Supervised ML(regression) - Appliance Energy Prediction

Technical Document

Mohammed Akifuddin
akifkashif007@gmail.com

Introduction

Currently there is an uncontrollable damage to the environment because of rapid consumption of natural resources of Earth. The increase of C02 is rampant and the damage to the ozone layer is critical. The usage of appliances in daily households also are contributing to the damage of Earth and its environment. Tracking the usage and the amount of energy can be very useful in curbing the problems by keeping the usage in control. We are tasked with tracking the usage using supervised ML algorithms.

Problem Statement

We are tasked with predicting the amount of energy consumed in watt per hour(wh) by tracking the usage of appliances in the household from the data collected throughout 4.5 months at every 10 minute interval to understand the trend and growth of energy consumption of residential buildings in Belgium.

Overview of the data

1. Date : Date and time of the Appliances usage recorded
2. Appliances : Values of appliance usage in Watt per hour(Wh)
3. Lights : Energy use of light fixtures in the house in Wh
4. T1: Temperature in kitchen area, in Celsius
5. RH1 : Humidity in kitchen area, in %
6. T2 : Temperature in living room area, in Celsius
7. RH2: Humidity in living room area, in %

8. T3 : Temperature in laundry room area
9. RH3 : Humidity in laundry room area, in %
10. T4 : Temperature in office room, in Celsius
11. RH4: Humidity in office room, in %
12. T5 : Temperature in bathroom, in Celsius
13. RH5 : Humidity in bathroom, in %
14. T6 : Temperature outside the building(north side), in %
15. RH6 : Humidity outside the building(north side), in %
16. T7: Temperature in ironing room , in Celsius
17. RH7: Humidity in ironing room, in %
18. T8: Temperature in teenager room 2, in Celsius
19. RH8: Humidity in teenager room 2, in %
20. T9: Temperature in parents room, in Celsius
21. RH9: Humidity in parents room, in %
22. T0: Temperature outside (from Chievres weather station), in Celsius
23. Pressure (from Chievres weather station), in mmHg
24. RHout, Humidity outside (from Chievres weather station), in %
25. Wind speed (from Chievres weather station), in m/s
26. Visibility (from Chievres weather station), in km
27. Tdewpoint (from Chievres weather station), Â°C
28. rv1, Random variable 1, nondimensional
29. rv2, Random variable 2, nondimensional

Steps involved

I. Performing EDA (exploratory data analysis)

- Extracting the head and tail of the dataset to get insights of the data.
- Extracting all columns of the dataset.
- Extracting the info and description of the dataset to interpret the data types and mean , median etc., values of the column.
- Extracting the months, hours and weeks from the date column by setting it as index
- Proceeded with feature engineering by calculating average energy load per hour and Creating more columns in our dataset which would be helpful for creating model.
- Checking for outliers using the boxplot
- Defining the interquartile range and filtering out the values outside of it.
- Extracting correlation heatmap and calculating VIF to remove correlated and multicollinear variables.

II. Drawing conclusions from the data

- On average, the household energy appliances usage was at its peak during the 18th hour of the day.
- The warmest temperatures are the laundry room(T3), teenager room(T8) and kitchen area(T1) respectively. And the coldest rooms are bathroom(T5) and parents room(T9)
- The energy usage stays consistent from temperatures 5° to 25°. At extreme high and low temperatures, the energy usage is relatively low
- There is little to no correlation between pressure and energy consumption
- The humidity is at its highest outside the building(RH_out) and lowest in the ironing room(RH_7)
- In Feb, Mar, April, the usage of appliances is more or less equal and that of Jan was the lowest one
- At week 14 i.e, the second week of April, the usage of appliances is the highest and on the other hand, its lowest in the last week of the period i.e, the fourth week of May

III. Training the model

- Assigning the dependent and independent variables
- Splitting the model into train and test sets.
- Transforming data using minmaxscaler.
- Fitting linear regression on train set.
- Getting the predicted dependent variable values from the model.

IV. Evaluating metrics of model

A. Getting MSE , RMSE , R2-SCORE , ADJUSTED-R2 SCORE for different models used.

a. MSE - the mean squared error or mean squared deviation of an estimator measures the average of the squares of the errors.

b. RMSE - Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are.

c. R2-SCORE - R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

d. ADJUSTED-R2 SCORE - Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

B. Comparing the r2 score of all models used , to get the desired prediction.

Models used

Linear regression:

Linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Logistic regression assumptions:

- There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.
- There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
- The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.

- The error terms must have constant variance. This phenomenon is known as homoscedasticity. The presence of non-constant variance is referred to as heteroskedasticity.
- The error terms must be normally distributed.

We have to train our model considering the above assumptions

Properties of Logistic Regression:

● The line reduces the sum of squared differences between observed

values and predicted values.

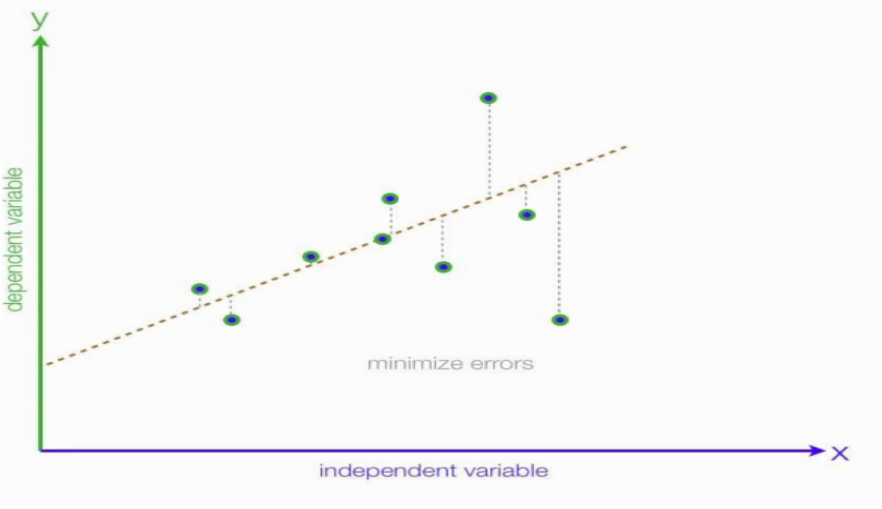
● The regression line passes through the mean of X and Y variable values

● The regression constant (b0) is equal to y-intercept the linear regression

● The regression coefficient (b0) is the slope of the regression line which is

equal to the average change in the dependent variable (Y) for a unit

change in the independent variable (X).



Lasso regression model:

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.The acronym “LASSO” stands for Least Absolute Shrinkage and Selection Operator.Lasso solutions are quadratic programming problems, which are best solved with software (like Matlab). The goal of the algorithm is to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Ridge regression model:

Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values. The cost function for ridge regression:

Min(||Y – X(theta)||^2 + λ||theta||^2)

Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. Higher the values of alpha, bigger is the penalty and therefore the magnitude of coefficients is reduced.

- It shrinks the parameters. Therefore, it is used to prevent multicollinearity
- It reduces the model complexity by coefficient shrinkage

Decision tree regression model:

Linear model trees combine linear models and decision trees to create a hybrid model that produces better predictions and leads to better insights than either model alone. A linear model tree is simply a decision tree with linear models at its nodes. This can be seen as a piecewise linear model with knots learned via a decision tree algorithm.

LMTs can be used for regression problems (e.g. with linear regression models instead of population means) or classification problems (e.g. with logistic regression instead of population modes).

Random forest regression model:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

Extra-trees regression model:

Extra Trees is an ensemble machine learning algorithm that combines the predictions from many decision trees. It is related to the widely used random forest algorithm. It can often achieve as good or better performance than the random forest algorithm, although it uses a simpler algorithm to construct the decision trees used as members of the ensemble. It is also easy to use given that it has few key hyperparameters and sensible heuristics for configuring these hyperparameters.

XGBoost Regression Model:

XGBoost is an ensemble learning and a gradient boosting algorithm for decision trees that uses a second-order approximation of the scoring function. This approximation allows XGBoost to calculate the optimal “if” condition and its impact on performance. XGBoost can then store these in its memory in the next decision tree to save recomputing it.

While training, the XGBoost algorithm constructs a graph that examines the input under various “if” statements (vertices in the graph). Whether the “if” condition is satisfied influences the next “if” condition and eventual prediction. XGBoost progressively adds more and more “if” conditions to the decision tree to build a stronger model. By doing so, the algorithm increases the number of tree levels, therefore, implementing a level-wise tree growth approach.

XGBoost learns a model faster than many other machine learning models (especially among the other ensemble methods) and works well on categorical data and limited datasets.

Challenges faced

1. Pre-processing the data was one of the challenges we faced which includes removing highly correlated variables from the data so as to not hinder the performance of our regression model.

2. Exploring all the columns and calculating VIF for multicollinearity was challenging because it might decrease the models performance.

3. Selecting the appropriate models to maximize the accuracy of our predictions was one of the challenges faced.

Conclusion

We are finally at the conclusion of our project! Coming from the beginning we did EDA on the dataset and also cleaned the data according to our needs.After that we were able to draw relevant conclusions from the given data and then we trained our model on linear regression and other models . Out of all models used , with the extra-trees regression model we were able to get the r2-score of 0.80.The model which performed poorly was Lasso Regression model with r2-score of 0.31. Given the size of data and the amount of irrelevance in the data , the above score is good.