Building Machine Learning Models

Before getting into this we need to first understand that why we need to build a model? Is there any future use or something like this…? Let’s understand

A machine-learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data. And in the future, we can use the same model to get the results and we can apply them to our business.

Let’s understand with an example. You have an old car shop and you have to sell the cars to maximum price and you don’t know the what should be cars price… That will attract the customer and also don’t put you the loss…

For this, you go to a Data Scientist and ask him to find your problem solutions. What they will do is…

He will first collect the old cars data and his price then he will clean the data pre-process the data generate the insights and then he will build the model for you and give it to you…. So in the future, if you want to predict a car price you simply have to enter details in the application like

1. Car model
2. Car make
3. How old is this
4. Build Year
5. Company name
6. Horse Power
7. Engine capacity
8. Color
9. Etc.

After entering the details if you get the car price that will not be too high or too low. So that’ the model. We build the model to predict things again and again. And we store this in a file that is known as a model

## So How to build a model?

Model building requests lots of steps to be performed carefully to get a good prediction. Building a model needs the following steps:

1. Collect Data
2. Prepare the data
3. Choose the model
4. Train your machine model
5. Evaluation
6. Parameter Tuning
7. Saving the Model
8. Prediction or Inference

So let get into all one by one to understand all the process

For the demonstration, we are gonna use the Big Data Mart Sales Problem

## Step 1: Collect Data

To solve the given problem you need to collect the required data that you will use to train your model. You will get the data in either Structured form or in Unstructured form depending on the organization. Those data you will receive could be collected by some volunteer by following methods

1. Interviews
2. Questionnaires and surveys
3. Observations
4. Documents and records
5. Focus groups
6. Oral histories

These data can be collected in any form of media storage like Databases or in text format and many more But you have to get data from

* Databases
* API’s
* Text Files
* Images
* etc

Data could be internal or external based on the organization if it will be the internal data you can easily access that data like employee data, sale data, Market data of the organization,

But also something you may need to fetch External data that is available over the internet. So you may scrap the external data using selenium or beautifulsoup in python.

Data can be numeric or categorical so you have to take care of that…

Numeric data includes

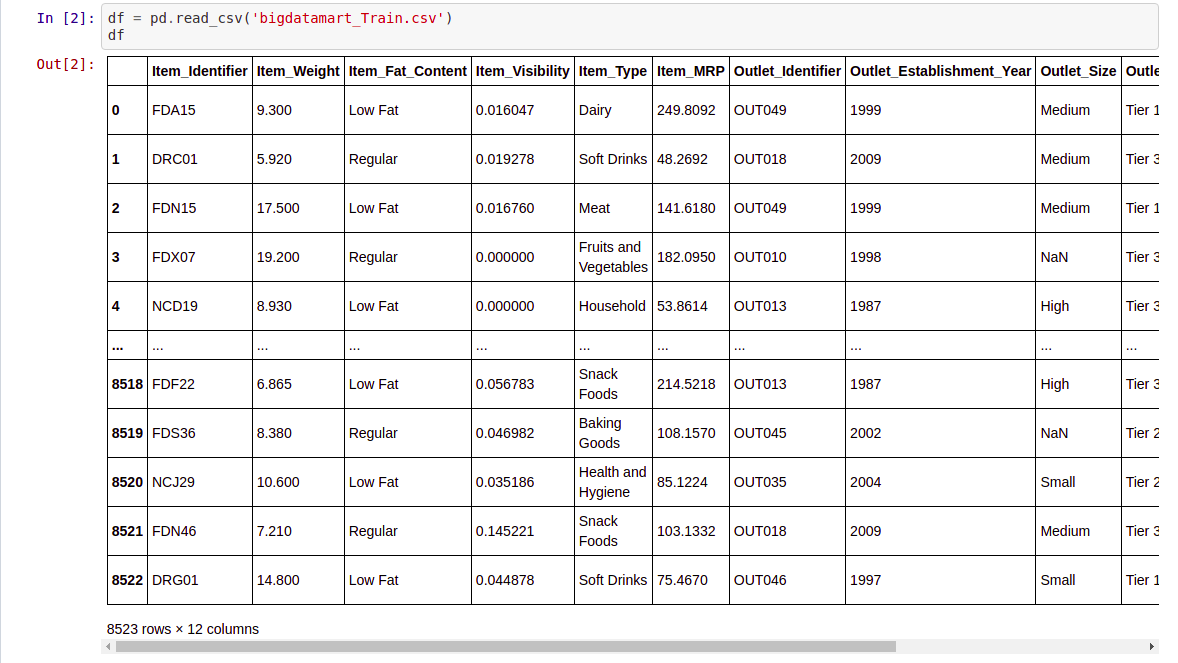
* Temperature
* Loan amount
* Customer retention rate
* Employe salary
* Car prices
* Etc.

And Categorical data may include

* Gender
* Color
* Highest degree earned
* Product is defective or not

After collecting the data we need to get into preparing the data

In Our case, we already have the data so let’s load the Data



## Step 2: Prepare the data

This step is connected with transforming the raw data that was collected into a form of data that can be used in modeling. Converting raw data into an appropriate form is known as Data Preparation.

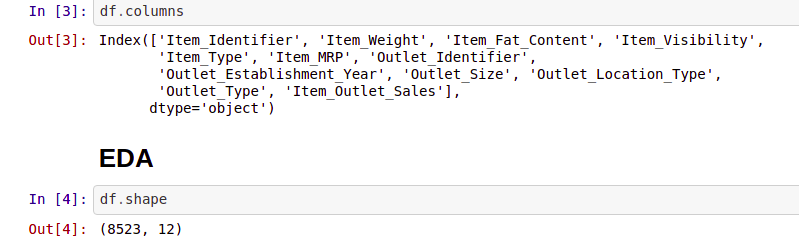
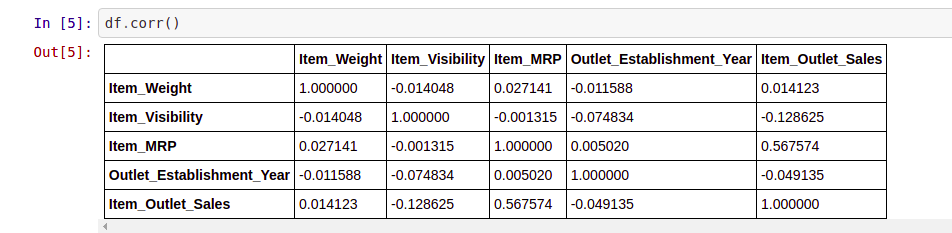
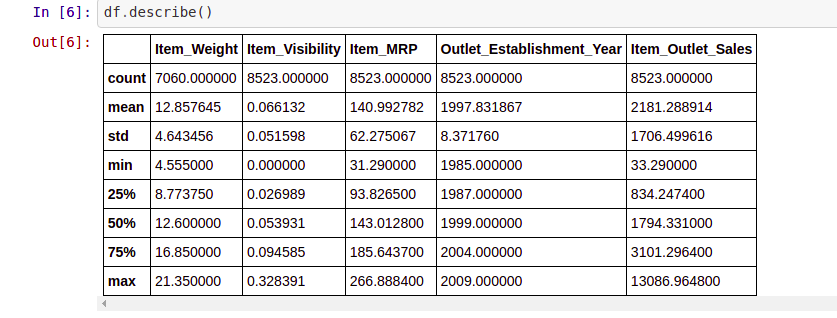
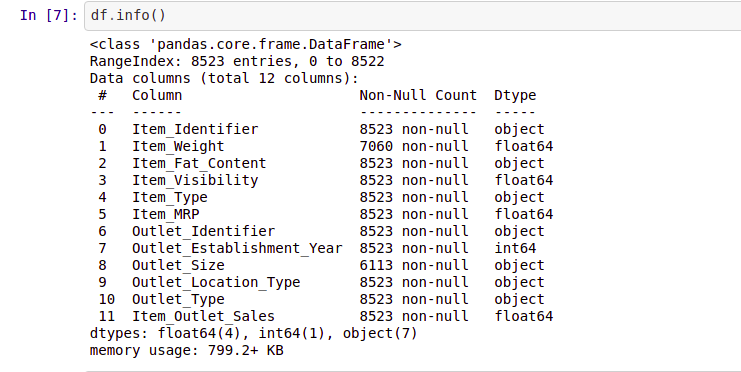
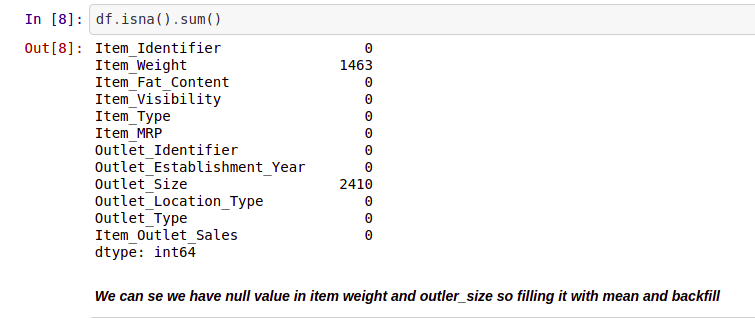
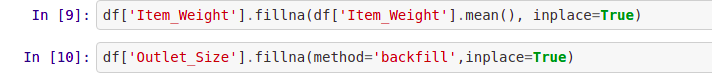
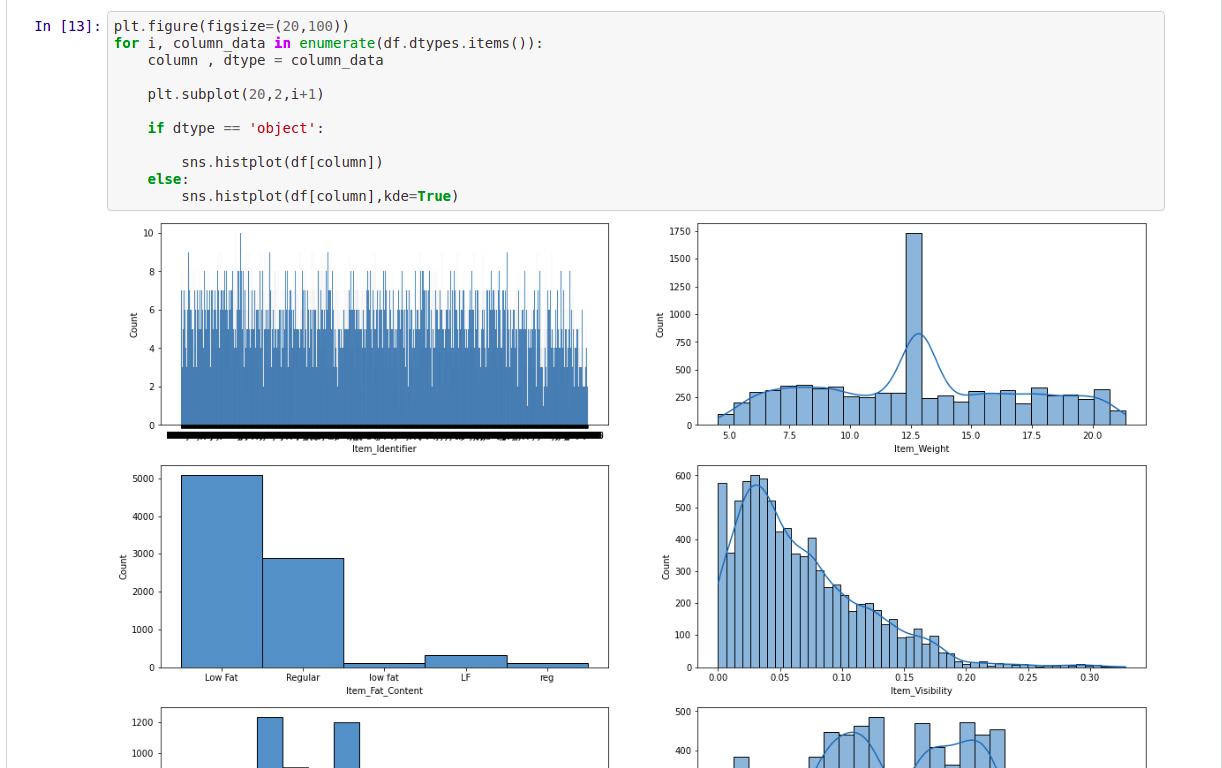
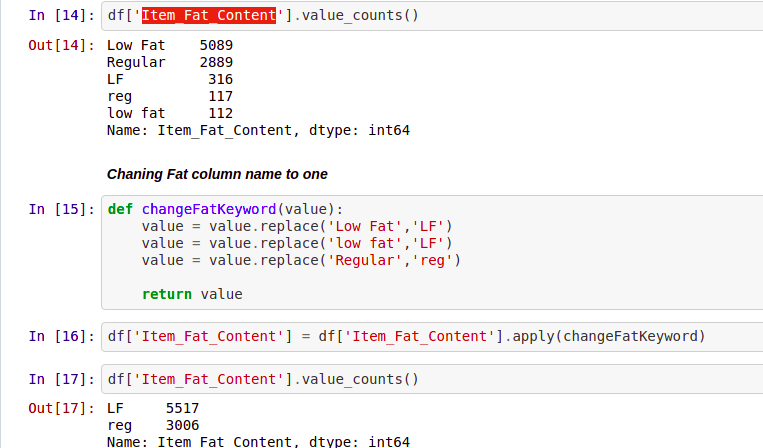
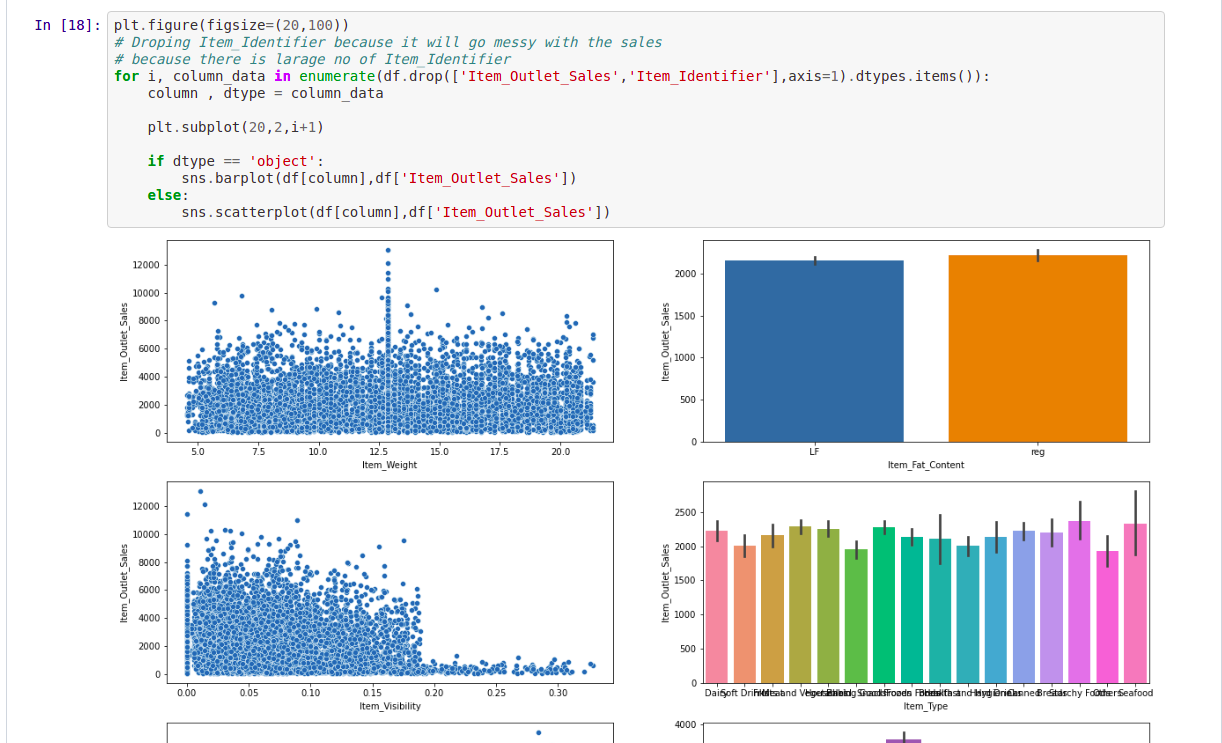
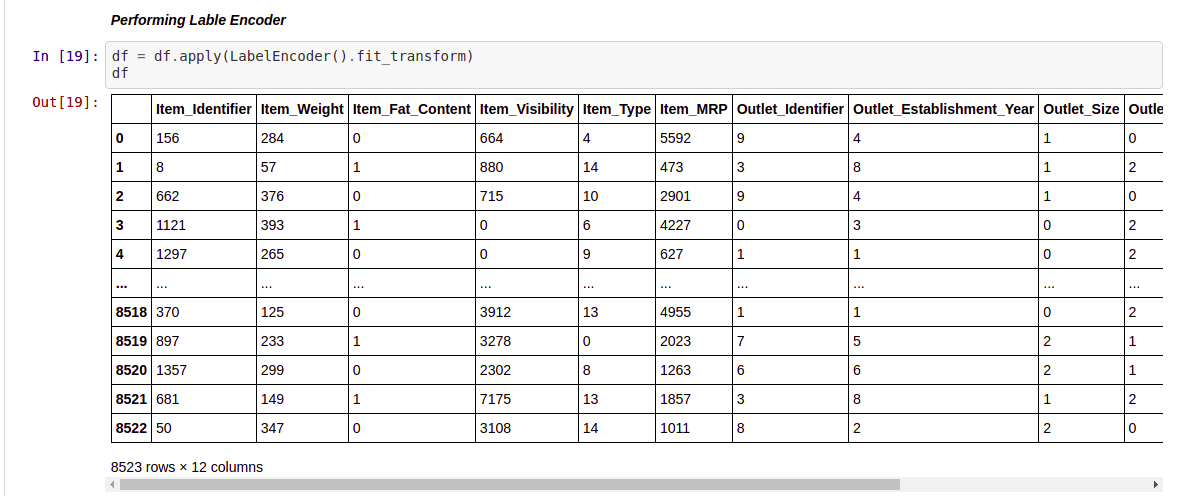
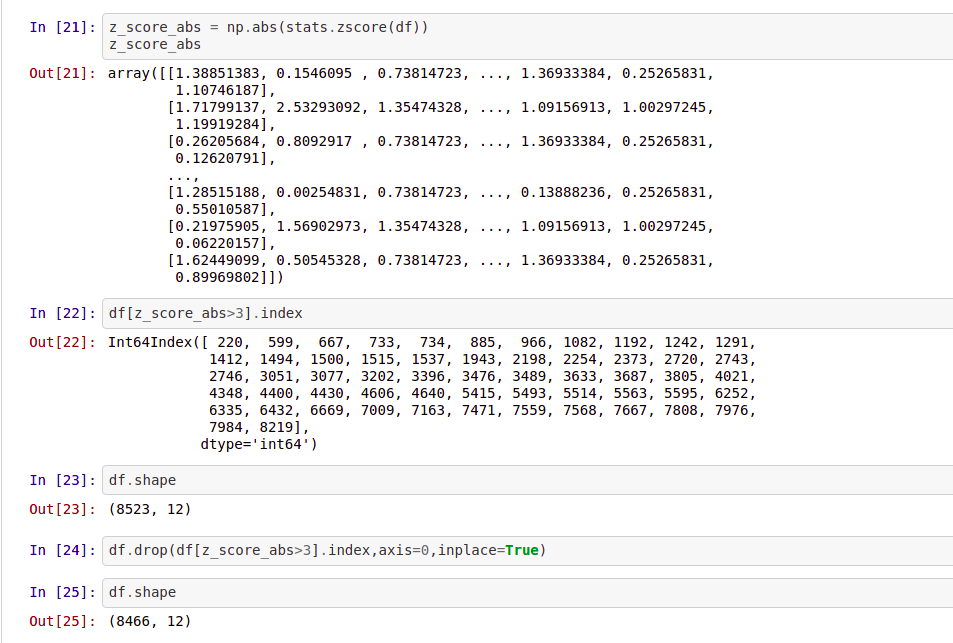
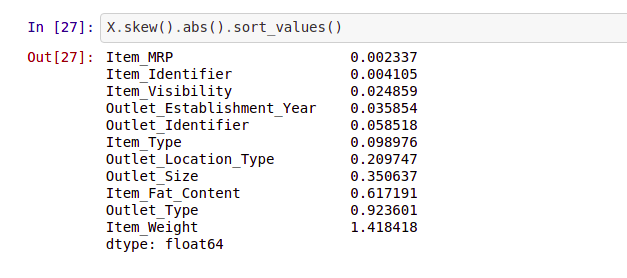
This step involves lots of smaller steps like

* Validating the data - Validity exists when we measure something like we have the length of the table and it never could be zero
* Questionnaire Checking - Parts of the answer to a question may include the wrong or inadequate answers like in the question of “Do you own a car?” People may answer like “Yes”, “No”, “I have”, “I don’t” and many more could be different answers possible.
* Edit acceptable questionnaire- Editing the answer could be done by applying some function and converting “I have” to “Yes” and “I don’t” to “No”
* Handling missing values - Handling missing values could be done in many ways like if the data set is large and there are very few missing rows we can remove the row. If the column contains 70% or more null values we can drop the whole column from the data
* Statistically adjust the data- This step requires normalizing the data transforming the data so it will scale down the data in normal form
* Store the data set for analysis - After performing all the above steps we can store the data for analysis.
* Analyze data - In this phase we analyze the data to understand the data better and to find insight from it. Basically, we perform 2 types of analysis on the data
  + Descriptive Analysis - It refers to the description of the data for a particular sample. It summarises and describes it. In Descriptive analysis we use

1. Frequency Distribution
2. The measure of Central Tendency (Mean, Median, and Mode)
3. The measure of Variability( Range, Standard Deviation)
4. Bivariate Descriptive analysis (Contingency tables and correlation)
   * Inferential analysis - It is used to test the significant relationship among the variable or to find the statistical support for the hypothesis. We may perform the following test in that
5. T-test
6. Analysis of variance (ANOVA)
7. Chi-square Test
8. Hypothesis Tes

After preparing the data we need to choose the model.

Here we needed to perform many things lets see one by one

1. Checking Shape and columns
2. Checking Correlation
3. Describing the Data
4. Get info of the dataset
5. Checking for null values 
6. Filling null values
7. Doing Univariate analysis
8. Removing duplicate keyword for Column “Item\_Fat\_Content”
9. Doing Bivariate analysis
10. Performing label encoder 
11. Checking for outliers and dropping it
12. Checking for Co-linearity
13. Checking for skewness

## Step 3: Choose the model

There are several models that you can choose according to the objective. There is a set of parameters that need to check before selecting the model. You have checked the following parameter:

* The accuracy of the model.
* The interpretability of the model.
* The complexity of the model.
* The scalability of the model.
* How long does it take to build, train, and test the model?
* How long does it take to make predictions using the model?
* Does the model meet the business goal?

So these sorts of questions you have to ask before you choose the model.

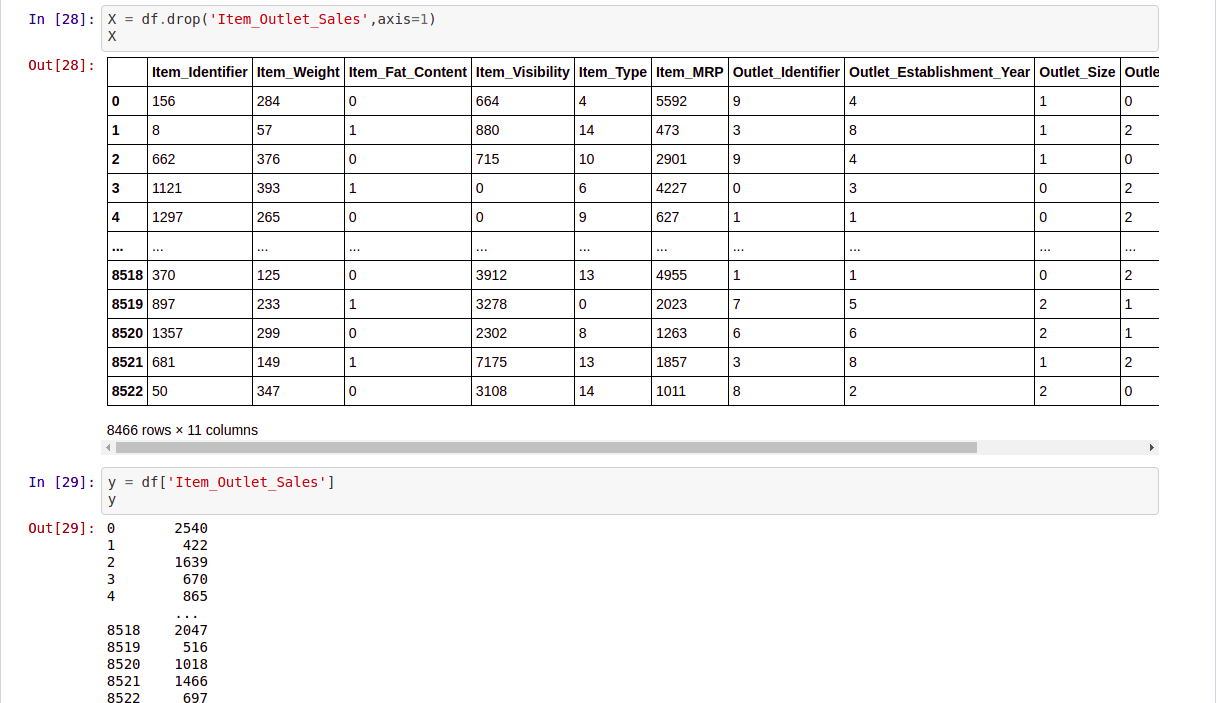
For our problem, we don’t have a specific goal so we are choosing all the popular models which could be used in Liner Regression as our target variable is continuous in nature

## Step 4: Split The Data

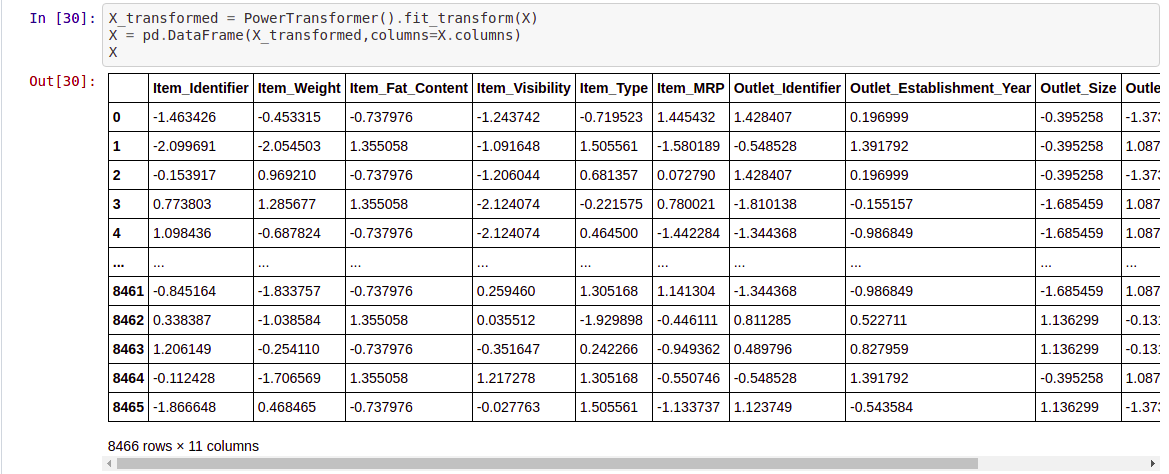
The train test split is used to estimate the performance of the model when they are used to make predictions on data that is not used for training it. But something it is advisable not to perform a train test split like when the dataset is too small.

This process involves taking the data and diving into the 2 subsets one is for training the data is called trained dataset and the second is used for testing the data is known as the testing dataset for the testing we feed features column into the model and predict the target column and then we compare the predicted value with the actual value we have in the dataset.

The main objective is to estimate the performance of the model on new data/ un-seen data (Data that is not used for training the model)

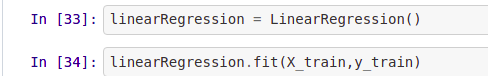
Splitting the data into training and test data 

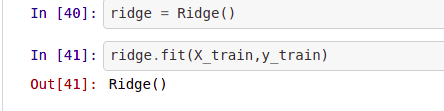
We are also performing the Power Transformation to transform the data

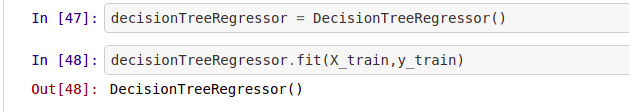
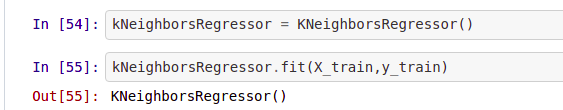


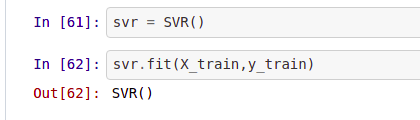
## Step 5: Train your machine model

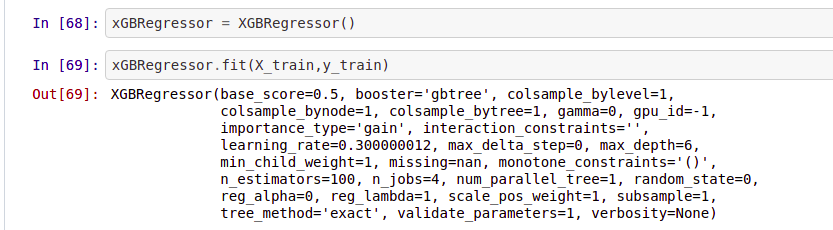
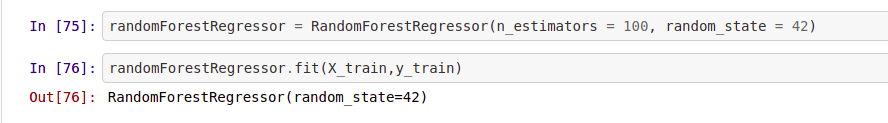
The training model is the process of feeding data into the ML algorithm so that the ML algorithm can learn from the past data and predict the data.

1. Linear Regression
2. Ridge



1. Decision Tree Regressor 
2. K-Neighbors Regressor 
3. SVR



1. XGB Regressor 
2. Random Forest Regressor

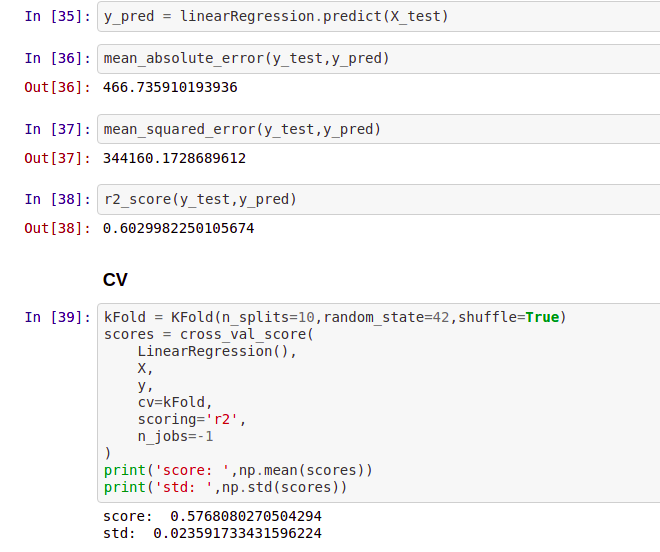
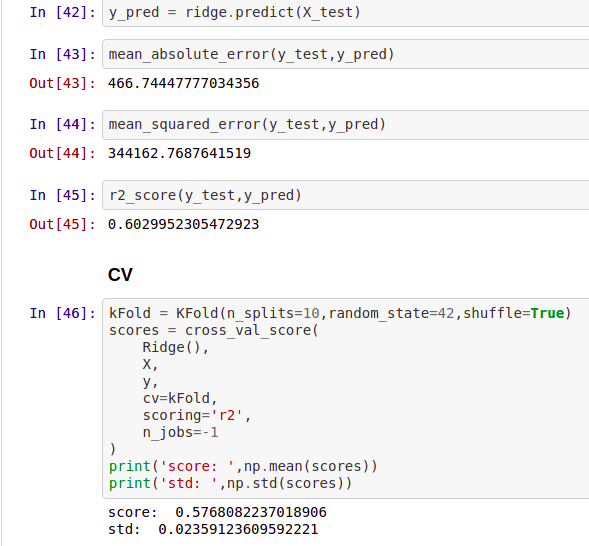
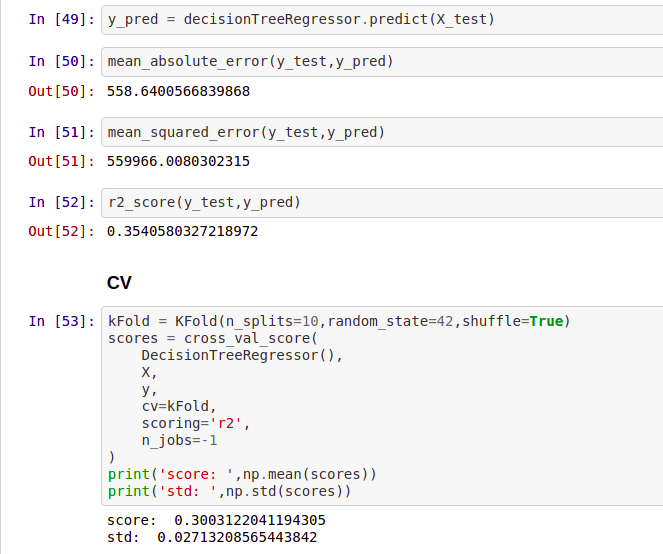
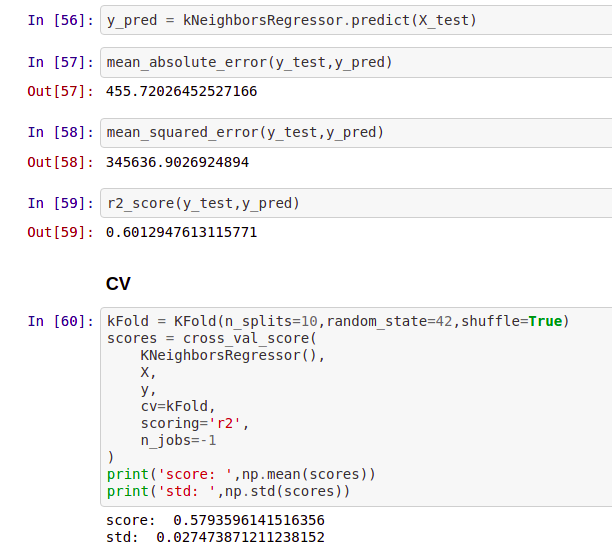
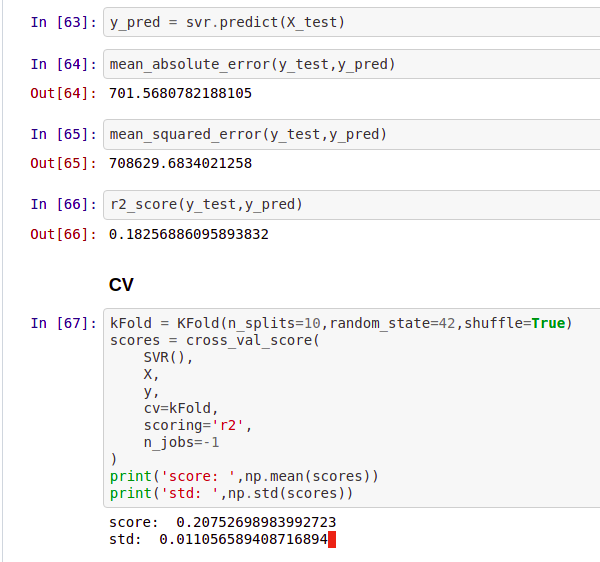
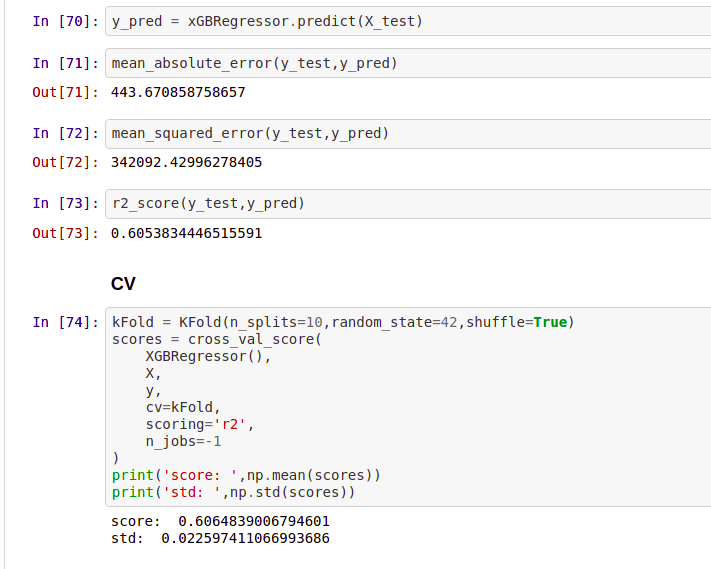
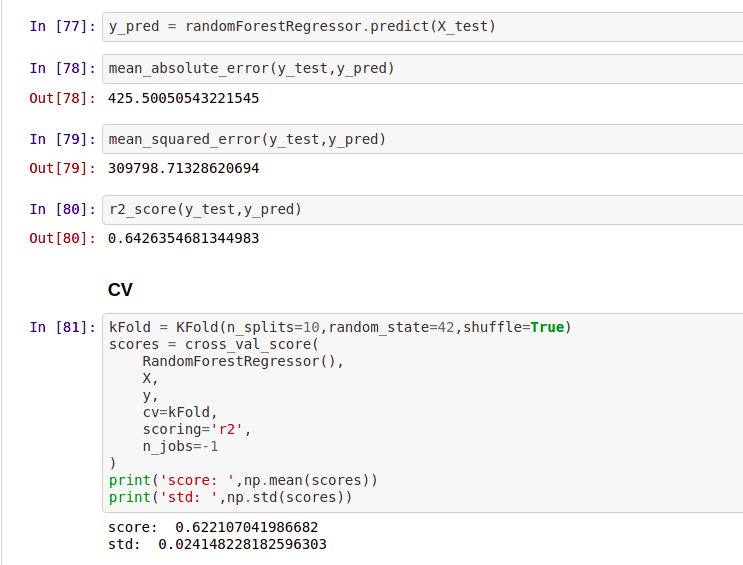
## Step 6: Test Model with Test data and Doing CV

It’s time to test the model with test data to check the model performance and also perform the CV to check the model efficiency

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample

Cross-validation is primarily used in applied machine learning models to estimate the skill of machine learning models on unseen data

Cross-validation steps: -

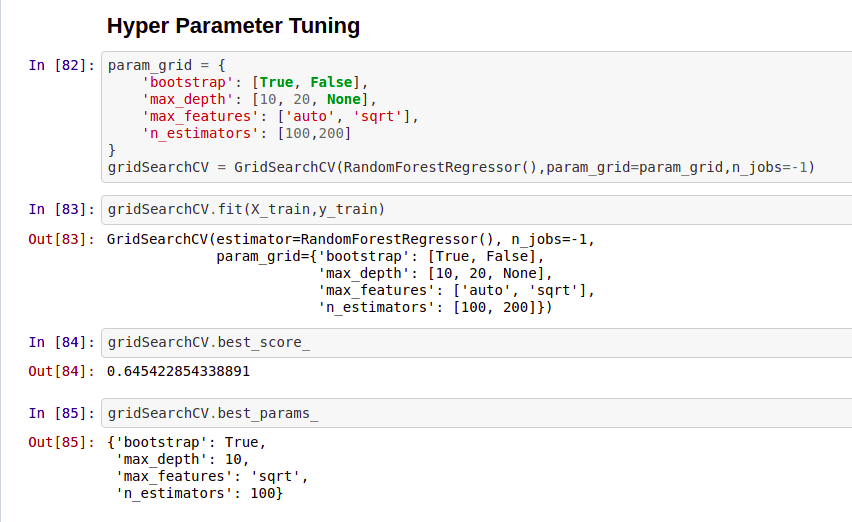
1. Shuffle the Data
2. Split the Data into K groups
3. For each group
   1. Take one group as test data
   2. Take remaining group as Train Data
   3. Fit the data on training data and test it on test data
   4. Store the score of the model and do it for other groups
4. Summarise the skill of the model using all the stored scores of the model
5. Linear Regression
6. Ridge
7. Decision Tree Regressor 
8. K-Neighbors Regressor 
9. SVR
10. XGB Regressor 
11. Random Forest Regressor

## Step 7: Parameter Tuning

Parameter tunning is the last part of improving the accuracy we give a set of parameters to model and check for the best parameter for the model. After checking the accuracy and CV score we got we can easily find the best model and then after we can perform parameter tunning on that particular model to improve more accuracy

Hyperparameter tuning is used by the machine learning model when it’s learning but they are not part of the resulting model. At the end of the learning process, we have the trained model parameter which effective is what we refer to as the model. The Hyperparameter that was used during training is not part of the model.

Anything in machine learning that you decide their values or choose their configuration those values remain same when training ends is known as a hyperparameter



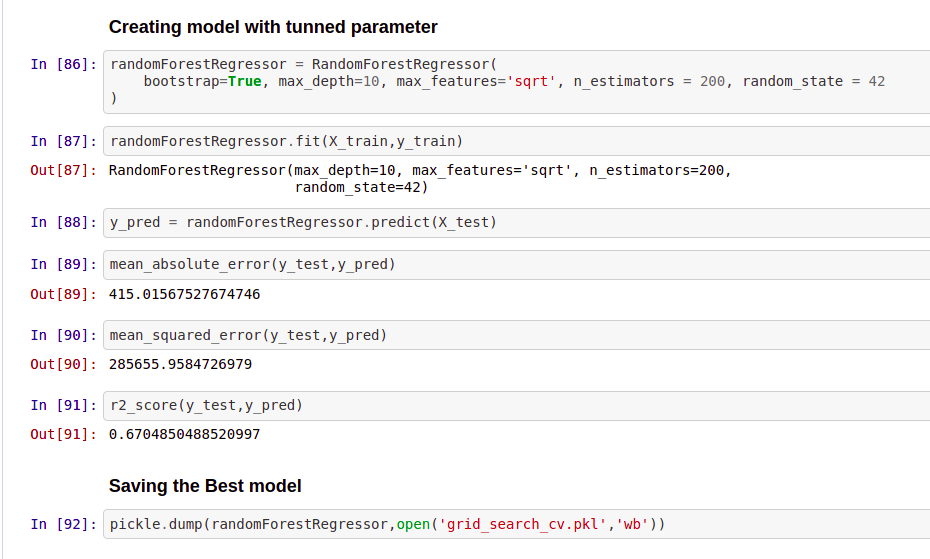
## Step 8: Saving the Model

After doing everything we need to save the model for future use we can save the model with a pickle. We are creating the bested tunned model and saving it with the pickle.

In machine learning, while working with models we need to save the trained model in a file and restore them in order to reuse it to compare the model with other models, to test on new data, or for future use. The saving of data is called serialization while restoring the data back is called Deserialization.

We have 2 very popular ways to perform the saving model is

1. Pickle: The pickle model implements a fundamental but powerful algorithm for serializing and deserializing a Python object structure. It basically has 2 function pickles. dump and pickle. load to load and dump the model
2. Joblib: Joblib is a replacement for pickles as it is more efficient on objects that carry large NumPy arrays. These functions also accept file-like objects instead of filenames.



## Step 9: Deploying the Model

There are different approaches to putting models into the production, with benefits that can vary dependent on the specific use case.

There are 2 ways we deploy the model is

1. Train: one-off batch and real-time :

Models don’t need to be continuously trained in order to be pushed to production. Quite often a model can be just trained ad-hoc by the data scientist and pushed to production until its performance deteriorates enough that they are called upon to refresh it

1. Serve: Batch Realtime:

While not fully necessary to implement a model in production batch training allows having a constantly refreshed version of your model based on the latest train

Successful ML deployment generally takes advantage of this principle which builds the following pillars.

1. Tracking: Keeping track of the software artifacts
2. Automation: For an ML application the pipeline should not only automate the training model but also automate model retraining.
3. Monitoring: ML application should also be monitored for invalid prediction or data drift which may require the model to be re-trained
4. Reliability: It’s important that ML applications function as expected and are resilient to failure. Getting reliability right for ML requires some special consideration around security and testing.