Dhruv Sawhney

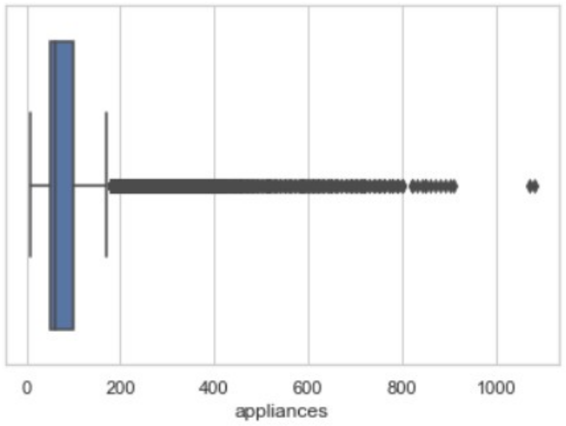
Appliances Electricity Usage Prediction

Summary

- The objective of this project is to accurately predict the amount of energy consumed by appliances in a household.
- Performed outlier treatment, segmented datetime in hours, day, week and month, and visualized them in respect to the appliances variable to check for the dependency on time.
- Built Correlation Matrix to check for the correlation between the variables and pair plot to examine the linear dependence among some features of our data set.
- Ran a random forest model with all the original features to examine the importance of each one to predict appliances consumption.
- Divided features into 3 models (Linear Model/Support Vector Regression Model/Random Forest Model)
- Performed scaling on the Train and Test dataset to normalize the data for better accuracy
- Implemented Linear/SVR/RF regression techniques and compared their R^2 and the Accuracy
- Used timeseries Split Cross Validation on all the three models
- Optimized the performance of the selected model using parameter tuning.
- Concluded Random Forest Regression as the best suited model to predict the target variance-Appliance electricity usage.

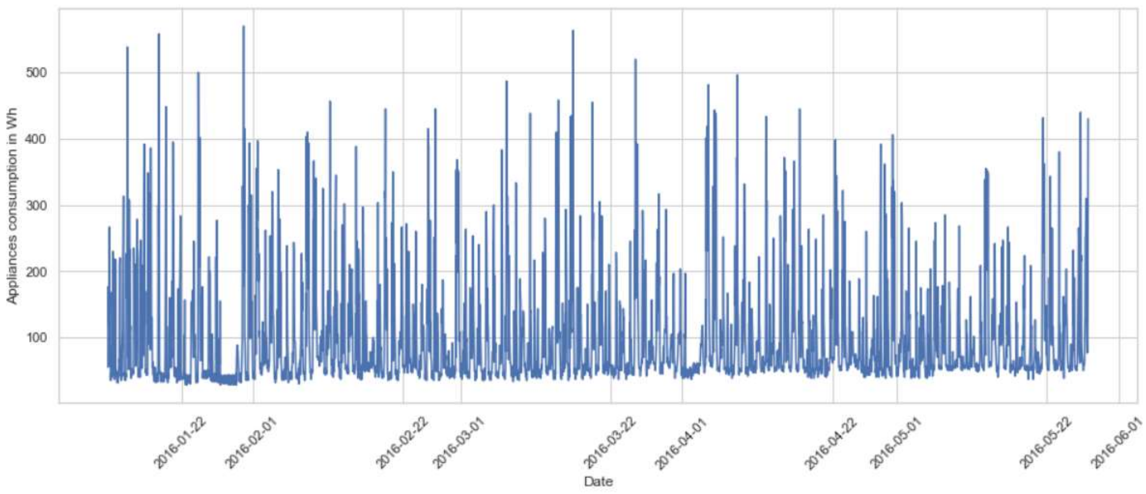
Data Preparation and Exploratory Data Analysis

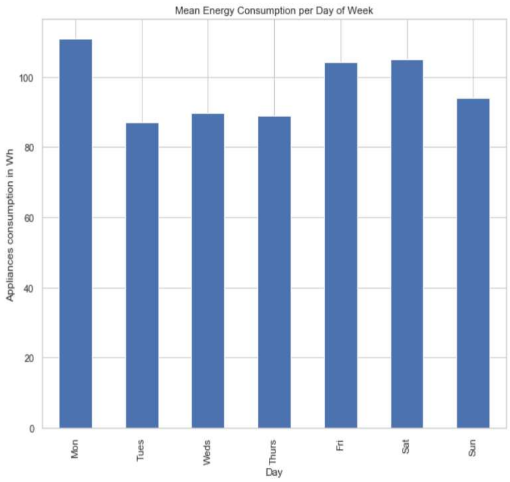
The first step was to look for the **missing values** in the dataset. I noticed that there were no misisng values. Then we checked for the outliers in the appliances variable. Below is the plot for the same

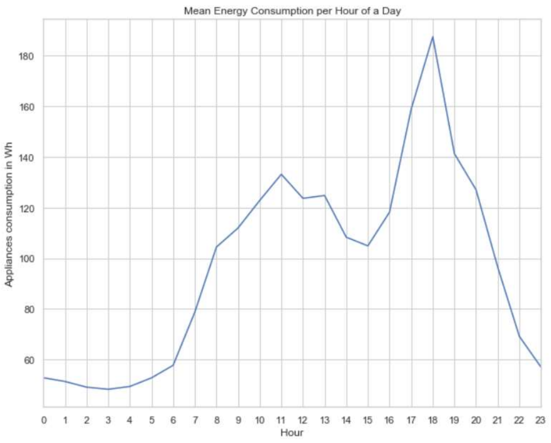


As load greater than 800Wh doesn't seem to be logical, we excluded the absolute 0 values and 1% top values of the appliance's variable.

After the outlier treatment, we segmented the time in hours, day, week and month as the appliances load is dependent on them and further plotted mean energy consumption over 4.5 months, Mean energy consumption per day per week, and mean energy consumption per hour in a day. Below are the plots for the same.

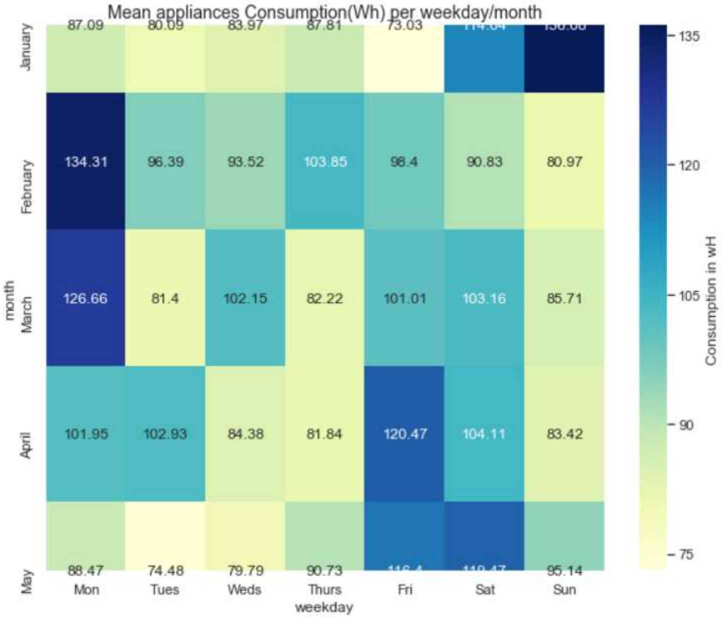




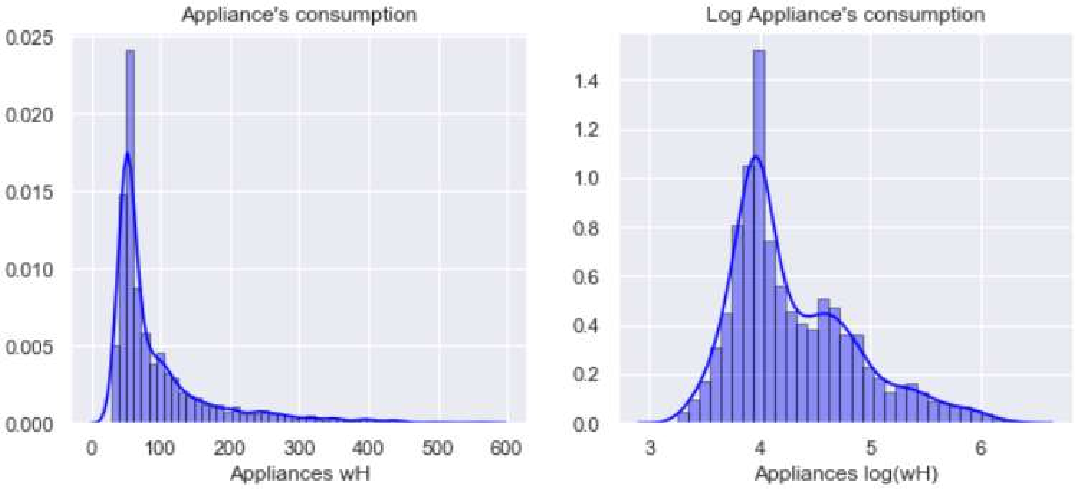


The bar plot on the left-hand side signifies that Monday has the highest consumption followed by weekends and Tuesday has the lowest out of all the seven days. The plot on the right-hand side signifies that people mostly do not use any appliances during the night hours. The appliances consumptions is below 80 Wh from 10 pm to 7 am. During morning hours, the consumptions is between 120- 135 Wh. The highest consumption is between 4pm to 8pm where the consumption ranges from 120 Wh to 185 Wh, when families are usually at home and using the appliances to the fullest.

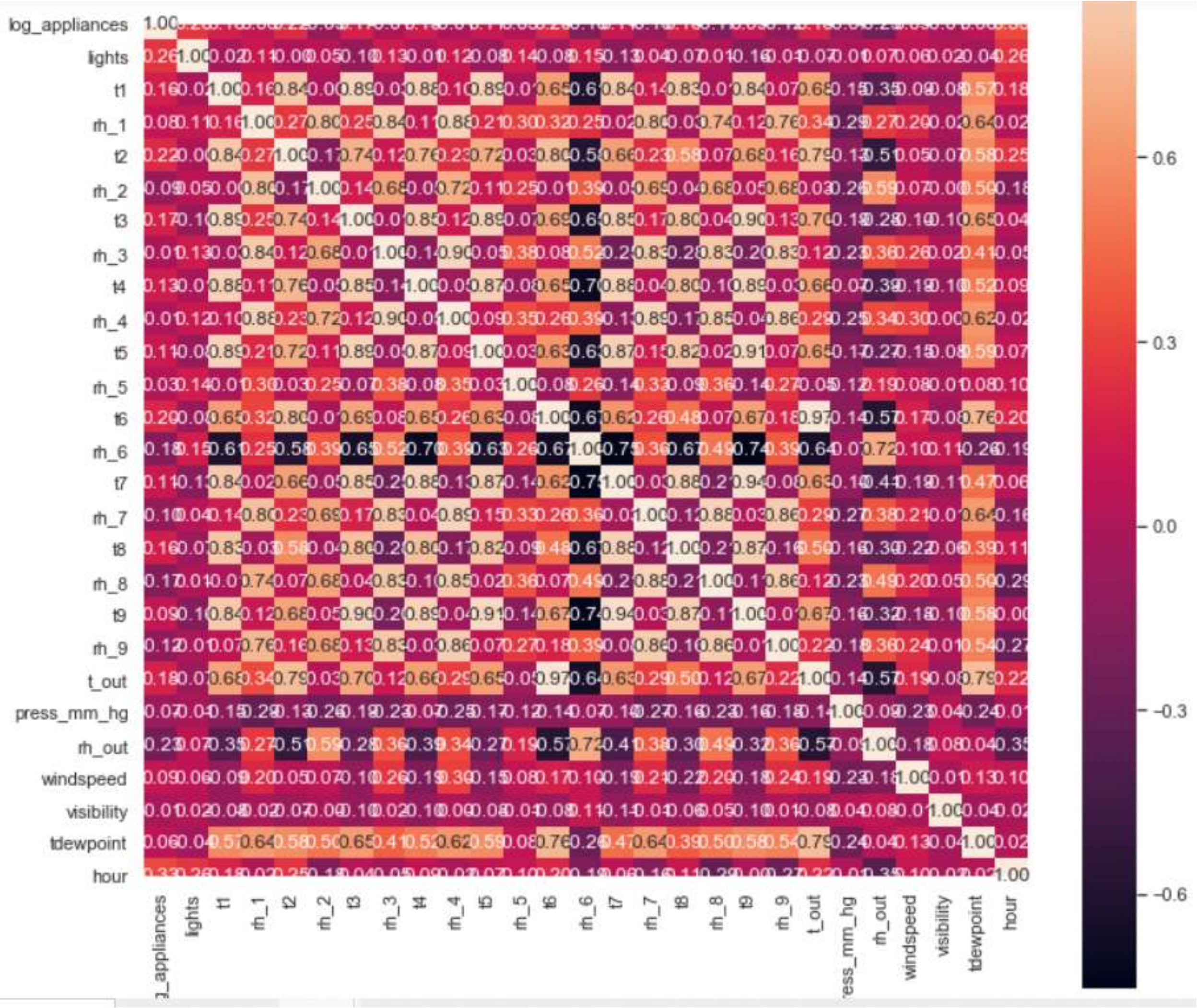
We further plotted the heatmap to verify the same.



Additionally, Histogram of the appliance's consumption was plotted. The distribution of the energy consumption came out to be right skewed, thus, to make a better and accurate model, we performed log normal conversion on the appliance's variable. Below are the plots for the same.

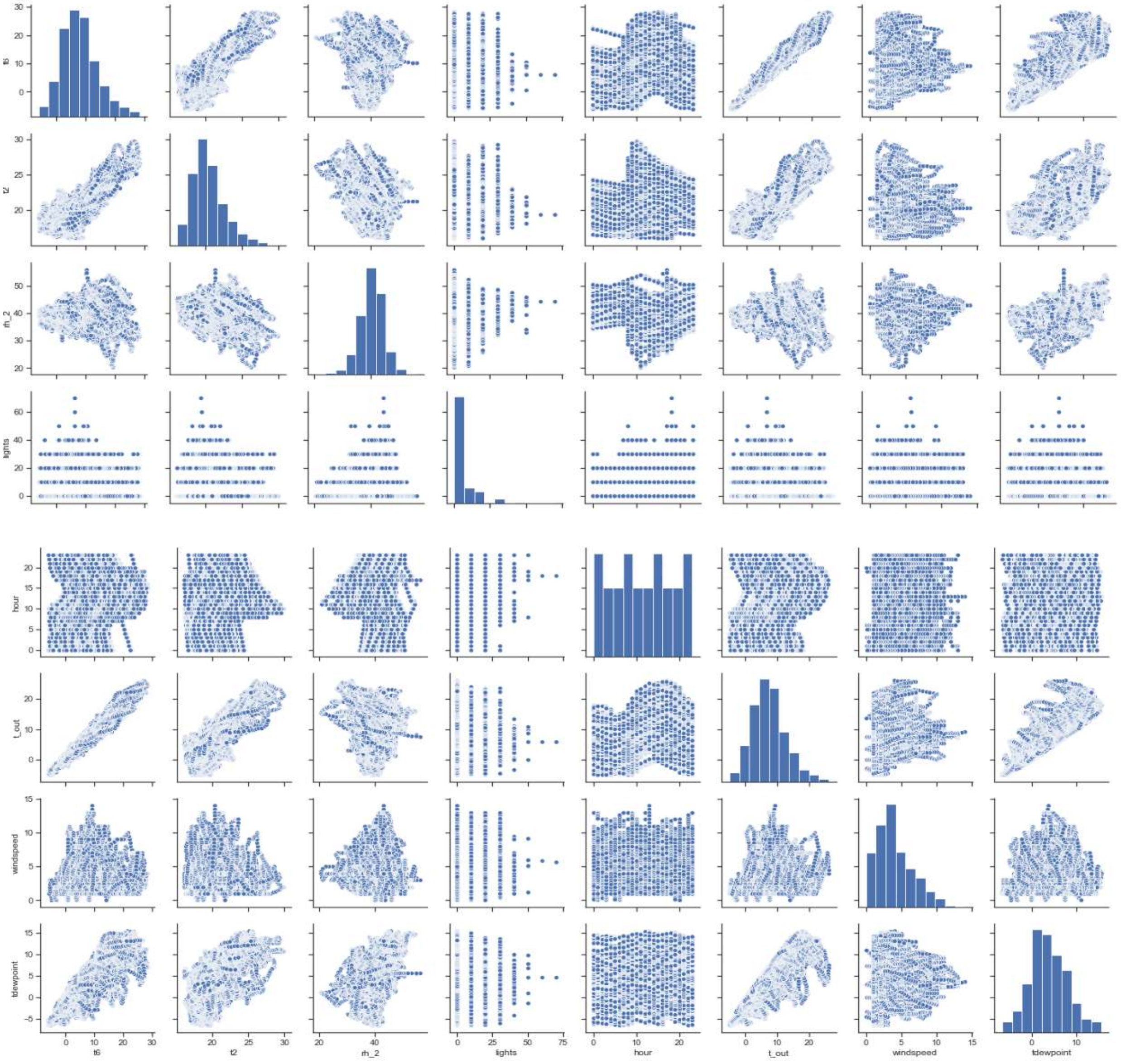


Moreover, we made a correlation matrix to check for the correlation between the features.



The correlation plot shows high correlation between humidity and temperature features between different rooms which was expected. There is an unexpected negative correlation of RH_6(Humidity outside the building, in %) with other humidity measures. The trend of value shows that feature is highly variant from 100% to 0% from one date to another date. This can be due to some calibration issues of the device. In the further analysis, it might be better to remove this variable.

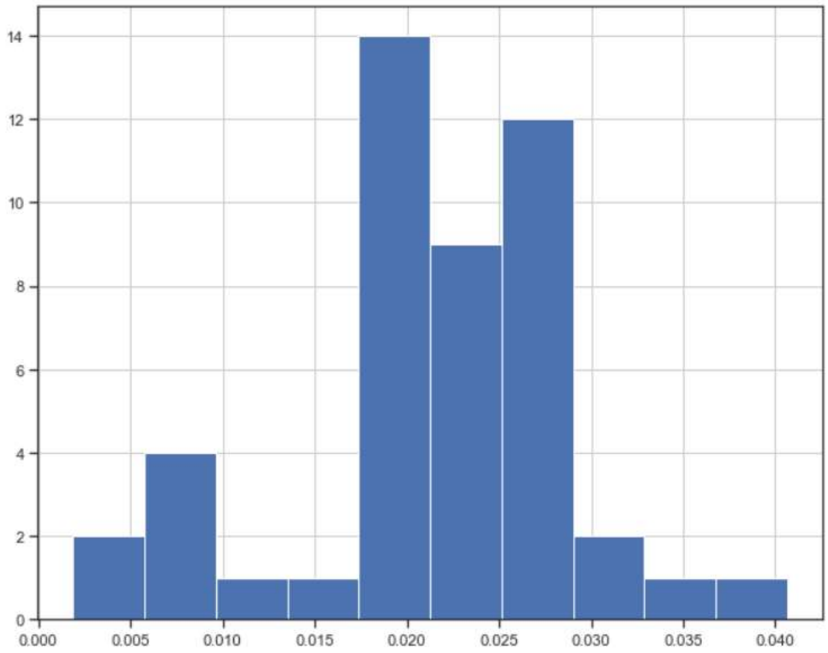
Next step is to examine linear dependence among some features of our data set. In a linear regression problem, linear independent variables are used as features to explain energy consumption to avoid **multicollinearity issues**.



From the above plotted pair-plots, we noticed that the temperature(in or out) features and tdewpoint have a linear relationship. Since, temperature in and temperature out are highly correlated thus, we can use temperature in the house as a feature in the linear model in order to optimize model performance.

Model Selection and Evaluation

Next, we ran a random forest model with all the original features to examine the importance of each feature required to predict the appliances' consumption. Below is the plot for the frequency distribution of the importance score of the features.



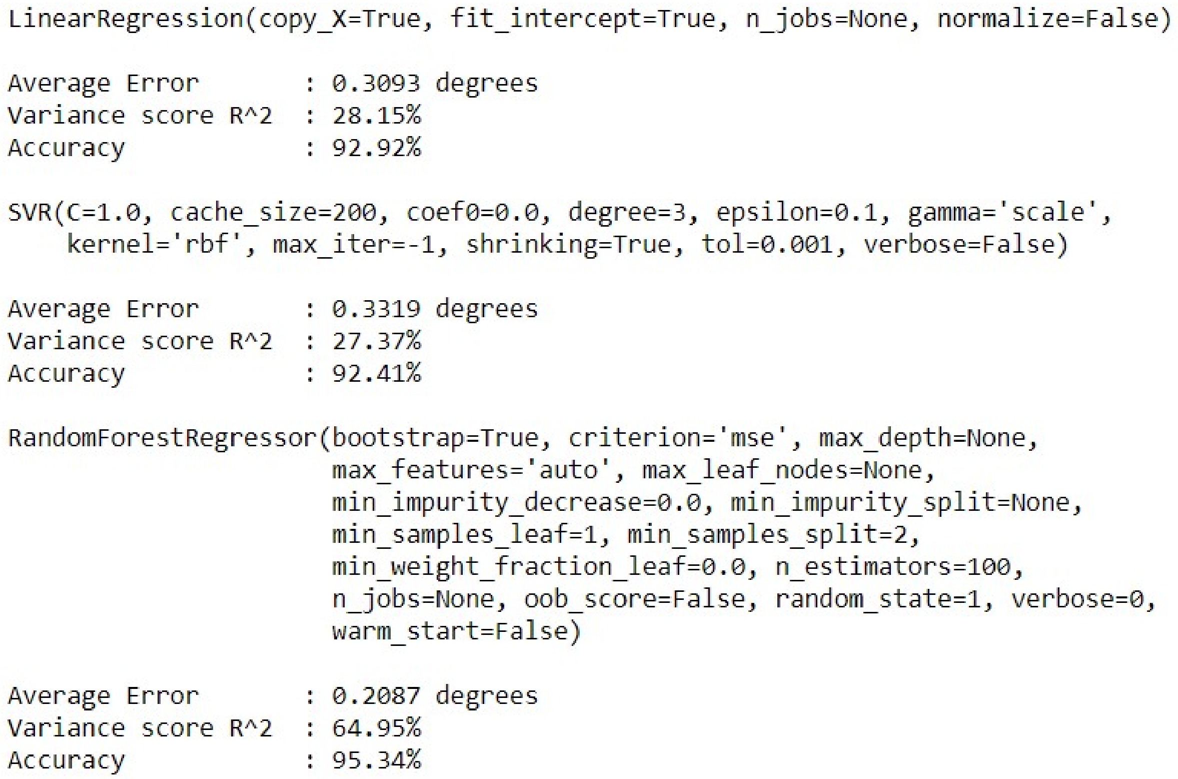
Subsequently, we found out from the Random Forest Classifier output that there are 25 significant features.

Justification for the chosen regression techniques

As observed from the correlation graph in the EDA process, independent variables were varying linearly with the dependent(appliances) variable. As a result, we chose Linear Regression as the benchmark algorithm for my analysis and prediction. In addition, we incorporated SVR and RandomForest for identifying the best model amongst the all.

Next, we developed three models, namely Linear Regression Model, Support Vector Regression Model and Random Forest Regressor Model.

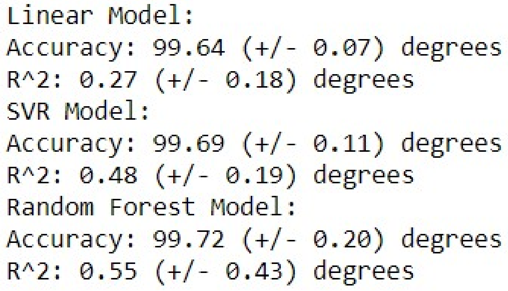
Below image depicts the performance report of the three models.



```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

Average Error       : 0.3093 degrees
Variance score R^2  : 28.15%
Accuracy            : 92.92%

SVR(C=1.0, cache_size=200, coef0=0.0, degree=3, epsilon=0.1, gamma='scale',
    kernel='rbf', max_iter=-1, shrinking=True, tol=0.001, verbose=False)

Average Error       : 0.3319 degrees
Variance score R^2  : 27.37%
Accuracy            : 92.41%

RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
                      max_features='auto', max_leaf_nodes=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=100,
                      n_jobs=None, oob_score=False, random_state=1, verbose=0,
                      warm_start=False)

Average Error       : 0.2087 degrees
Variance score R^2  : 64.95%
Accuracy            : 95.34%
```

*As shown above, Random Forest Regressor performed significantly better than the other two models with Variance score (R^2) being **64.95%** and Accuracy being **95.34%.***

Furthermore, we used timeseries Split Cross Validation on all the three models which significantly increased the accuracy by approximately 6.5% for all the three models.



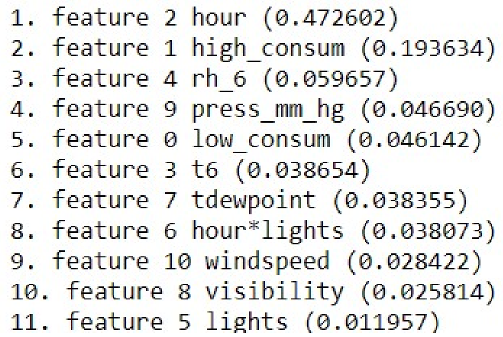
```
Linear Model:
Accuracy: 99.64 (+/- 0.07) degrees
R^2: 0.27 (+/- 0.18) degrees
SVR Model:
Accuracy: 99.69 (+/- 0.11) degrees
R^2: 0.48 (+/- 0.19) degrees
Random Forest Model:
Accuracy: 99.72 (+/- 0.20) degrees
R^2: 0.55 (+/- 0.43) degrees
```

From the above results, the Random Forest Model is the best of the three, having highest R^2 value and accuracy being almost the same for all the three models.

R^2 and accuracy metric is considered among the most accurate technique for choosing the best model because R^2 depicts the variation explained by a model whereas accuracy depicts how well our model will perform on the unseen data. We chose these metrics for all the 3 models using the TimeSeriesSplit cross validation technique. This significantly increased the accuracy of all the 3 models to 99%.

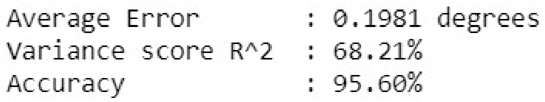
Optimization

Implemented feature importance on the Random Forest Model(RF) so find out the most important variables to predict the energy consumption.



```
1. feature 2 hour (0.472602)
2. feature 1 high_consum (0.193634)
3. feature 4 rh_6 (0.059657)
4. feature 9 press_mm_hg (0.046690)
5. feature 0 low_consum (0.046142)
6. feature 3 t6 (0.038654)
7. feature 7 tdewpoint (0.038355)
8. feature 6 hour*lights (0.038073)
9. feature 10 windspeed (0.028422)
10. feature 8 visibility (0.025814)
11. feature 5 lights (0.011957)
```

In order to optimize the variance score of RF model, Parameter tuning was executed using GridsearchCV. In Addition, we evaluated the model and observed significant increase in the variance score from 54.95% to 68.21%.



```
Average Error       : 0.1981 degrees
Variance score R^2  : 68.21%
Accuracy            : 95.60%
```

From the above analysis, we concluded that Random Forest Regression should be used to predict the target variance-Appliance electricity usage.

For further analysis, we can implement MLP Regressor, XGBoost, Multiple random forest regression models using different combinations of feature engineering to explore the best suited model for Appliance energy usage prediction.

Approximate time taken to complete the whole exercise including the writeup would be around 11-12 hours.