**­­Appliance Energy Prediction**

**Ruchitha Bijja, Data Science Trainee,**

**AlmaBetter**

**Abstract:**

**Energy conservation** is the effort to reduce wasteful energy consumption by using fewer energy services. This can be done by using energy more effectively (using less energy for continuous service) or changing one's behavior to use less service. Energy conservation can be achieved through energy efficiency, which has a number of advantages, including a reduction in greenhouse gas emissions, a smaller carbon footprint, and cost, water, and energy savings.

Energy analytics will help us in determining the relevant market business forecasts and predict future discrepancies in the market. This is possible thanks to replicable, resilient and evidence-based assessments at any division.

***Keywords: Machine Learning, Appliance Energy Prediction, Exploratory Data Analysis, Regression***

**1.Problem Statement:**

When it comes to residential energy consumption, people are constantly striving for ways to reduce their monthly bills and energy usage and wastage. We use energy every day in a variety of areas of our daily lives. In this project we are considering appliance usage by analyzing the data driven from 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.

**The dataset that we are using consists of a number of features:**

* **Temperature:** The temperatures are recorded for various places (using sensors) like living room, bathroom, kitchen, laundry and also outside area.
* **Humidity:** Similar to temperature, humidity is also measured for living room, bathroom, kitchen, laundry and outside area.
* **Pressure:** The pressures are recorded in mmHg.
* **Visibility:** The visibility is present in km.
* **T Dewpoint:** This tells us about the dew point temperature.
* **Appliances:** The energy consumed in Wh. This column is our dependent variable.

**2. Introduction:**

The dataset is used for data driven predictive modelling of the energy usage. Data used here contains assessment of the temperature and humidity using wireless sensors, weathering conditions like pressure, windspeed, visibility is taken from a nearby airport

It is important to understand the energy consumption behaviors in the residential areas and predict the energy usage by home appliances to decide the energy management and reduce the consumptions. This project focuses on predicting the energy consumption of appliances based on temperature, pressure, humidity, windspeed, visibility.

The main objective of this project is to predict the energy consumption by the home appliances. With the oncoming of smart homes and the rising need for energy management, existing smart systems can benefit us with accurate prediction. If the energy usage can be predicted for every possible state of appliances, then device control can be optimized for energy savings as well.

In this case of Regression analysis (which is part of the Supervised Learning problem) Appliance energy usage is the dependent or the target variable while sensor data and weather data are the features.

**3. Reasons for Energy Management:**

* Energy management saves cost.
* It reduces the risk of energy scarcity.
* Reduces greenhouse effect.
* Maintains energy price.
* Meet Statutory compliances.
* Reducing personal utility bills.

This actually gave us an opportunity t built a model which will predict the energy consumed.

**4.Steps Involved:**

* **Exploratory Data Analysis:**

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it.

This process helped us understanding the basic relationship, features is having with the label i.e., dependent variable. It helped us creating a proper intuition in order to reduce any computational error.

* **Null Value Treatment:**

One way of handling missing values is the deletion of the rows or columns having null values.

Our dataset is very much reliable as it contains no null values.

* **Data Visualization:**

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain. We have understood the behaviors of features and label using visualization techniques.

* **Feature selection:**

In this phase we have selected the most important features used for our model. Here we have made the use of correlation heatmaps and mathematically variance inflation factor to understand multicollinearity and respectively make our decisions.

* **Standardization of Features:**

Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. While we implement any machine learning algorithm it could be a possibility that objective functions will not work properly without normalization.

* **Algorithm Fitting:**

We have made the use of the following algorithms.

1. **Linear Regression**
2. **Lasso Regression**
3. **Ridge Regression**
4. **ElasticNet Regression**
5. **Random Forest**
6. **Gradient Boosting**
7. **Extreme Gradient Boosting.**

* **Hyperparameter Tuning:**

For better accuracy and to reduce overfitting it is an important step we have made the use of gridsearch cross validation to achieve the best parameters in order to enhance the predicting capabilities of our model.

* **Model Explainability:**

This process helps us understanding the involvement of features and their impact on the target variable mostly use for complex model interpretability. We have made the use of ELI5 technique to achieve the outcome of this step.

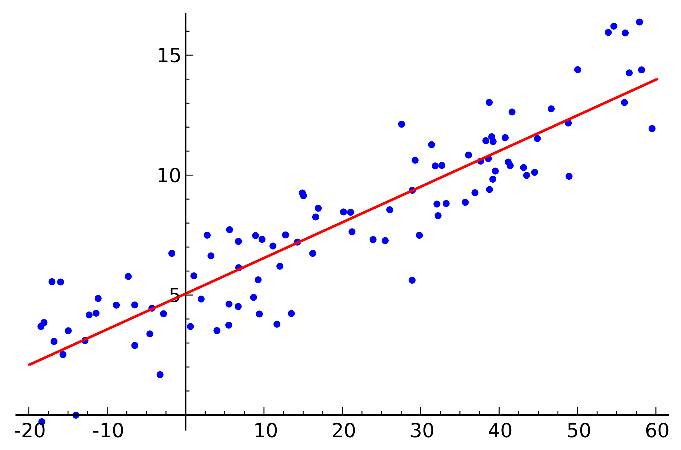
**5. Algorithms:**

1. **Linear Regression:**

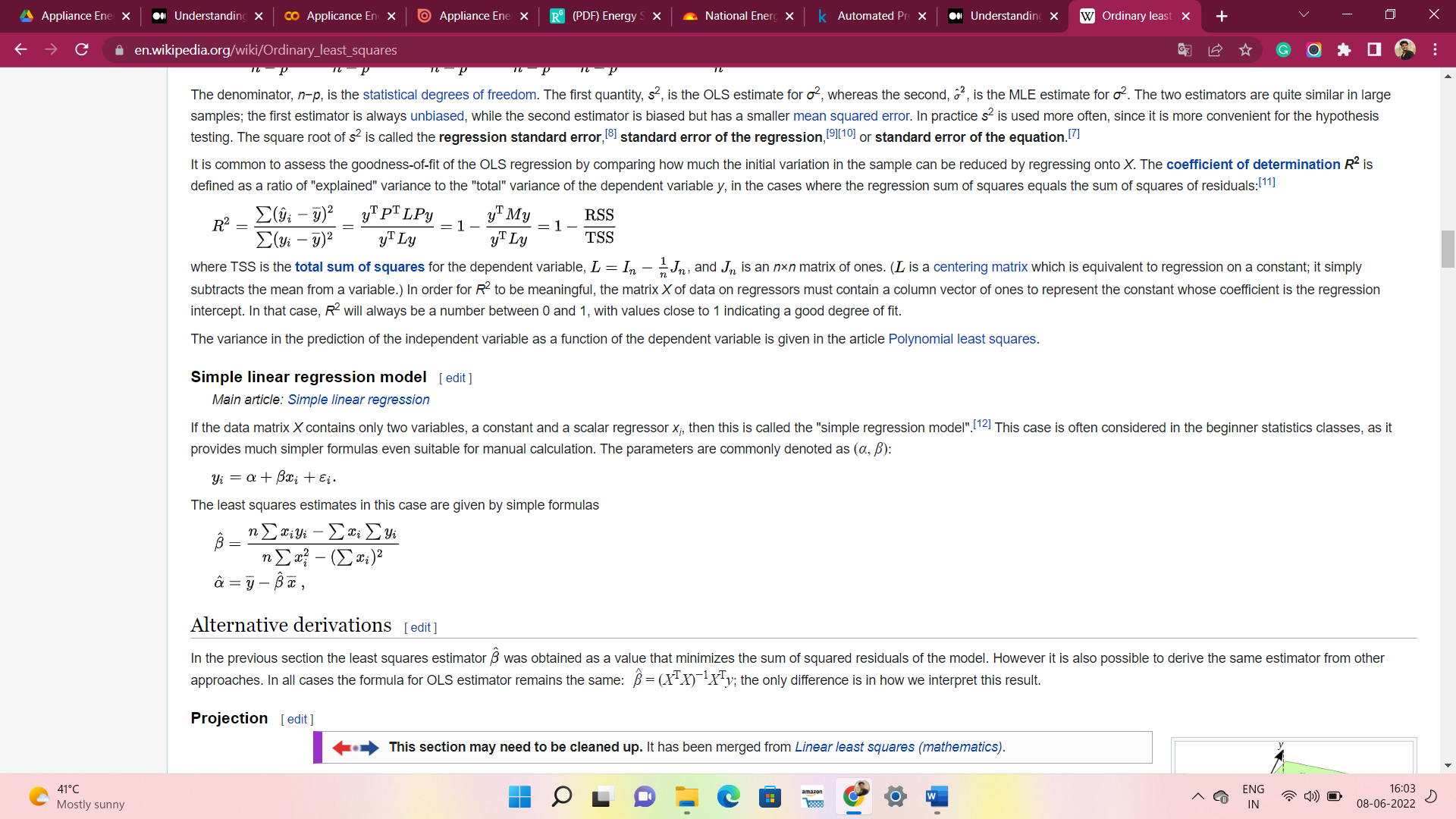
Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable we want to predict is called the dependent variable. The variable we are using to predict the other variable's value is called the independent variable.

The function used in Logistic Regression is:

f(x)= mx + c

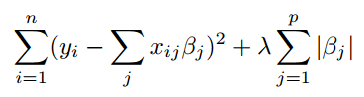


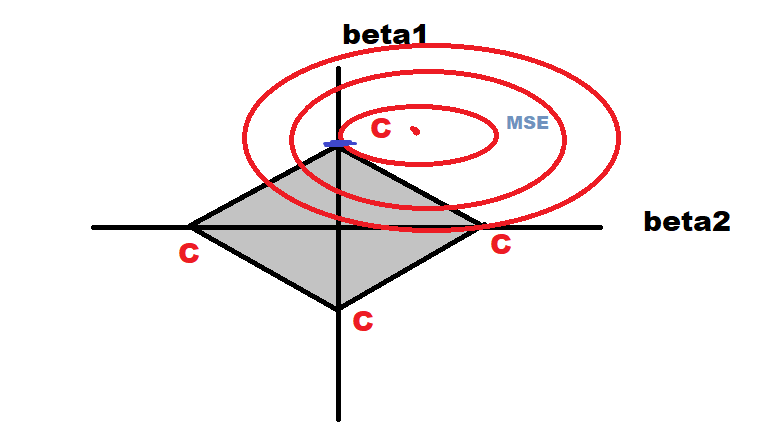
The basic model for linear regression is:



1. **Lasso Regression:**

 It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters).

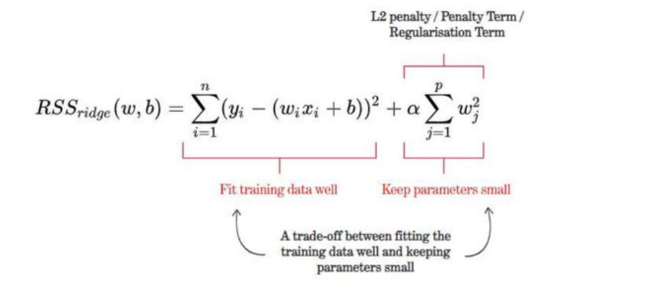


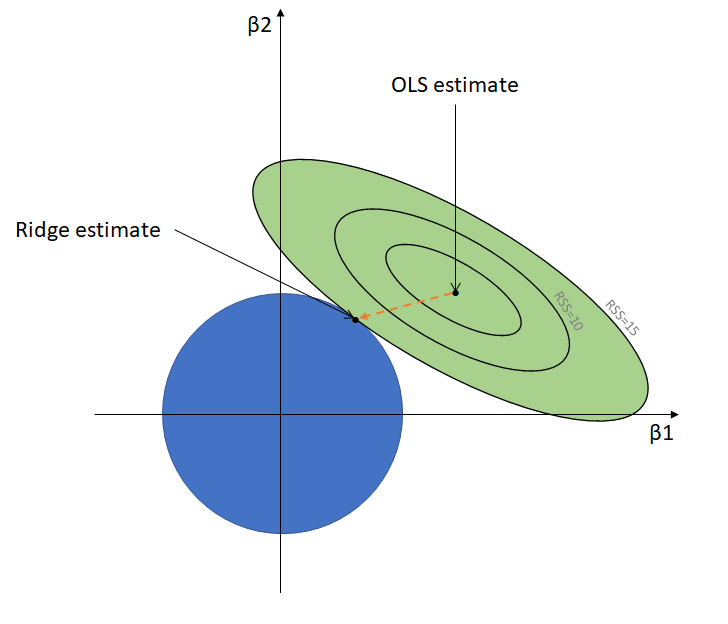


1. **Ridge Regression:**

Ridge regression is a model tuning method that is used to analyses any data that suffers from multicollinearity. This method performs L2 regularization.

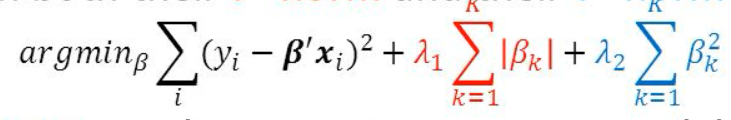
It shrinks the coordinates with respect to the orthonormal basis formed by the principal components.

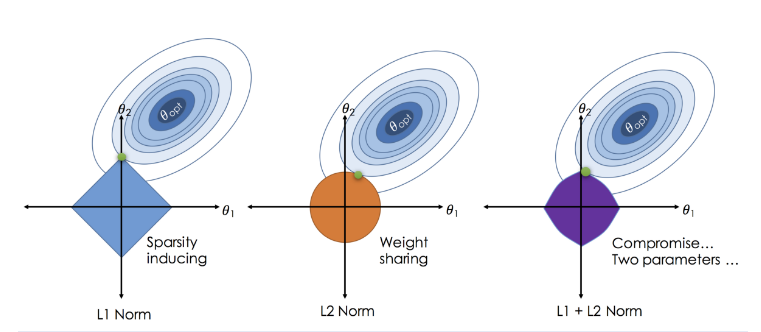


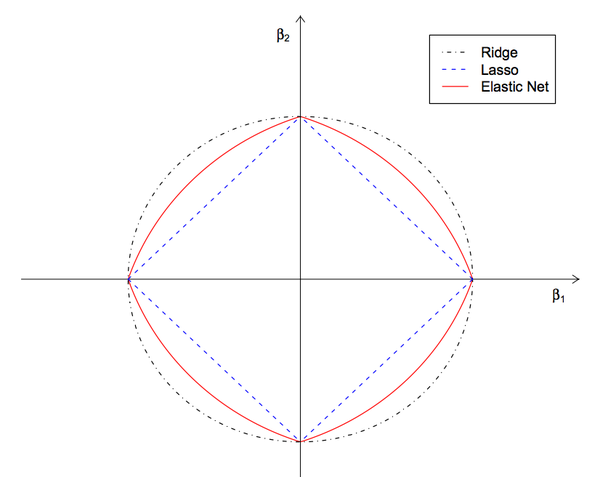


1. **ElasticNet Regression:**

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models.







1. **Random Forest Regression:**

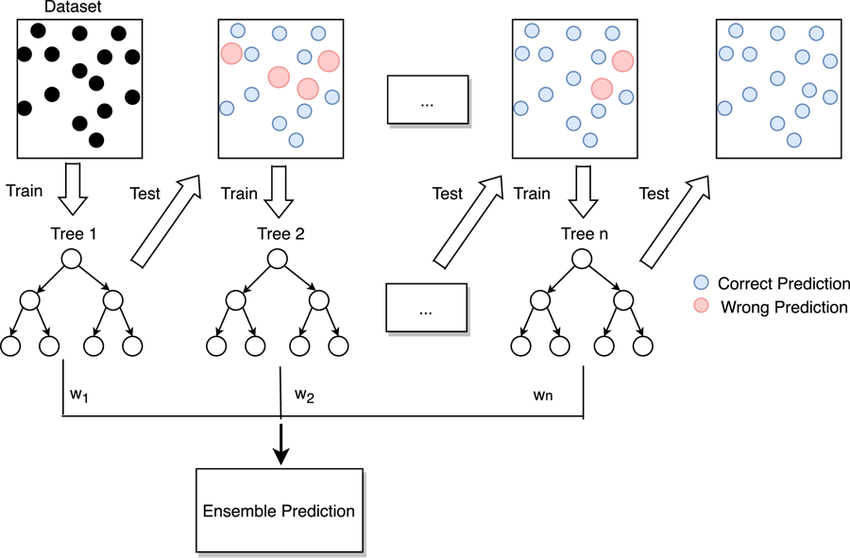
Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.



1. **Gradient Boosting:**

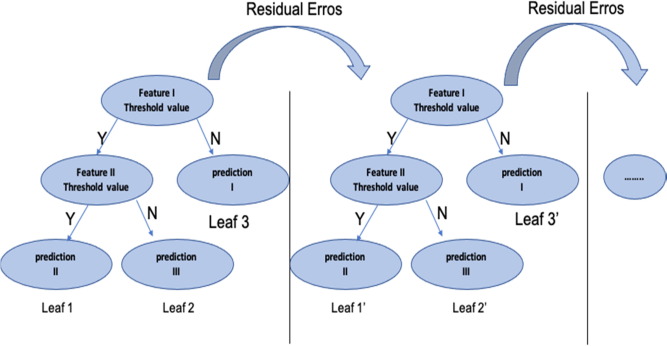
Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. Machine learning solutions for businesses, this algorithm has produced the best results. We already know that errors play a major role in any machine learning algorithm. There are mainly two types of error, bias error and variance error. Gradient boost algorithm helps us minimize bias error of the model.



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.



**6. Model performance parameters:**

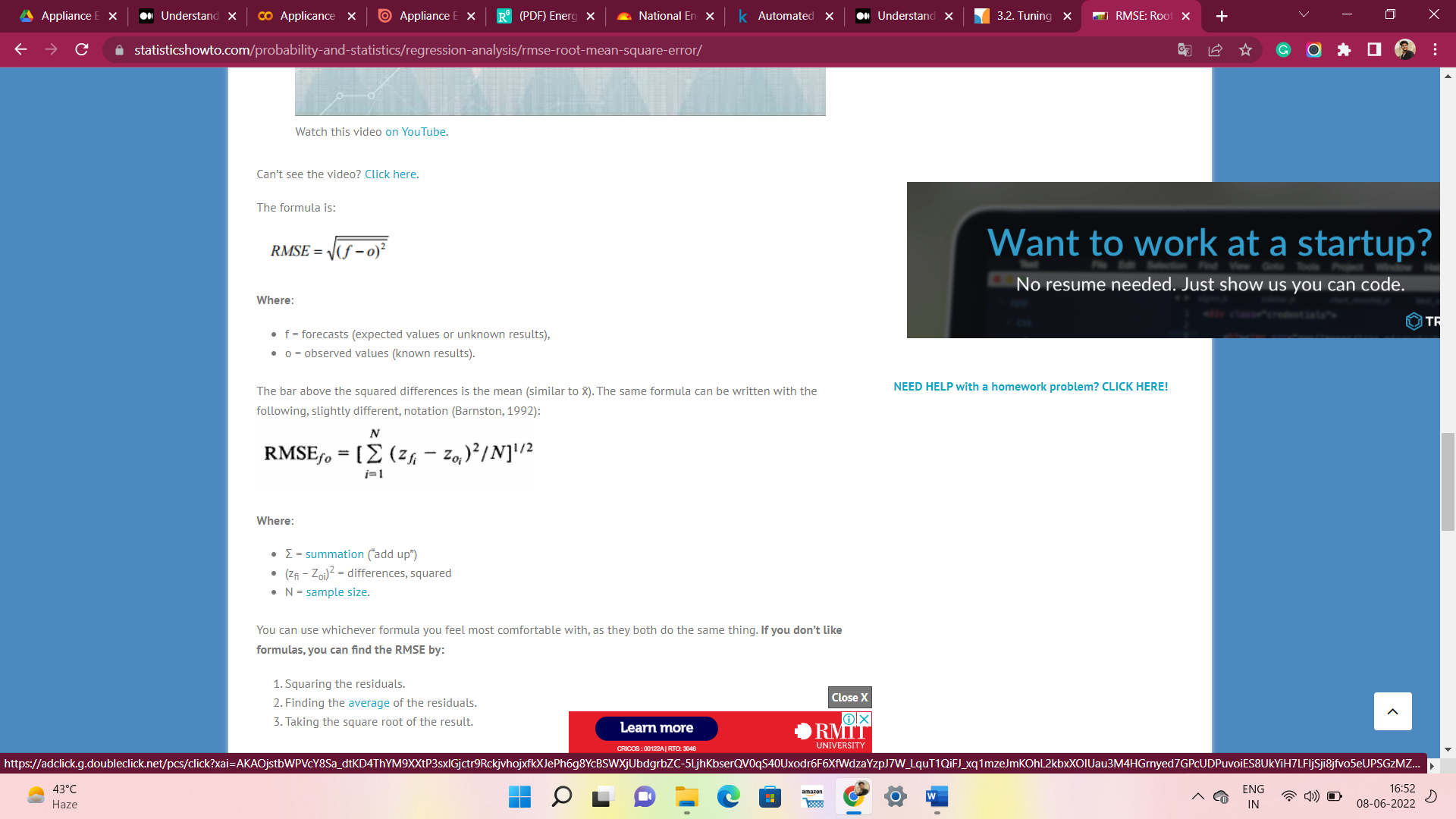
Model can be evaluated by various metrics such as:

**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.

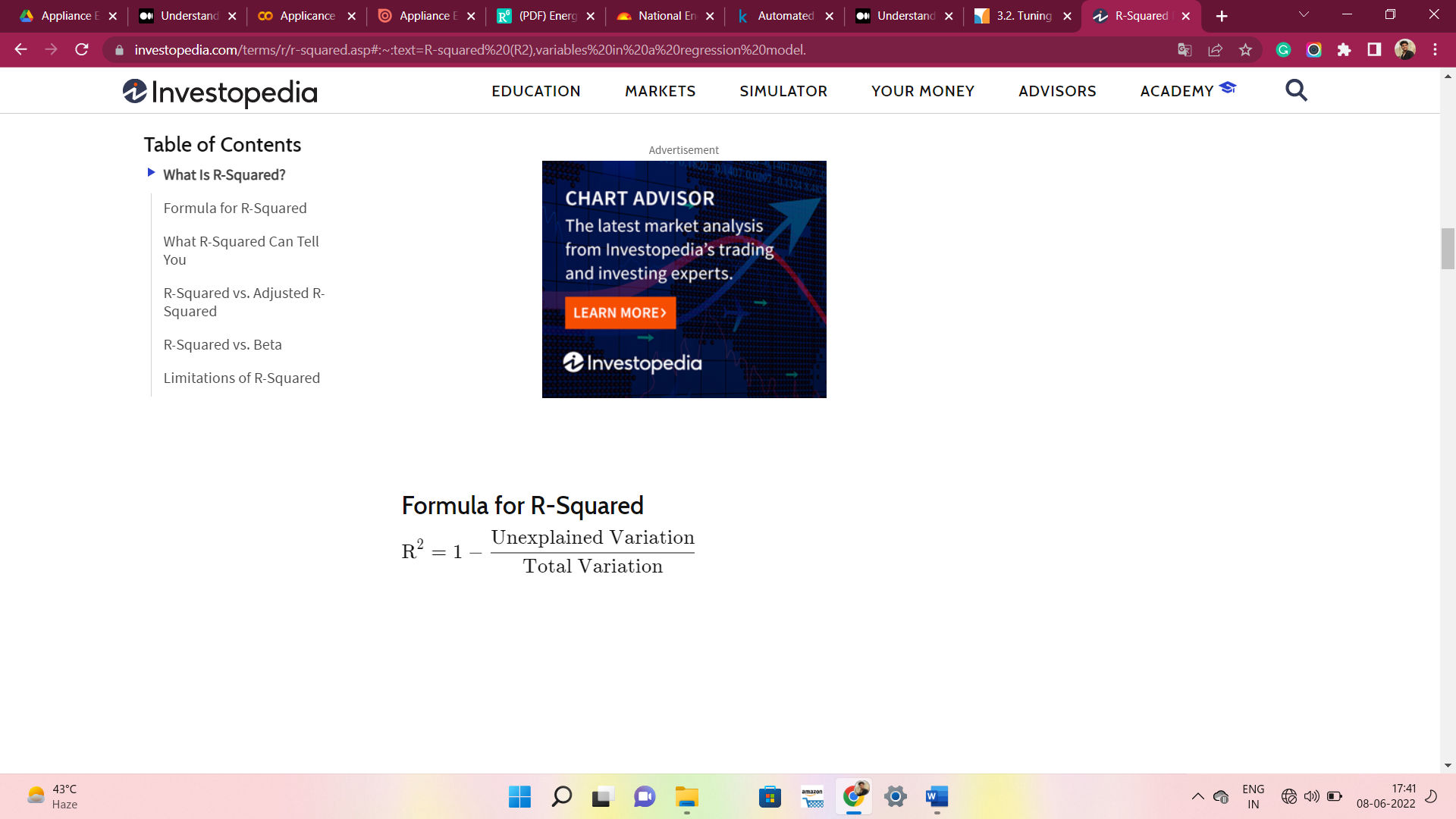
**2.Root Mean Squared Error (RMSE)**:

**Root Mean Square Error**(RMSE) is the Standard deviation of the residuals. 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



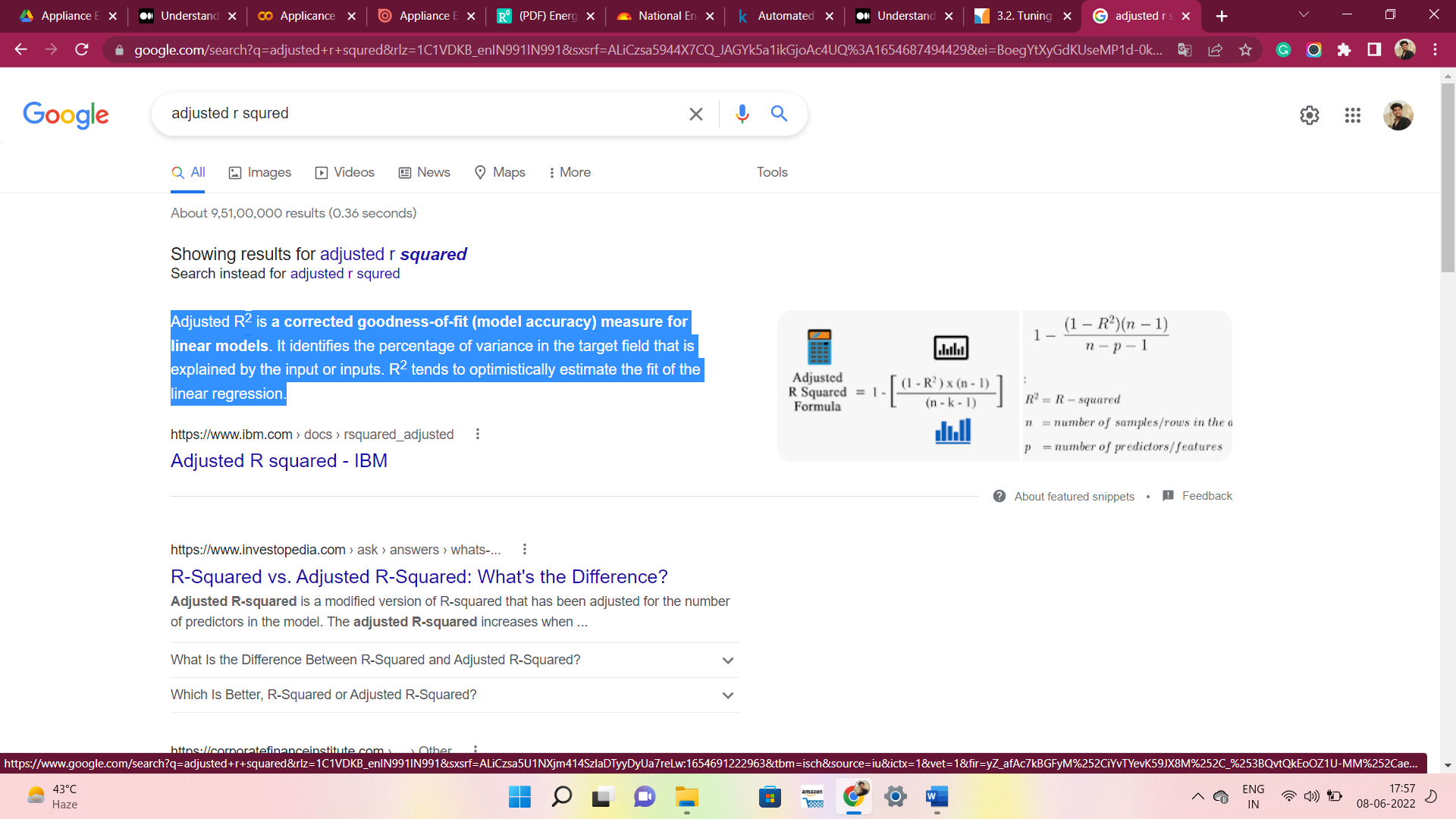
**3.R-squared**:

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.



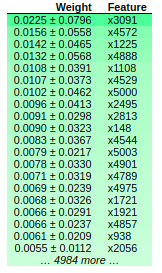
**4.Adjusted R-squared**:

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.



1. **Model Explainability:**

ELI5 is a python package that is used to inspect ML classifiers and explain their predictions. It is popularly used to debug algorithms such as sklearn regressors and classifiers, XGBoost, CatBoost, Keras.



1. **Conclusion:**

* Our main objective is to predict the Energy usage by the Appliance to achieve this we have Implemented the XGBoosting, Decision tree, Randomforest, Gradient Boosting, LinearRegression and Regularized Linear Regression algorithms was done along with cross validation and hyperparameter adjustment.
* In a comparison of all models, the RandomForest regressor is the best, having a high r2 score, a low MSE, and a low RMSE value (Comparing to other models). The model explainability Eli5 approach is used to determine which attributes are crucial for predicting output and understanding the model.

**Acknowledgement:**

This project was completed by Dipanshu Kumar. I am extremely grateful to the celebrated authors whose precious works have been consulted and referred in this project work. I also wish to convey my appreciation to our peers who provided encouragement and timely support in the hour of need.

**References:**

* Analyticsvidya.com
* Renewableenergyworld.com
* Mygreatlearning.com
* datapine.com
* Some ideas are also referred from Wikipedia.