**Appliance Energy Prediction**

**Rohit Kumar sharma, Data science trainee,**

**Alma Better, Bangalore**

**Abstract:**

With the rapid development in each sector demand of electricity is increasing by substantial rate. Energy saved is energy generated. For energy saving we need to calculate how much we consume. This is possible if you install metering devices in apartments.

Home energy monitoring by appliance-level information can provide consumers awareness on energy saving. The system can be implemented through a smart meter which requires an efficient data analysis algorithm for providing an accurate energy consumption profile.

This project aims to provide insight into reducing energy consumption by identifying trends and appliances involved.

***Keywords: machine learning, Appliance energy prediction, Regression analysis***

**1.Problem Statement**

In this time of global uncertainty world needs energy and in increasing quantities to support economic and social progress and build a better quality of life, in particular in developing countries. But even in today’s time there are many places especially in developing world where there are outages. These outages are primary because of excess load consumed by appliances at home. Heating and cooling appliances takes most power in house. In this project we will be analyzing the appliance usage in the house gathered via home sensors. All readings are taken at 10 mins intervals for 4.5 months. The goal is to predict energy consumption by appliances. In the age of smart homes, ability to predict energy consumption can not only save money for end user but can also help in generating money for user by giving excess energy back to Grid (in case of solar panels usage). In this case regression analysis will be used to predict Appliance energy usage based on data collected from various sensors. The main objective is to build a predictive model, which could help them in predicting the Energy usage proactively.

There are lots of attributes taken to predict the consumption:

* **Temperatures**: There are various temperatures for kitchen, living room, bathroom, laundry room, outside area
* **Humidity**: It contains various humidity for kitchen, laundry room, bathroom, living room.
* **Press\_mm\_Hg**: It is collection of pressure in mm Hg.
* **Visibility**: It tells about visibility.
* **Windspeed**: It tells about windspeed in that area.
* **T-dew point**: It tells about dew point temperature.
* **Appliances:** It is collection of energy usage and it is dependent variable.

**2. Introduction**

### This Dataset is for data-driven predictive models for the energy use of appliances. Data used include measurements of temperature and humidity sensors from a wireless network, weather from a nearby airport station and recorded energy use of lighting fixtures.

### data filtering to remove non-predictive parameters and feature ranking is a typical task in this dataset. From the wireless network, the data from the kitchen, laundry and living room are considered to be of highest in importance for the energy prediction.

### The prediction models with only the weather data, selected the atmospheric pressure (which is correlated to wind speed) as the most relevant weather data variable in the prediction. Therefore, atmospheric pressure may be important to include in energy prediction models and for building performance modeling.

### Our goal here is to build a predictive model, which could help consumers in predicting energy usage by appliances proactively.

## **3. Reasons for Appliance Energy Prediction.**

Some of the specific reasons are:

* Un-Even distribution of energy in various locations
* It gives an idea of energy consumed
* Excess use of energy by heating and cooling device
* Demand of high-power consumption is increasing
* Significantly reduce your utility bills

**3. Steps involved:**

* **Exploratory Data Analysis**

After loading the dataset, we performed this method by comparing our target variable that is Appliances with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

* **Null values Treatment**

Our dataset does not contain null values so we don’t have to handle missing values and it save our time.

* **Encoding of Date column**

We have extracted weekday and month from date column after converting date object to date type. This helps in finding direct relation between weekday, month and various features.

* **Feature Selection**

In this step we have used correlation heatmap to remove multicollinearity and select top features. Which feature is more important compared to our model and which is of less importance.

* **Standardization of features**

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

* **Fitting different models**

For modelling we tried various classification algorithms like:

1. **Gradient Boosting Regressor**
2. **Random forest regressor**
3. **Linear regression**
4. **Support vector regressor**
5. **Xgboost**
6. **Adaboost**
7. **LightBGm**
8. **Decision Tree Regressor**

* **Tuning the hyperparameters for better accuracy**

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting. This is the process to find optimal parameters of each column. It plays a crucial role to improve prediction.

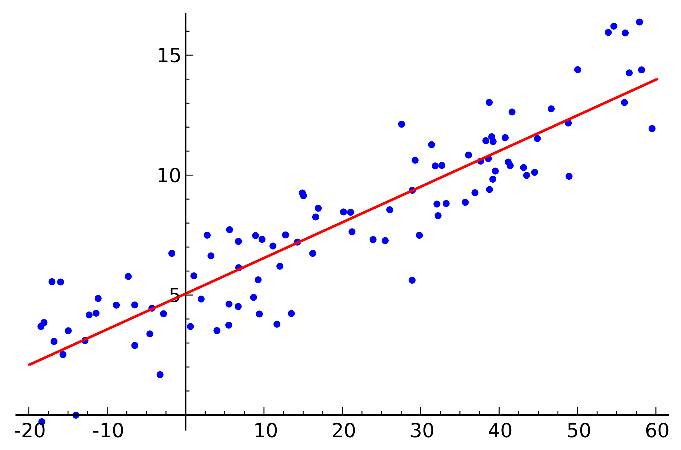
**7.1. Algorithms:**

1. **Linear Regression:**

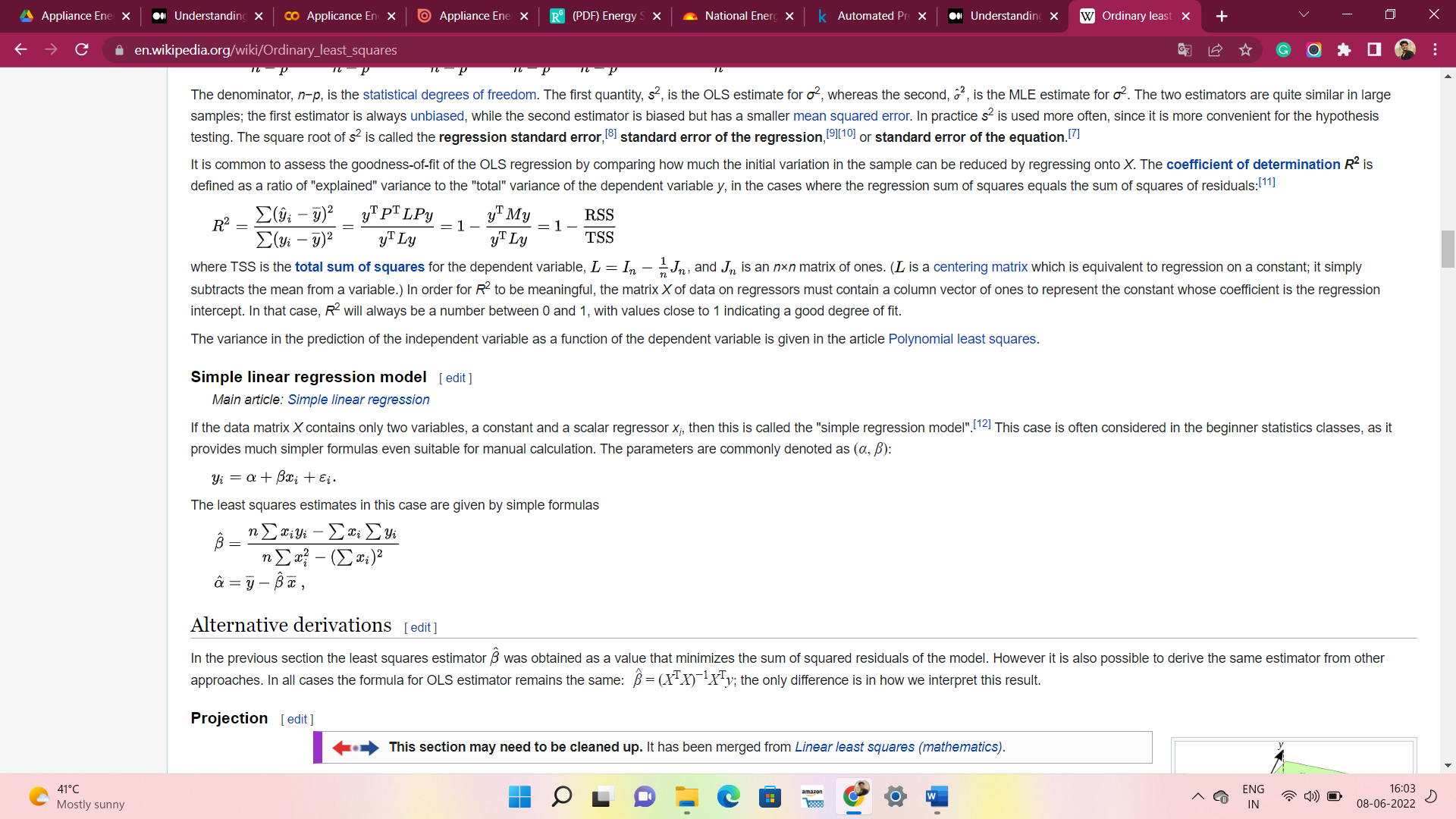
Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

The function used in Logistic Regression is:

f(x)= mx + c



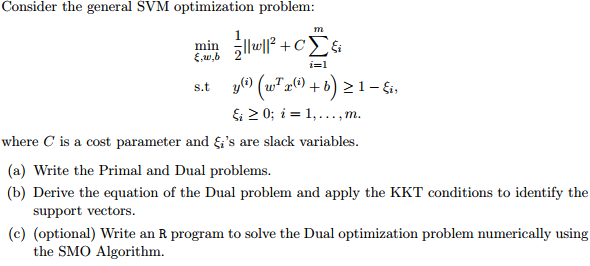
The basic model for linear regression is:

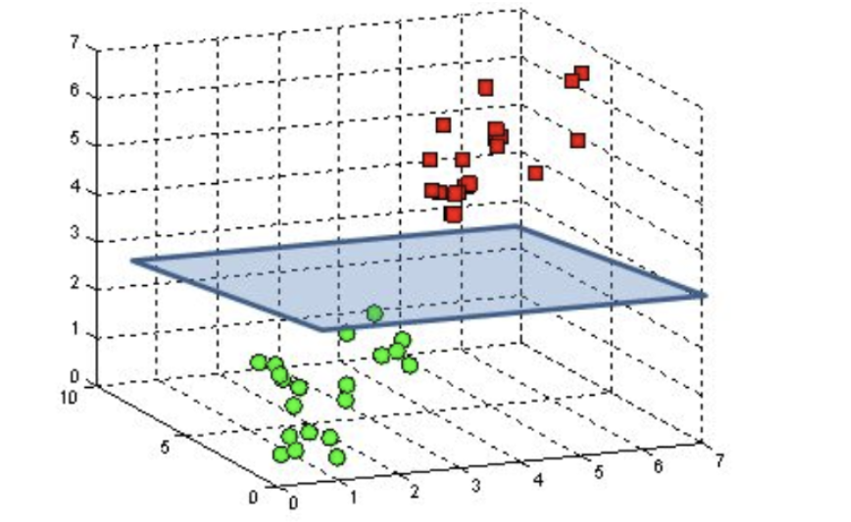


1. **Support Vector Machine Regressor:**

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

In SVM we use the optimization algorithm as:





We use hinge loss to deal with the noise when the data isn’t linearly separable.

Kernel functions can be used to map data to higher dimensions when there is inherent non linearity.

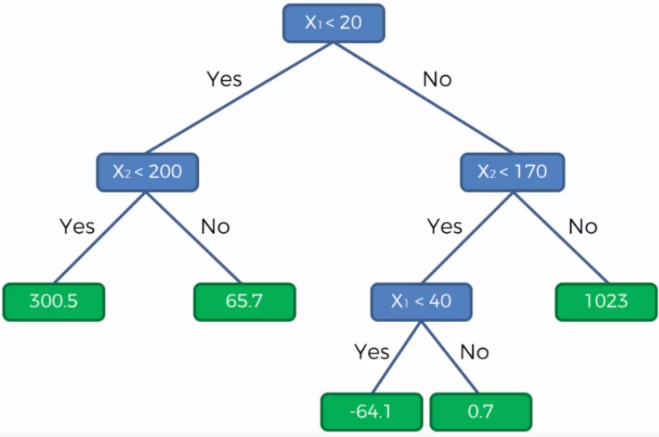
1. **XGBoost:**

XGBoost is a powerful approach for building supervised regression models. The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e., how far the model results are from the real values.

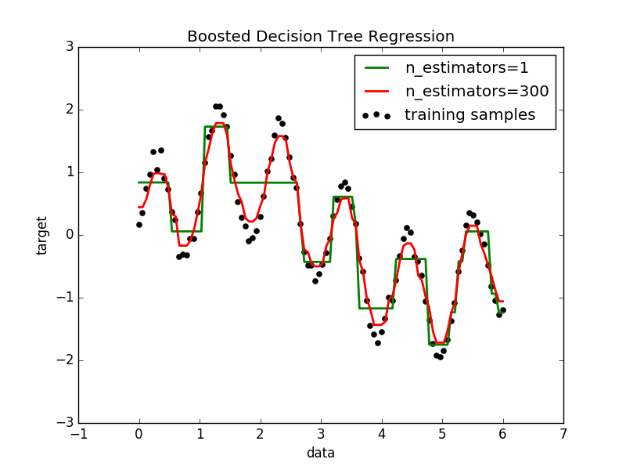
XGBoost expects to have the base learners which are uniformly bad at the remainder so that when all the predictions are combined, bad predictions cancel out and better one sums up to form final good predictions.

1. **Decision Tree Regressor-**

A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers the to the question; and the leaves represent the actual output or class label.

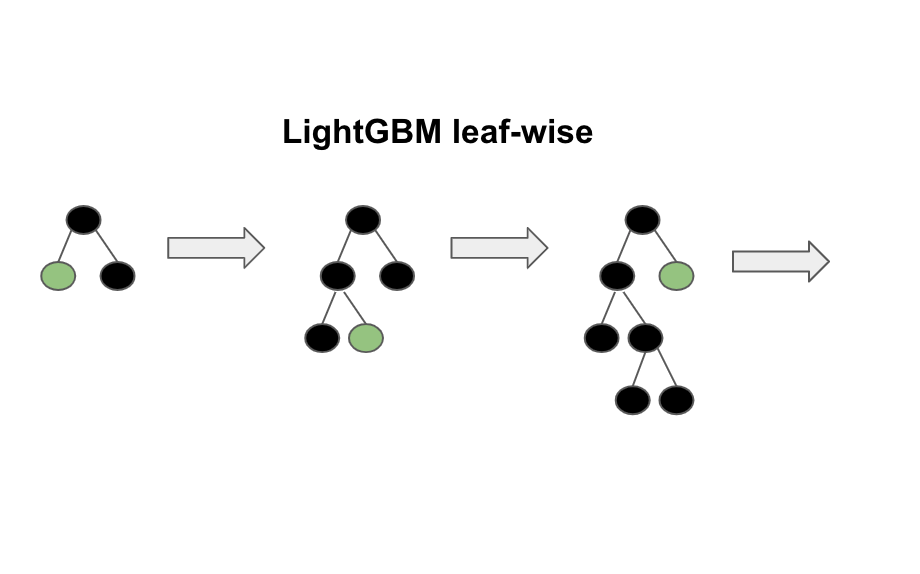
Decision trees classify the examples by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the exam

1. **AdaBoost:**

An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. As such, subsequent regressors focus more on difficult cases. 

1. **LightGBM:**

LGBM is an open-source gradient boosting library that has gained tremendous popularity and fondness among machine learning practitioners. It has also become one of the go-to libraries in Kaggle competitions. It can be used to train models on tabular data with incredible speed and accuracy. This performance is a result of the way LightGBM samples the data (**GOSS**— Gradient-based One-Sided Sampling) and reduces the number of features (**EFB**— Exclusive Feature Bundling) in sparse datasets during training.



**7.2. Model performance:**

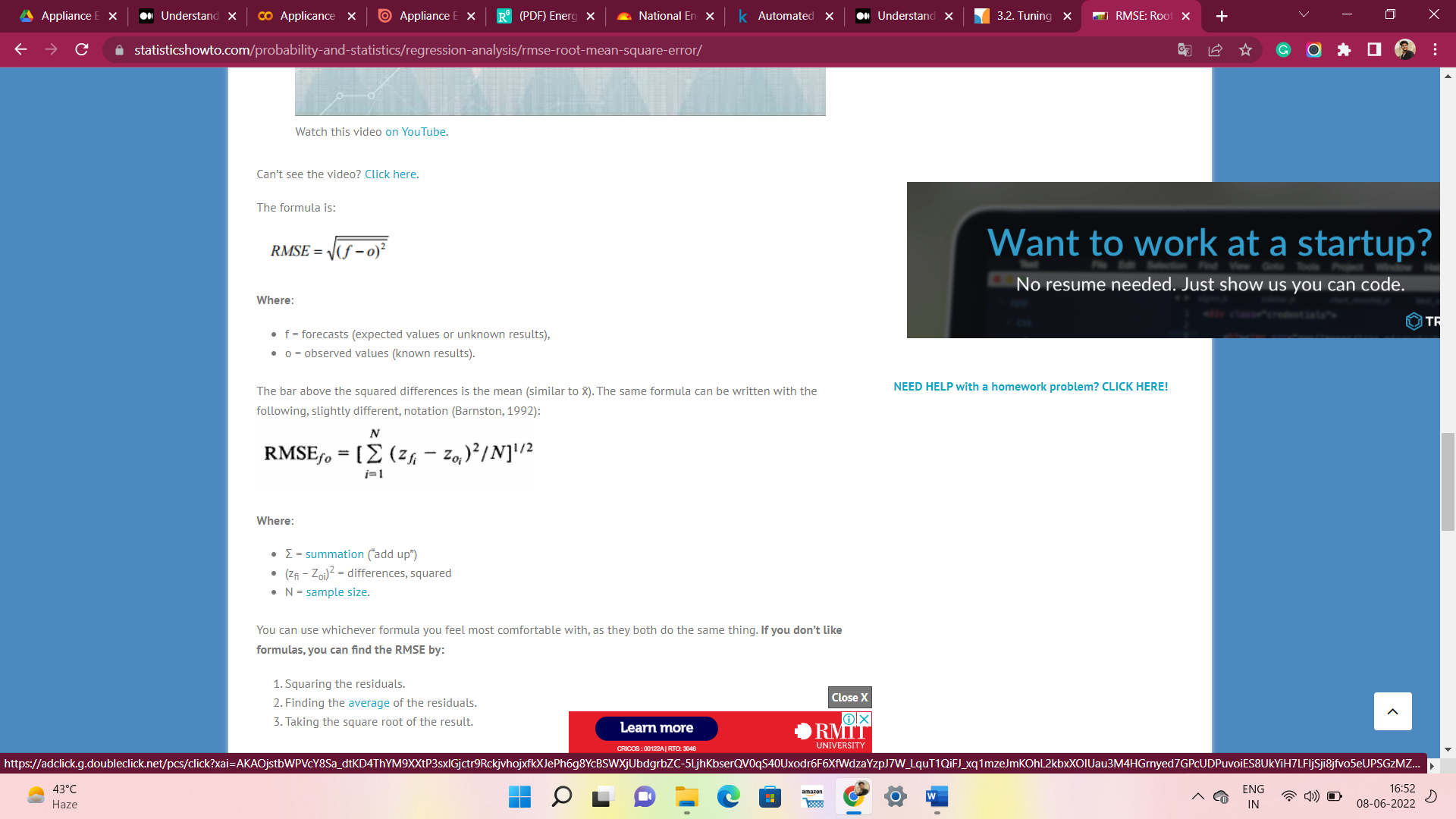
Model can be evaluated by various metrics such as:

1. **Mean squared Error**-

Mean squared error measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

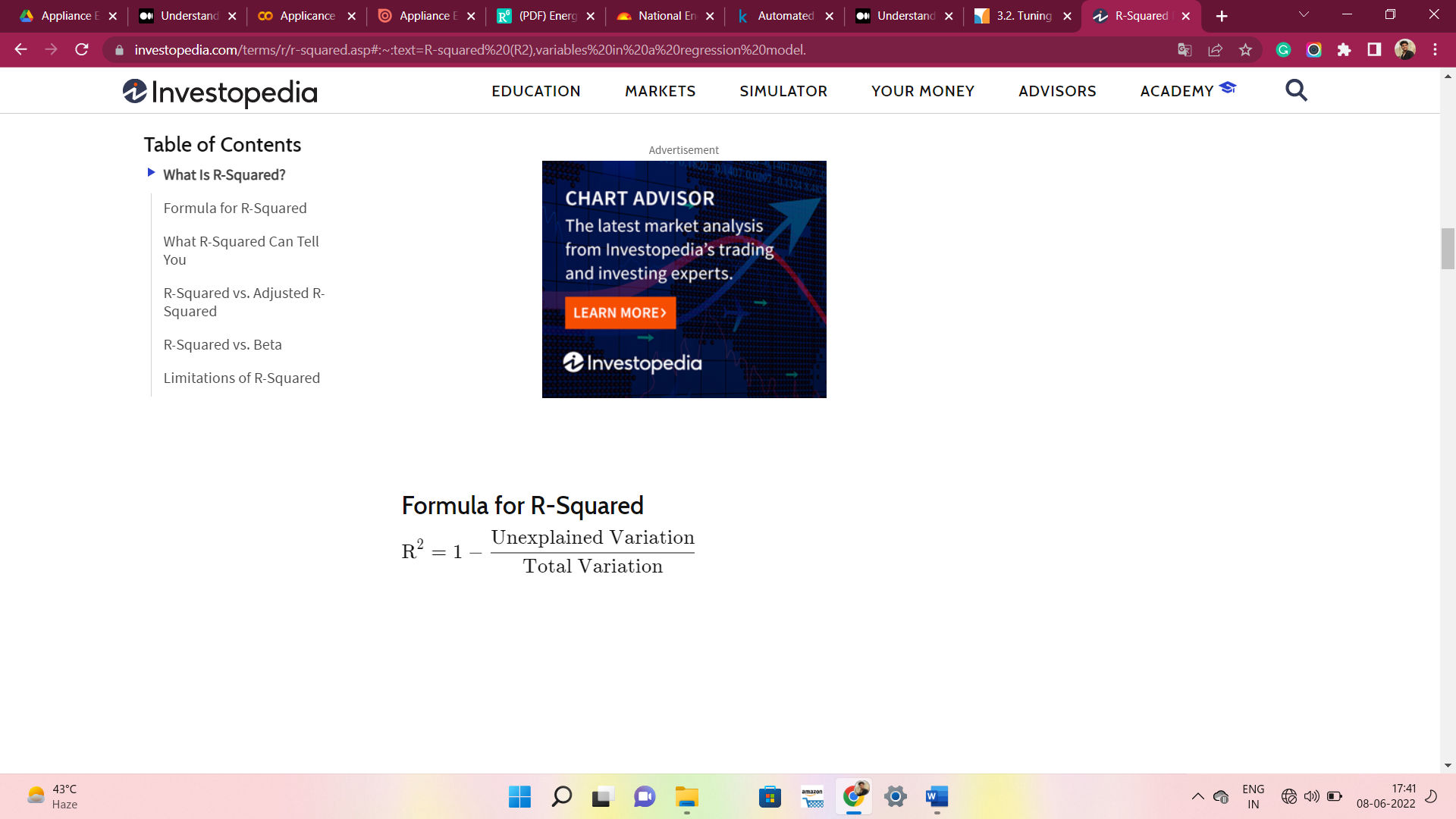
1. **Root Mean Squared Error (RMSE)**-

**Root Mean Square Error**(RMSE) is the [standard deviation](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) of the [residuals](https://www.statisticshowto.com/residual/) ([prediction errors](https://www.statisticshowto.com/prediction-error-definition/)). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the [line of best fit](https://www.statisticshowto.com/line-of-best-fit/).



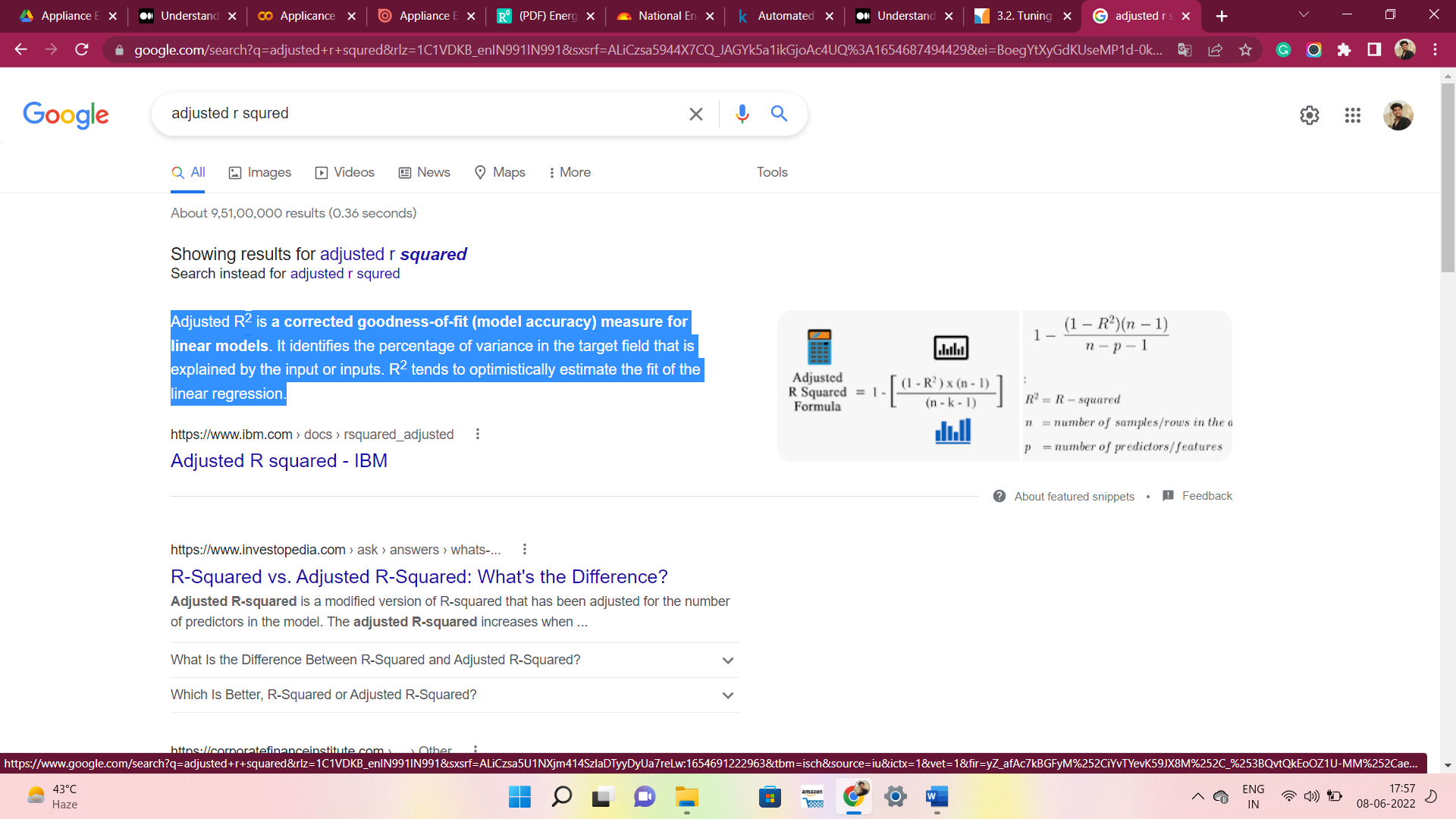
1. **R-squared**-

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



1. **Adjusted R-squared**-

Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R2 tends to optimistically estimate the fit of the linear regression.



**7.3. Hyper parameter tuning:**

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV and Randomized search for hyperparameter tuning. This also results in cross validation and in our case, we divided the dataset into different folds.

**Grid Search CV-**Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

**Randomized Search CV-** While using a grid of parameter settings is currently the most widely used method for parameter optimization, other search methods have more favorable properties. [RandomizedSearchCV](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html#sklearn.model_selection.RandomizedSearchCV) implements a randomized search over parameters, where each setting is sampled from a distribution over possible parameter values.

**8. Conclusion:**

That's it! We reached the end of our exercise.

Starting with loading the data so far, we have done EDA, encoding of categorical columns, feature selection, feature scaling and then model building.

In all of these models our accuracy revolves in the range of 70-90%. And there is no such improvement in accuracy score even after hyperparameter tuning. So, the accuracy of our best model is 90%.

* Linear Regression has the least Adjusted R-squared score hence it is underfitting.
* Random Forest Regressor is the best model which gives score of 0.89  before tuning and after hyperparameter tuning 0.9
* Gradient Boosting Regressor after  Hyperparameter Tuning is also  giving good score approximately .87
* Windspeed is little bit effecting appliances usage
* The month in which windspeed is low appliance usage is more may be due to more usage of cooler and fan.

**References-**

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya