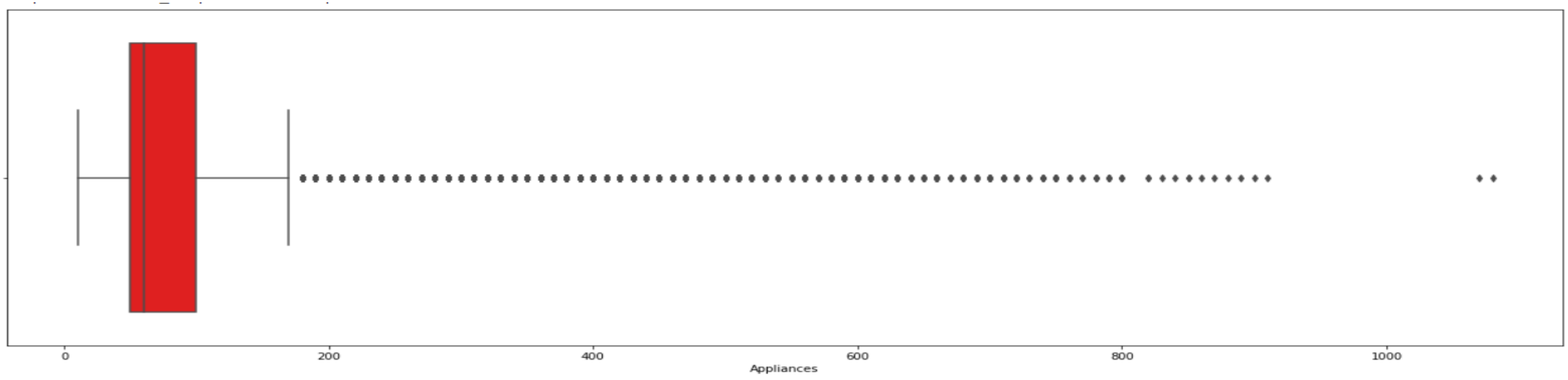
- Date - time year-month- day of energy consumption
- Appliances - energy use in WH (Dependent variable)
- Lights - energy use of light fixtures in the house in Wh (Drop this column)
- T1 - Temperature in kitchen area, in Celsius
- RH1 - Humidity in kitchen area, in %
- T2 - Temperature in living room area, in Celsius
- RH2 - Humidity in living room area, in %
- T3 - Temperature in laundry room area
- RH3 - Humidity in laundry room area, in %
- T4 - Temperature in office room, in Celsius
- RH4 - Humidity in office room, in %

Data Summary

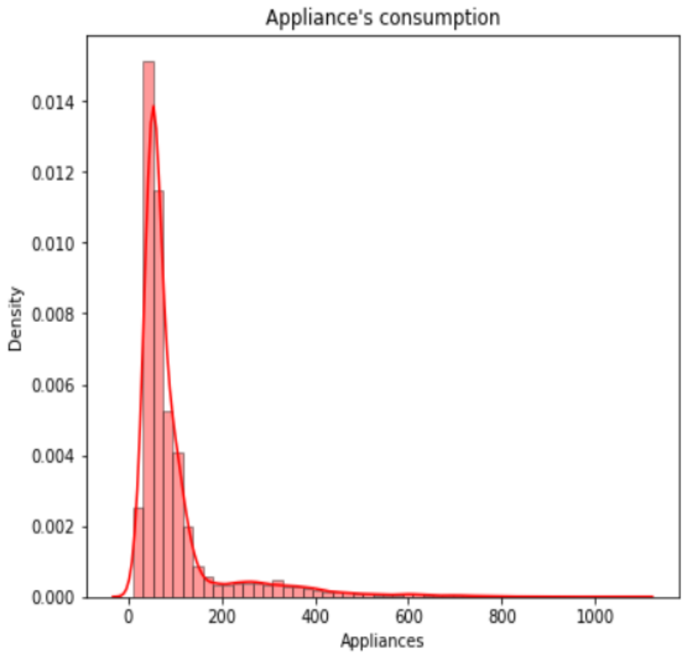
- T9 - Temperature in parent’s room, in Celsius
- RH9 - Humidity in parent’s room, in %
- T_out - Temperature outside (from Chievres weather station), in Celsius
  Press_mm_hg -Pressure (from Chievres weather station), in mm Hg
- Rhout - Humidity outside (from Chievres weather station), in %
- Wind speed - (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

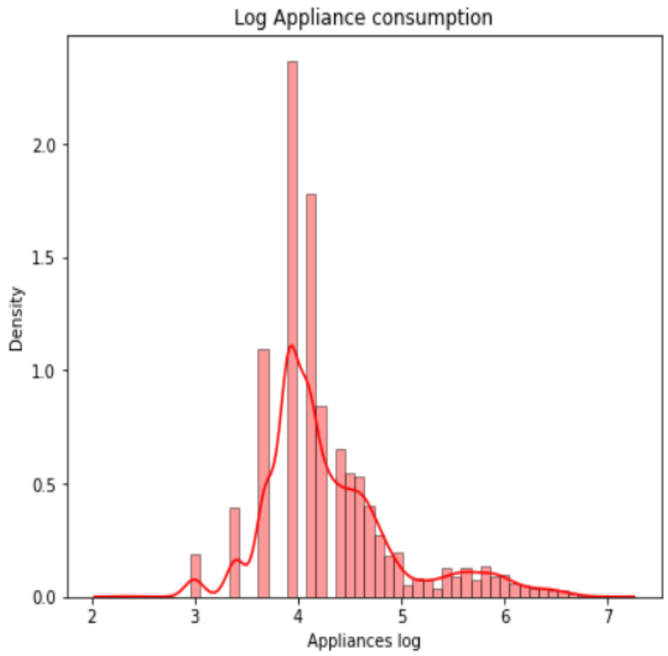
Target Variable

- 75% of Appliance consumption is less than 100 Wh
- With the maximum consumption of 1080 WH, there will be outliers in this column and there are small number of cases where consumption is very high
- This column is positively skewed, most the values are around mean 100 Wh



Normalising Outliers…





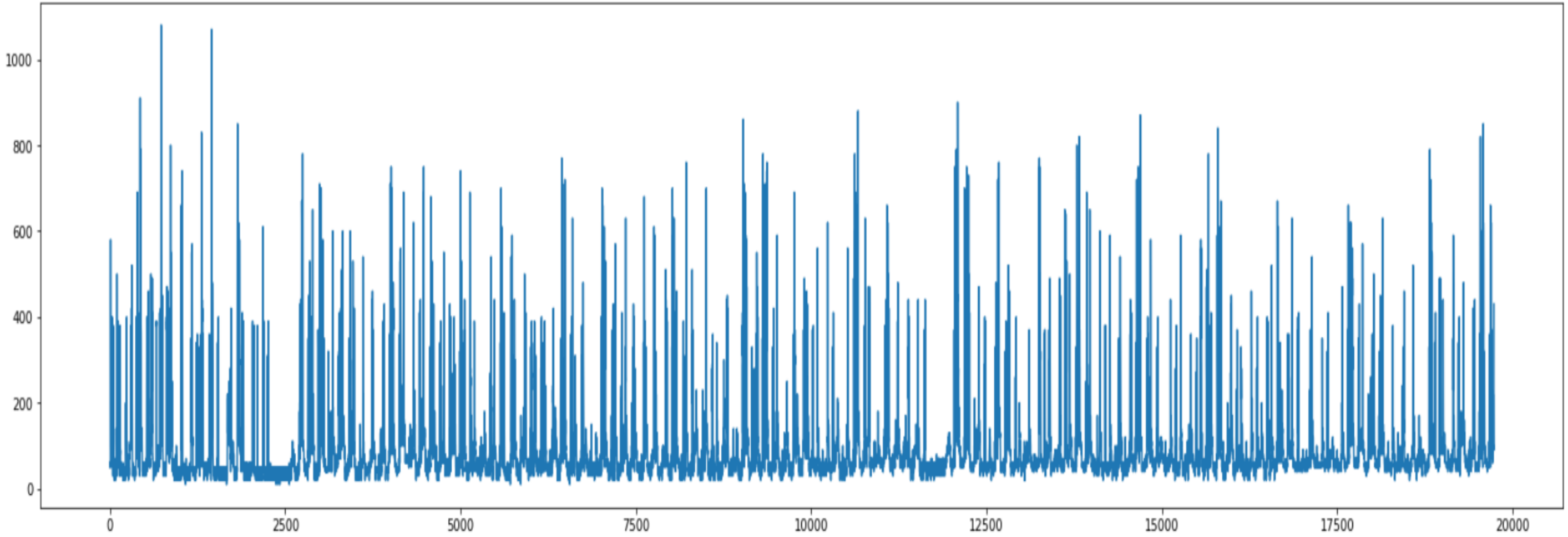
Feature extraction…

5. EXTRACTING NEW FEATURES FROM DATE COLUMN

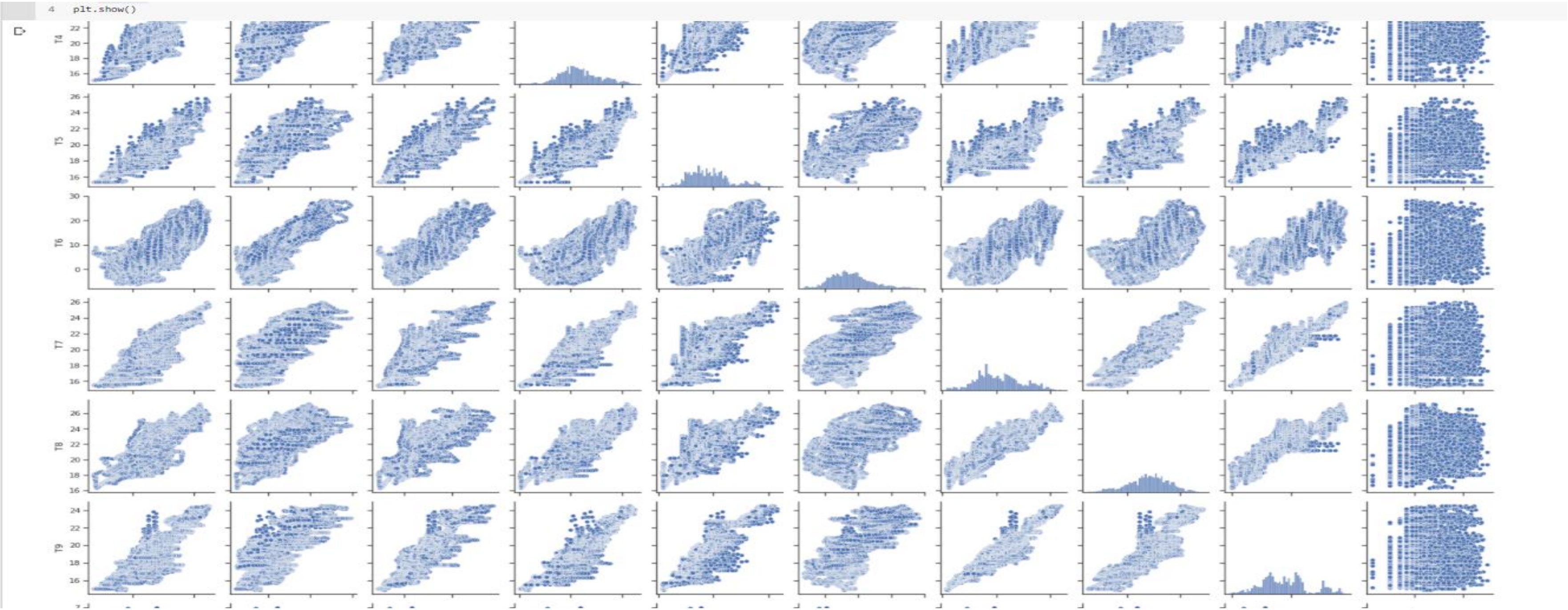
```
energy_df['week_day'] = ((pd.to_datetime(energy_df['date']).dt.dayofweek)// 5 == 1).astype(float)
energy_df['date']=pd.to_datetime(energy_df["date"])

energy_df['hours']=energy_df['date'].dt.hour
energy_df['month']=energy_df['date'].dt.month
energy_df['day']=energy_df['date'].dt.day
energy_df['week_of_month']=(energy_df['date'].dt.day//7)+1
```

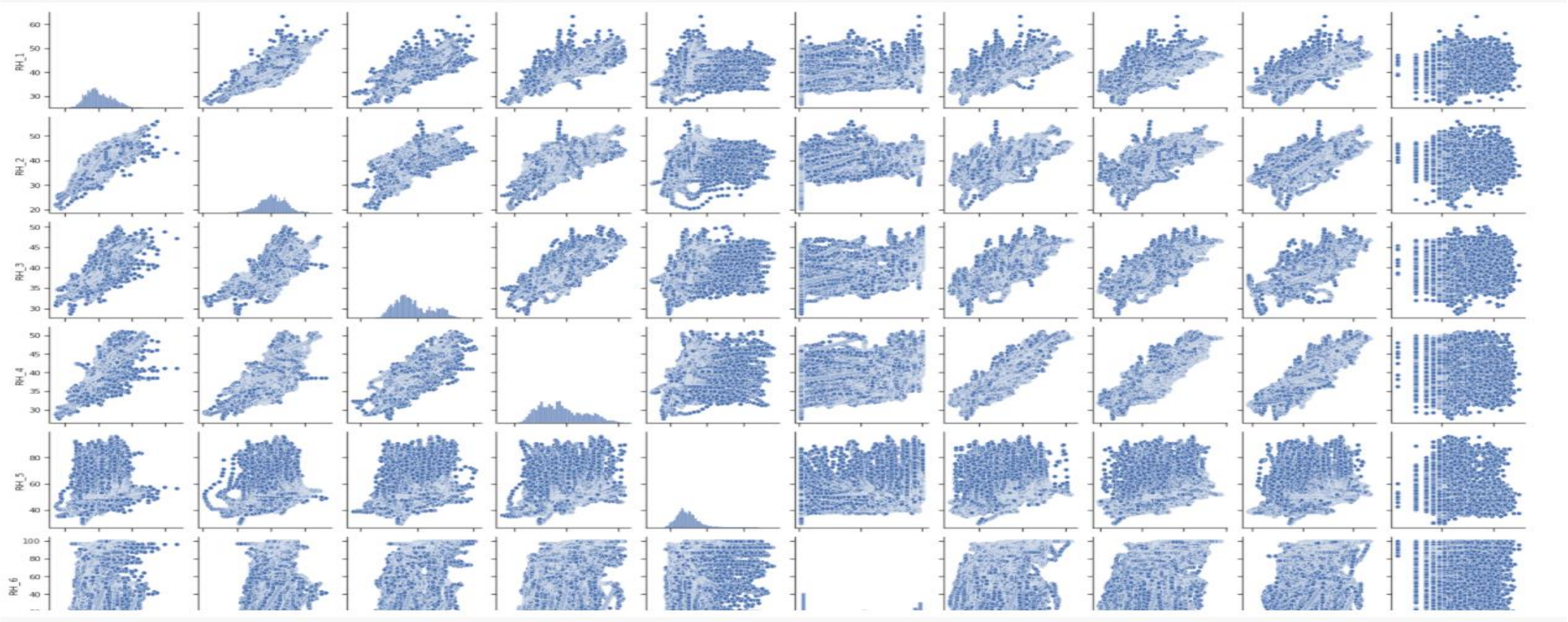
EDA (Seasonal…..I guess not)



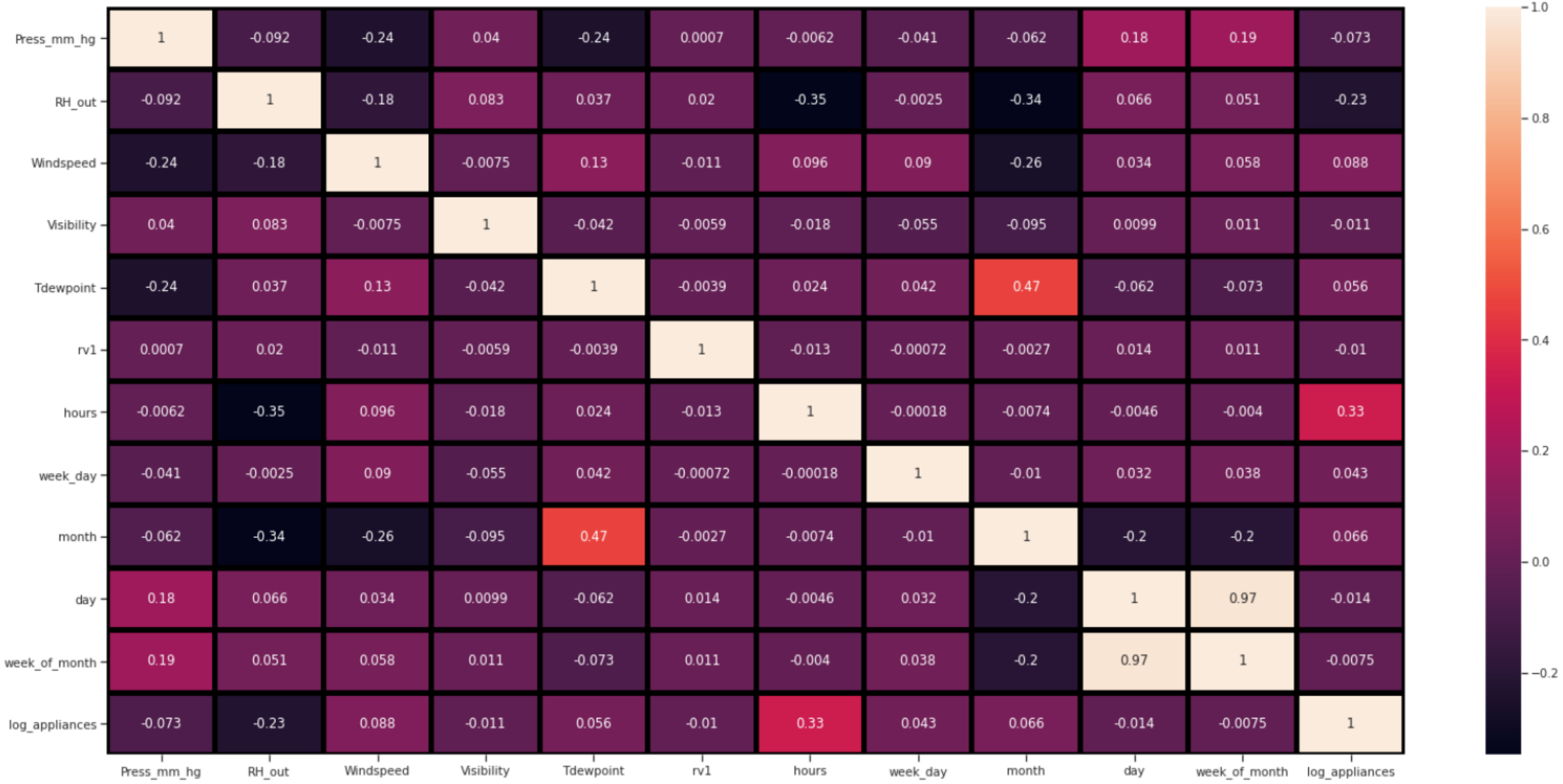
EDA (temperatures)



EDA (Humidities)



EDA

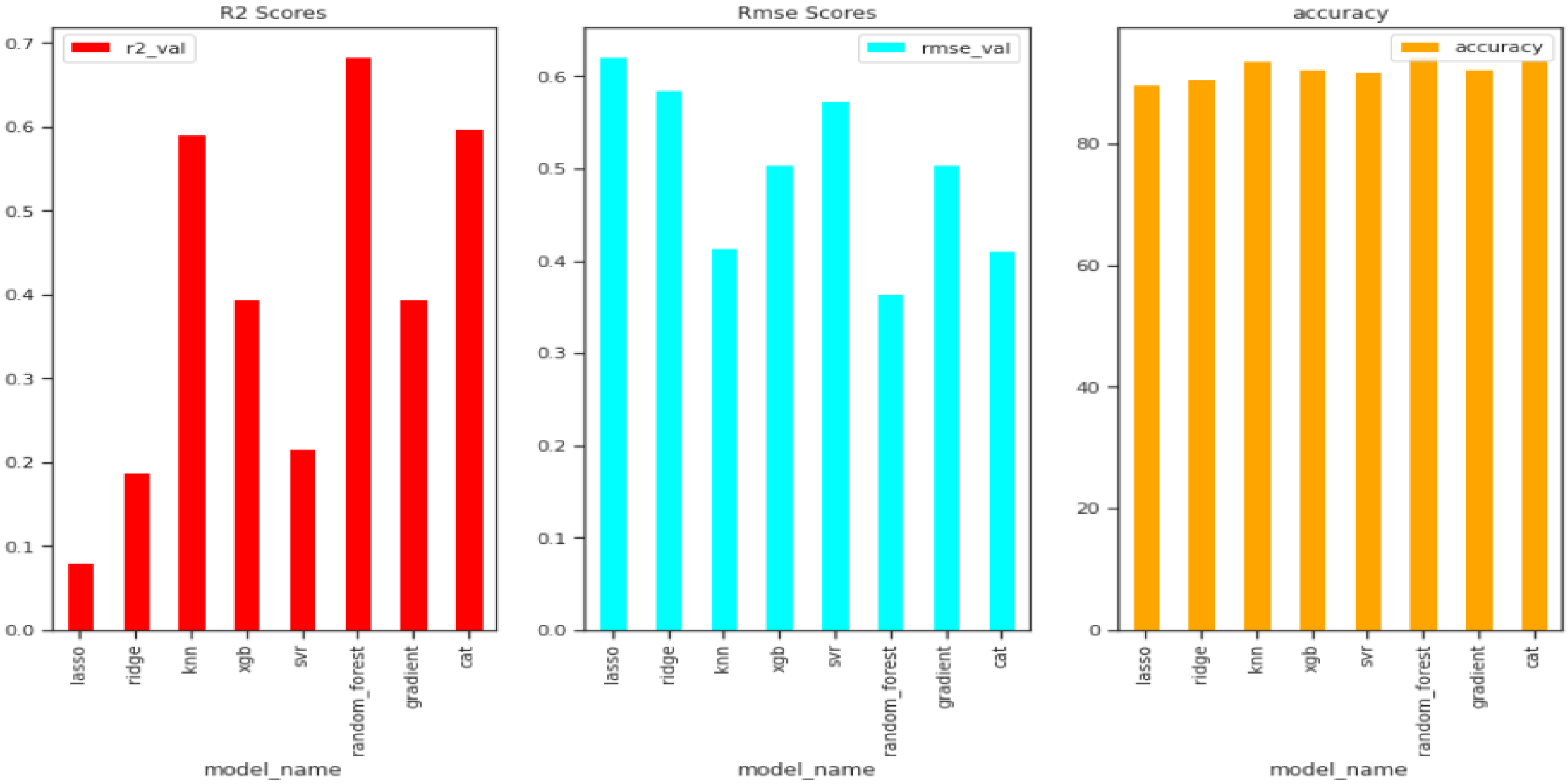

Press_mm_hg
1 -0.092 -0.24 0.04 -0.24 0.0007 -0.0062 -0.041 -0.062 0.18 0.19 -0.073
RH_out
-0.092 1 -0.18 0.083 0.037 0.02 -0.35 -0.0025 -0.34 0.066 0.051 -0.23
Windspeed
-0.24 -0.18 1 -0.0075 0.13 -0.011 0.096 0.09 -0.26 0.034 0.058 0.088
Visibility
0.04 0.083 -0.0075 1 -0.042 -0.0059 -0.018 -0.055 -0.095 0.0099 0.011 -0.011
Tdewpoint
-0.24 0.037 0.13 -0.042 1 -0.0039 0.024 0.042 0.47 -0.062 -0.073 0.056
rv1
0.0007 0.02 -0.011 -0.0059 -0.0039 1 -0.013 -0.00072 -0.0027 0.014 0.011 -0.01
hours
-0.0062 -0.35 0.096 -0.018 0.024 -0.013 1 -0.00018 -0.0074 -0.0046 -0.004 0.33
week_day
-0.041 -0.0025 0.09 -0.055 0.042 -0.00072 -0.00018 1 -0.01 0.032 0.038 0.043
month
-0.062 -0.34 -0.26 -0.095 0.47 -0.0027 -0.0074 -0.01 1 -0.2 -0.2 0.066
day
0.18 0.066 0.034 0.0099 -0.062 0.014 -0.0046 0.032 -0.2 1 0.97 -0.014
week_of_month
0.19 0.051 0.058 0.011 -0.073 0.011 -0.004 0.038 -0.2 0.97 1 -0.0075
log_appliances
-0.073 -0.23 0.088 -0.011 0.056 -0.01 0.33 0.043 0.066 -0.014 -0.0075 1
Press_mm_hg RH_out Windspeed Visibility Tdewpoint rv1 hours week_day month day week_of_month log_appliances
1.0 0.8 0.6 0.4 0.2 0.0 −0.2

Models

- Lasso
- Ridge
- Knn
- SVM
- Random Forest
- Gradient Boosting
- Catboost

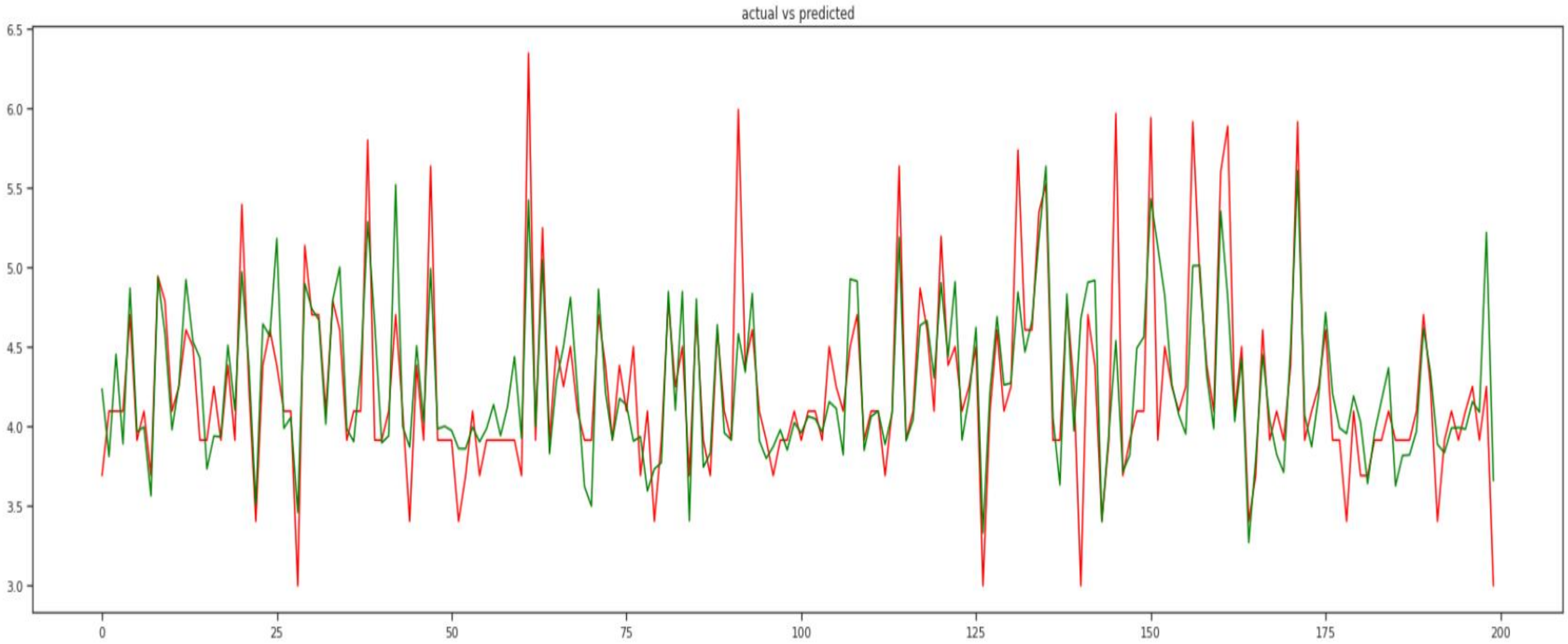
Validation and selection

	model_name	accuracy	r2_val	rmse_val
0	lasso	89.741500	0.080544	0.622252
1	ridge	90.694278	0.187442	0.584963
2	knn	93.697068	0.591920	0.414547
3	xgb	92.242783	0.394802	0.504836
4	svr	91.763680	0.216791	0.574301
5	random_forest	94.471283	0.684807	0.364325
6	gradient	92.235615	0.394756	0.504854
7	cat	93.607803	0.598229	0.411330



R2 Scores
r2_val
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
lasso
ridge
knn
xgb
svr
random_forest
gradient
cat
model_name
Rmse Scores
rmse_val
0.0
0.1
0.2
0.3
0.4
0.5
0.6
lasso
ridge
knn
xgb
svr
random_forest
gradient
cat
model_name
accuracy
accuracy
0
20
40
60
80
lasso
ridge
knn
xgb
svr
random_forest
gradient
cat
model_name

Actual vs Predicted



Hyperparameters

```
[ ]
from sklearn.model_selection import RandomizedSearchCV
# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 50, stop = 200, num = 10)
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
# Method of selecting samples for training each tree
bootstrap = [True, False]
# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
```

Slight Improvement

	NAME OF MODEL	R2 SCORES	ACCURACIES	RMSE
0	random_forest	0.684807	94.471283	0.364325
1	random_foresst_after_tuning	0.709293	94.724074	0.349888

Other things I tried

- Removing Outliers – Didn’t work well
- Including variables with very less correlation with target variable gave less accuracy and r2 scores
- Ridge and Lasso didn’t cross 91% accuracy

Conclusion

- Final accuracy was 94.7 with random forest
- Catboost also worked well
- Randomized SearchCV had slight improvement of 0.4%
- Independent Variable selection worked well
- It can improve with further tuning

Challenges

- Feature selection – so many features and their correlation was confusing
- Computation time in randomized search CV

Q & A