Introduction

The sale is a lifeline of every business sales prediction that significantly impacts companies. Accurate predictions benefit the organization to maintain the standard and increase the company's lifestyle by using different strategies. Typically, a prediction is based on the knowledge of previous studies with a deep focus on the conditions and then applies various factors, including customer's taste, culture, marketplace, and many more. In short, we can say that our prediction depends upon the previous study results. Every business needs to be good in profit, and profit does not mean that stock sales are at maximum but also avoid the extra stock. Every retailer must maintain the stock according to the requirements and check the flaws and drawbacks that lead the sales down.

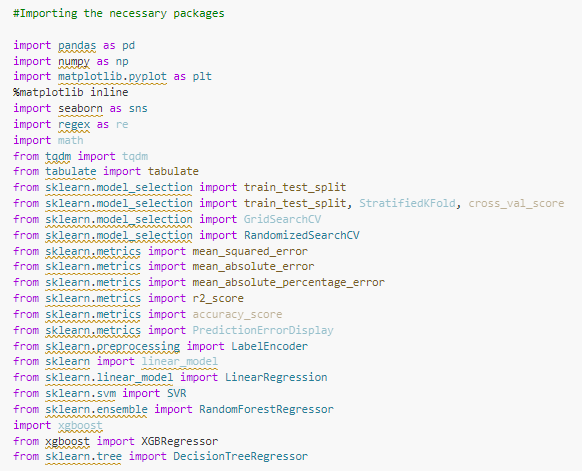
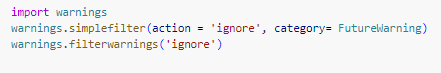
Problem Statement

The idea of the project is to extract information predict the sales to get an idea regarding the products.

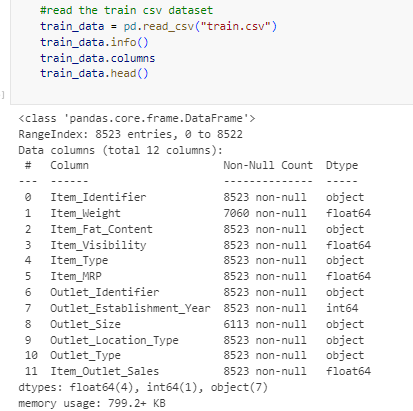
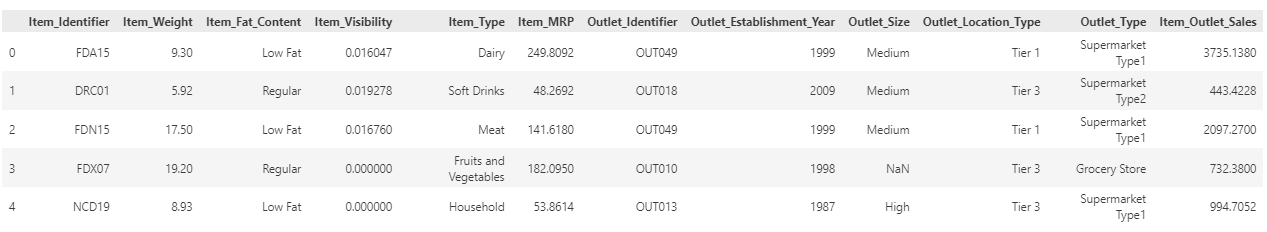
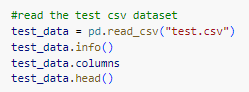
Analysis of Data

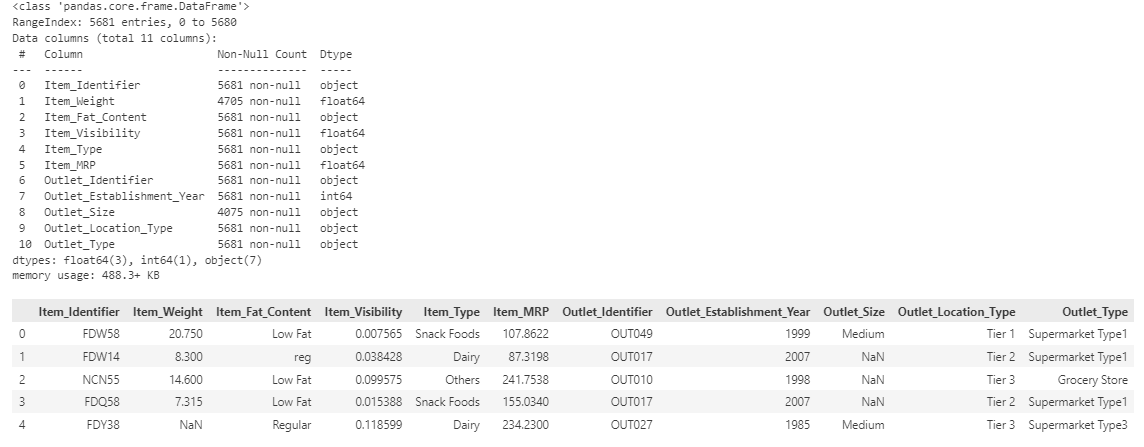
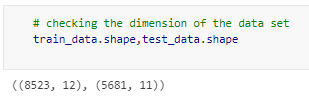
Before studying the data necessary packages of Python required for analysis is being called as follows:



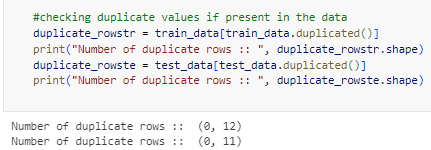
Explanation of Data:

The following step is to extract the data in Python as follows:   

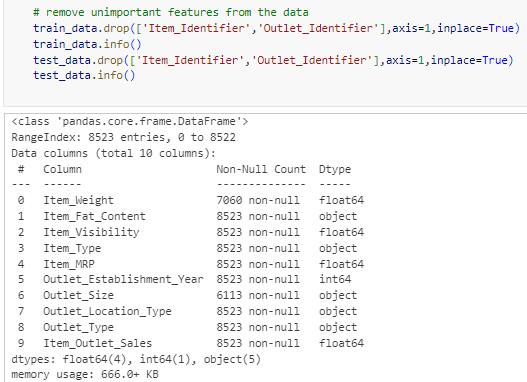
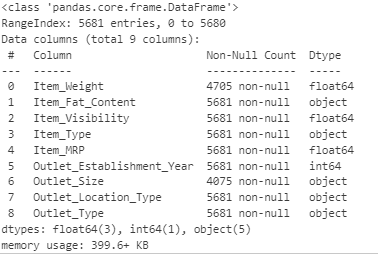
The test and train dataset is provided. The train data contains a combinations of 11 quantitative and qualitative features (explanatory variables) along with the quantitative predictor variable (Item\_Outlet\_Sales). While the test data contain only explanatory variables. Also, both the test and train dataset contain missing values.

Our object is to predict the sales of the product using the provided features such as item weight, item type, item MRP etc. However, before analyzing the data, the data needs to be *studied*, *cleaned* and *preprocessed*.

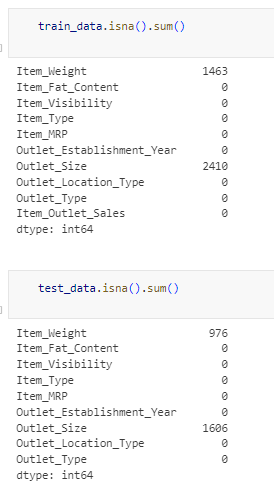
At first it needs to be find that whether the data contains any duplicate values or not. If present, it needs to be eliminated. 

As there is no such duplicates, we don’t need to take any action.

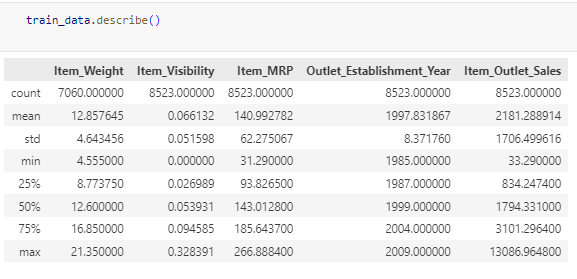
The column “Item\_Identifier” and “Outlet\_Identifier” are removed from both the train and test data, as it is not important. This is because, both are unique ids which are unique for each product. So, it will be of no help in predicting the loan status of a customer.

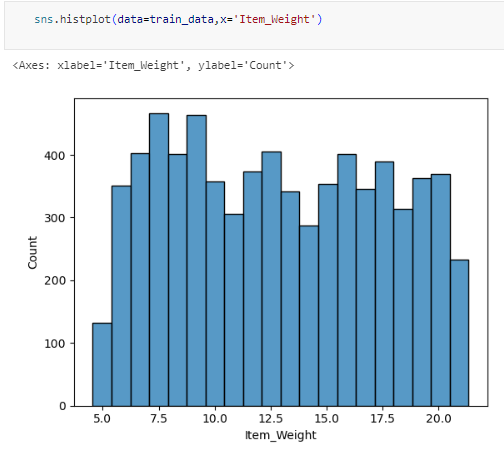
 

The number of null values are counted as follows:



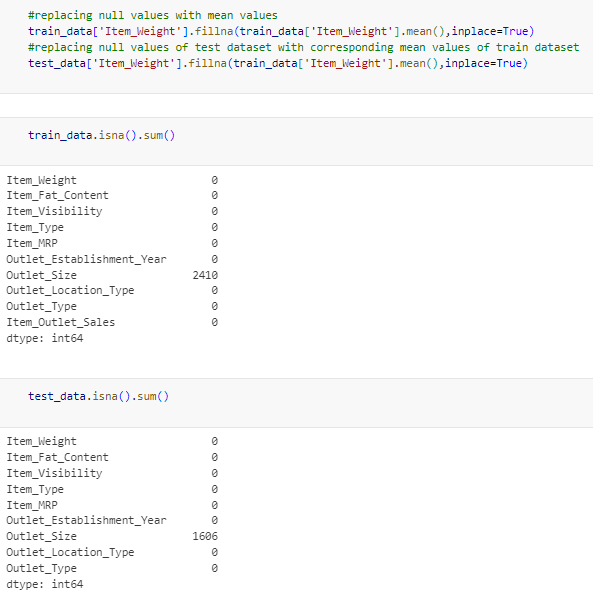
The summary statistics to study the distribution of the numerical data is obtained as follows:



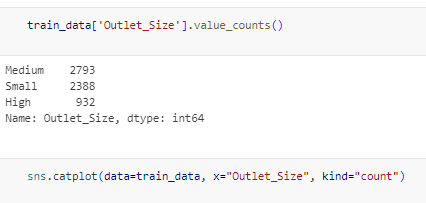
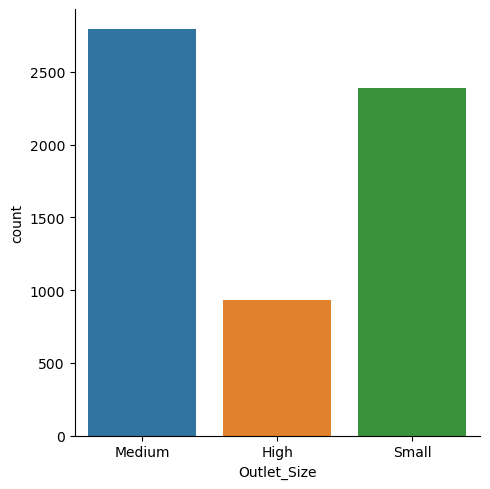
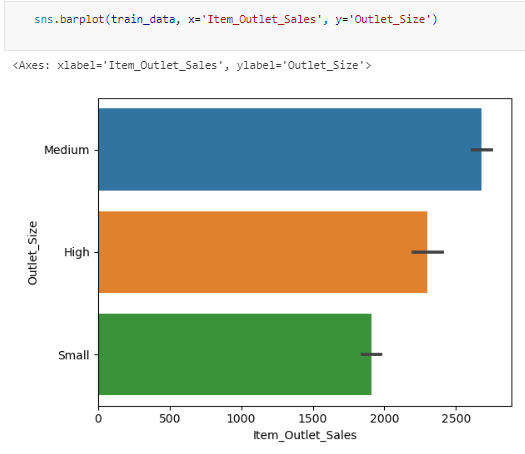
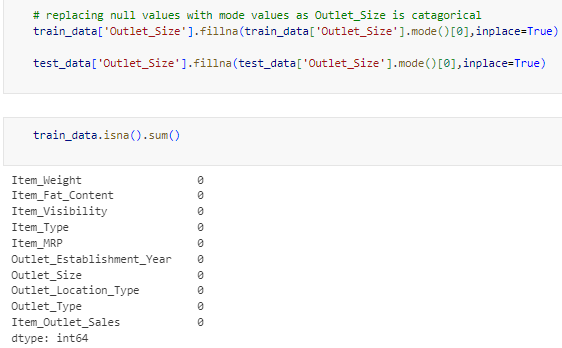
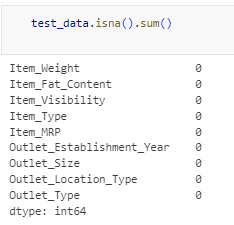
One can also visualize the distribution of the item weight using histogram as follows: 

From the above plot it can said that the distributions does not have outliers. Thus, mean can be used for replacing the null values in train data.

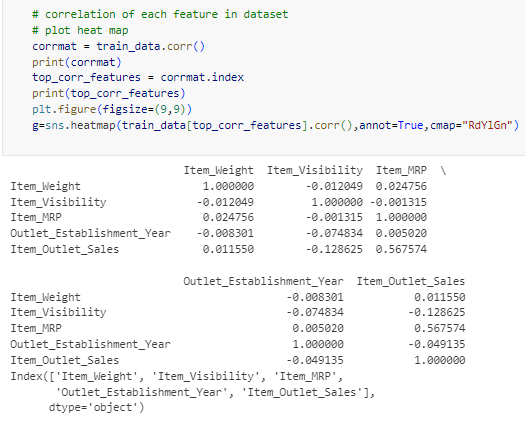
Note that, the same should be applied in test dataset to avoid baisness.

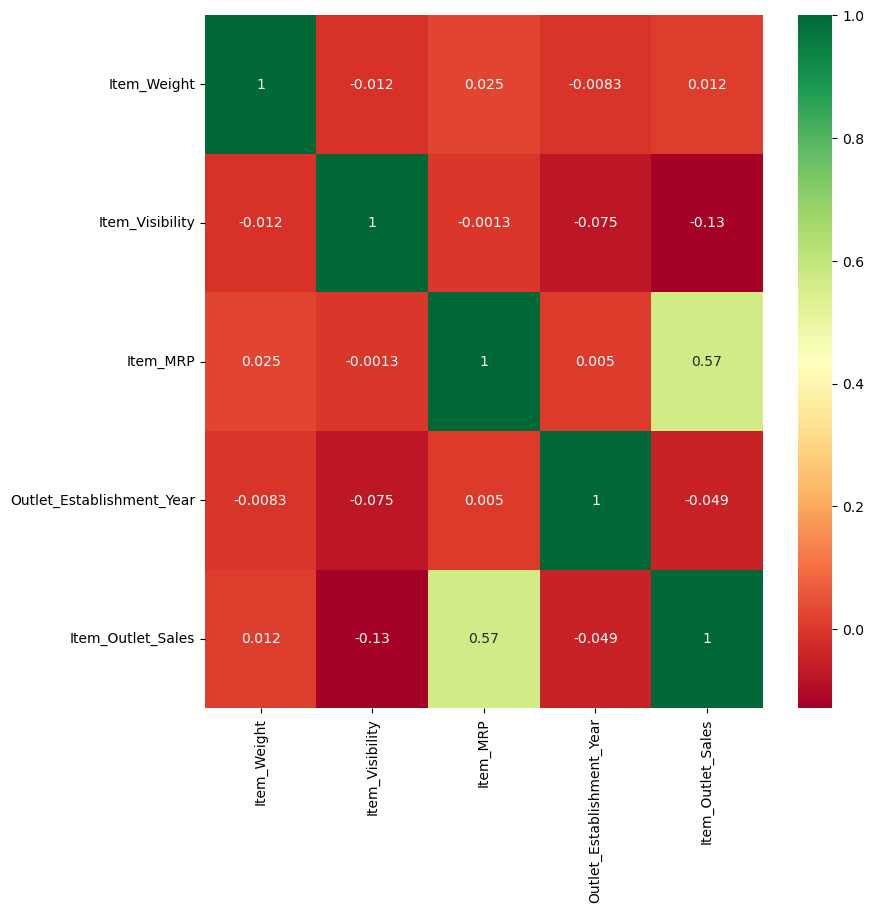


Now, only column “Outlet\_Size” has missing values. As it is categorical, we need to use mode to replace the null values. Thus, first the mode is observed and then it is replaced as follows:

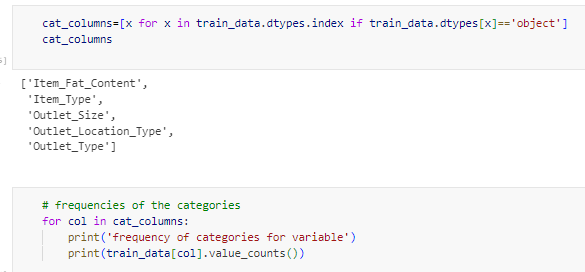
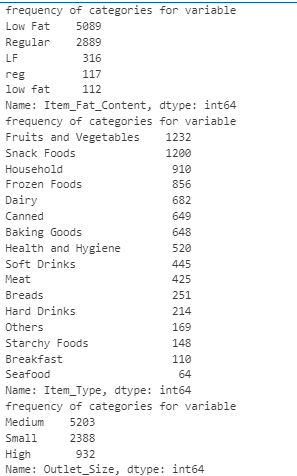
    

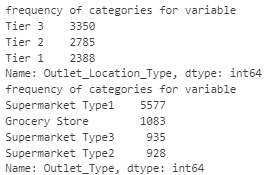
To check the correlation between different explanatory variables with predictor variable, one can use correlation matrix and heat map to visualize it.

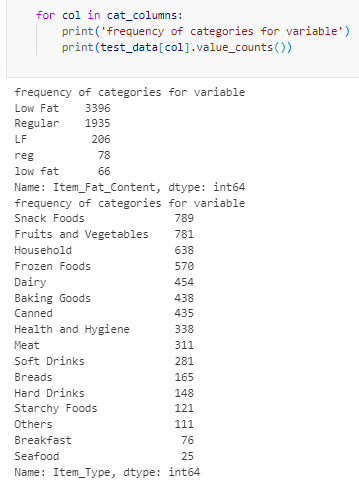


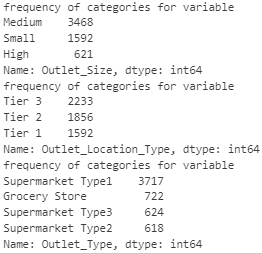


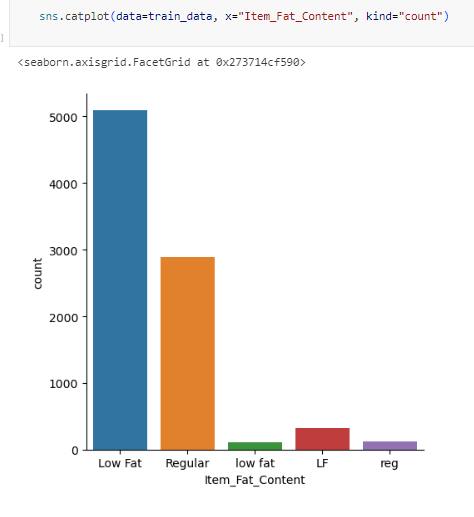
The frequency distribution of each qualitative features is as follows:

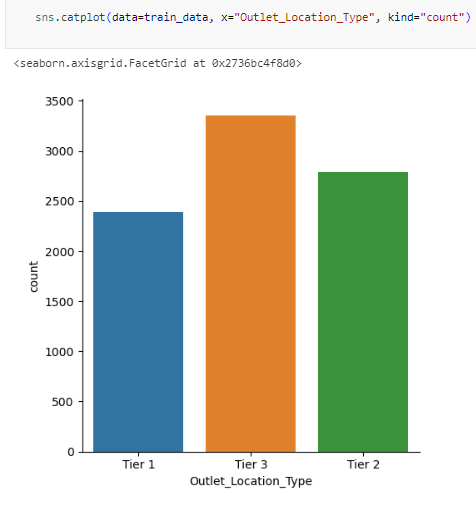
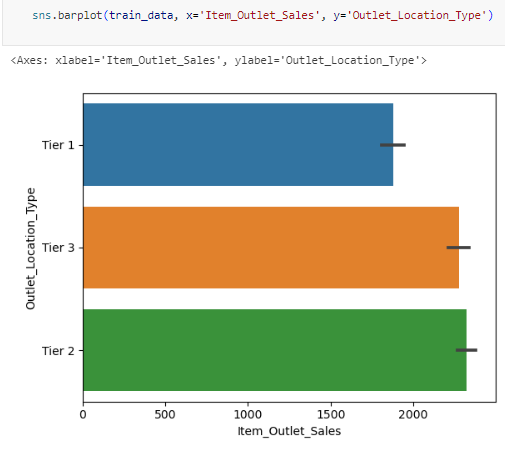
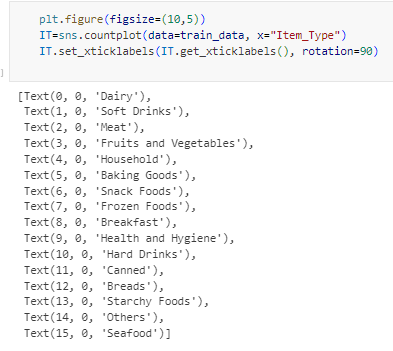
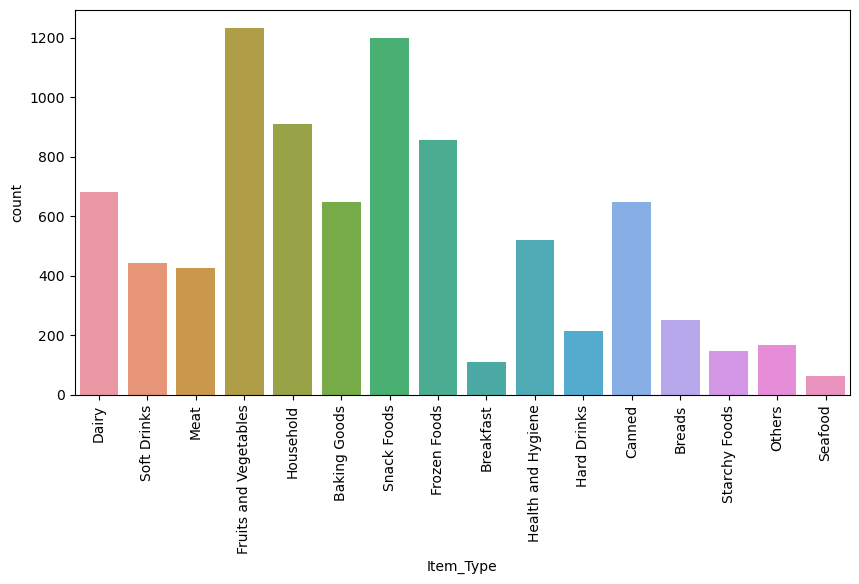
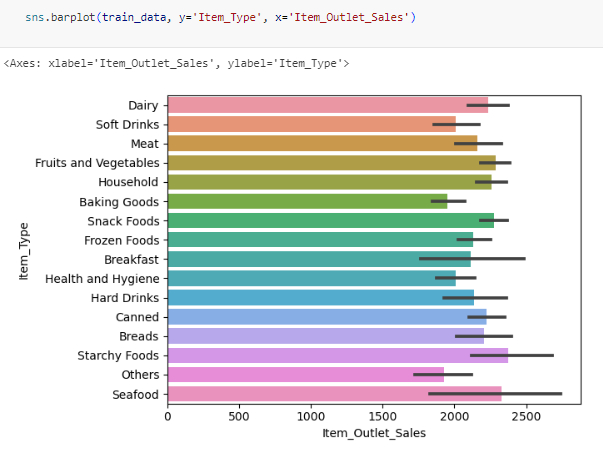
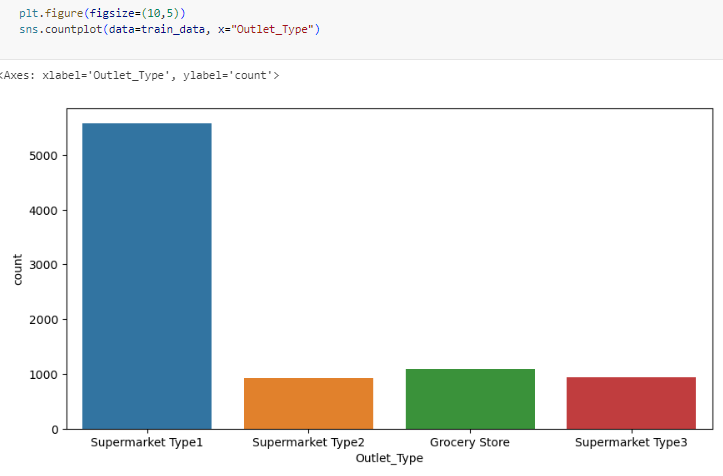
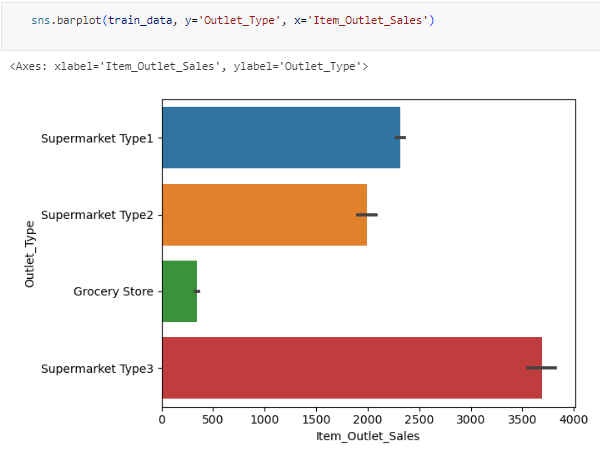




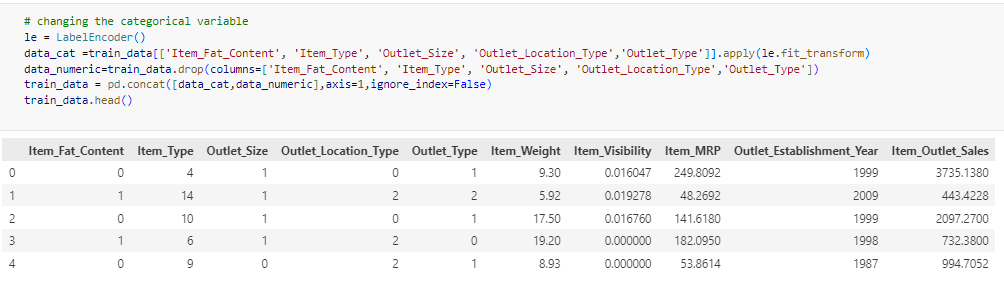


The item fat content is visualized as follows: 

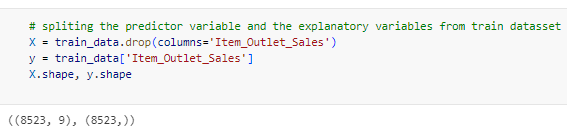
As it can be seen that, in item fat content the “low fat” and “regular” category is written is different pattern for those are taken as different category. Thus, it is being corrected both in train and test dataset as follows: 

Now, the other qualitative features with respcet to frequency (count) and item outlet sales are also visualized as follows:       

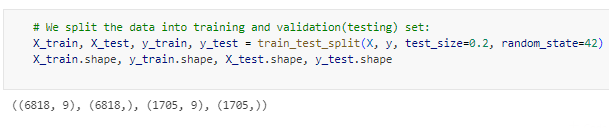
As it is not possible to carry any analysis using attributes. The attributes needs to be converted to quantitative variables (encoding) using “LabelEncoder” as follows:



Now, the train data is splitted into explanatory variables and predictor variables, so that we can do the analysis further.



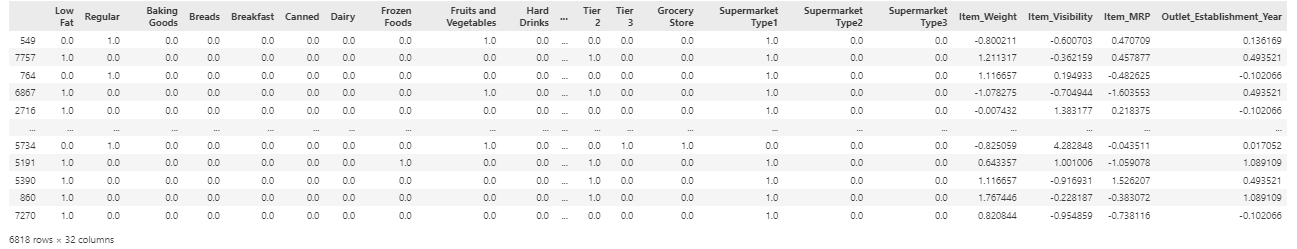
Then, the train data is spitted into 2 datasets. One will train the dataset to make models and another will validate the train dataset for the models.



Now, the categorical variables in both the train and test dataset are converted into array, so that the models can be applied on it.



The numerical variables or features are standardized as different variables are in different units of measurements. So it is important to make the numerical variables uniform. After that, the new changed categorical variables and numerical variables are merged to form the new processed data which is ready for analysis.



Data Analysis

Here, the predictor variable (Loan Status) is provided, thus it is Supervised Learning.

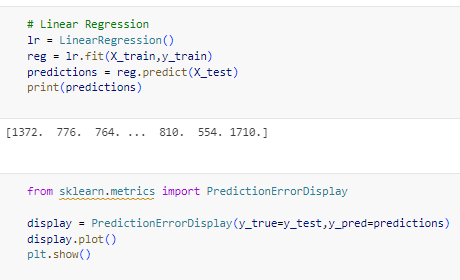
And as the predictor variable is a quantitative variable, so it is a Regression Problem.

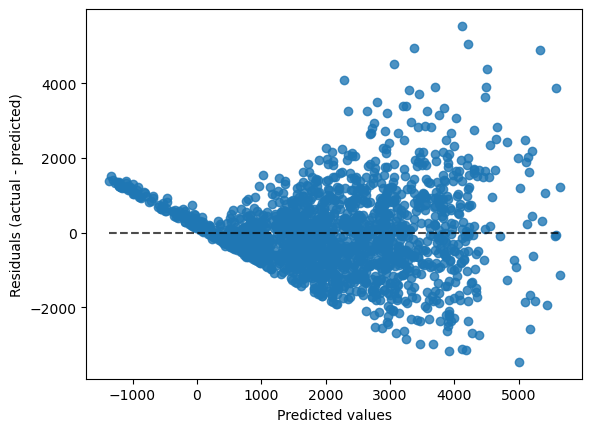
So different regression algorithm of supervised learning will be used.

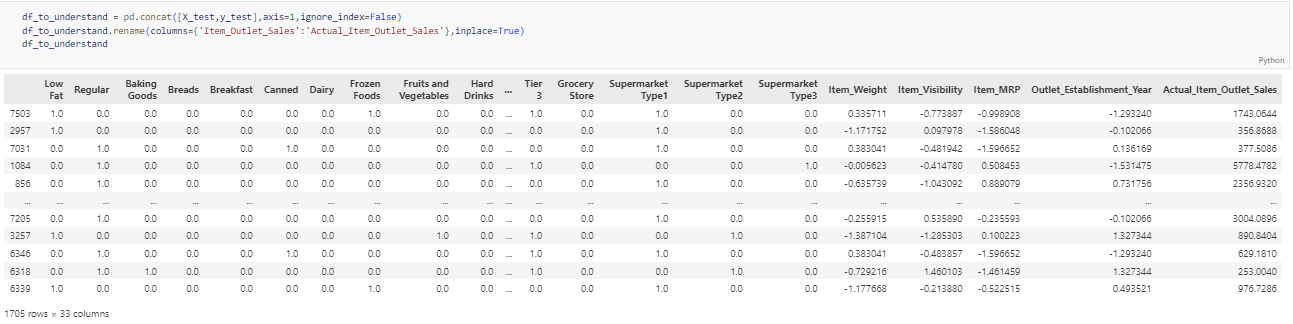
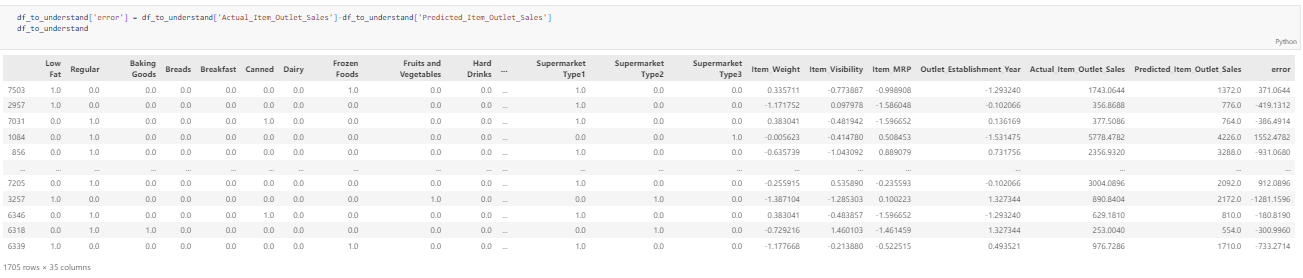
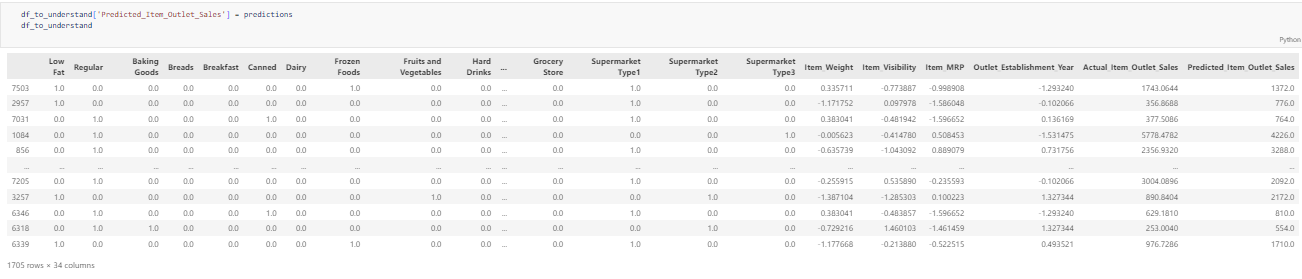
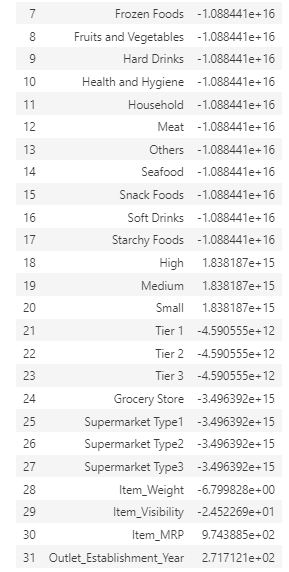
Regression:

1. Linear Regression:

It is the supervised learning technique. It is based on the relationship between independent variable (x) and dependent variable (y) and which is shown by equation of line which is linear in nature that why this approach is called linear regression. It gives a relation equation to predict a dependent variable value y (output) based on an independent variable value x (input).





1. Ridge Regression:

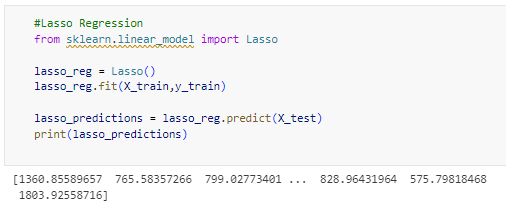
[Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are correlated.](https://www.bing.com/ck/a?!&&p=40dc603dff2f32bbJmltdHM9MTY5NDMwNDAwMCZpZ3VpZD0wY2UxNTlhOS1hOTU1LTYzMjAtMGFlOC00YmU5YTg4NzYyODUmaW5zaWQ9NTgyOA&ptn=3&hsh=3&fclid=0ce159a9-a955-6320-0ae8-4be9a8876285&psq=Ridge+Regression%3a&u=a1aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvUmlkZ2VfcmVncmVzc2lvbg&ntb=1" \o "en.wikipedia.org" \t "_blank)

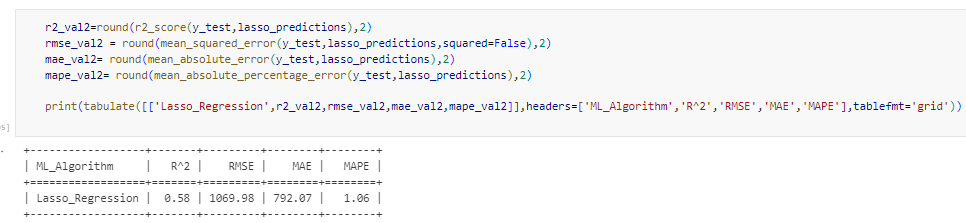
The basic idea of ridge regression is to introduce a little bias so that the variance can be substantially reduced, which leads to a lower overall MSE.



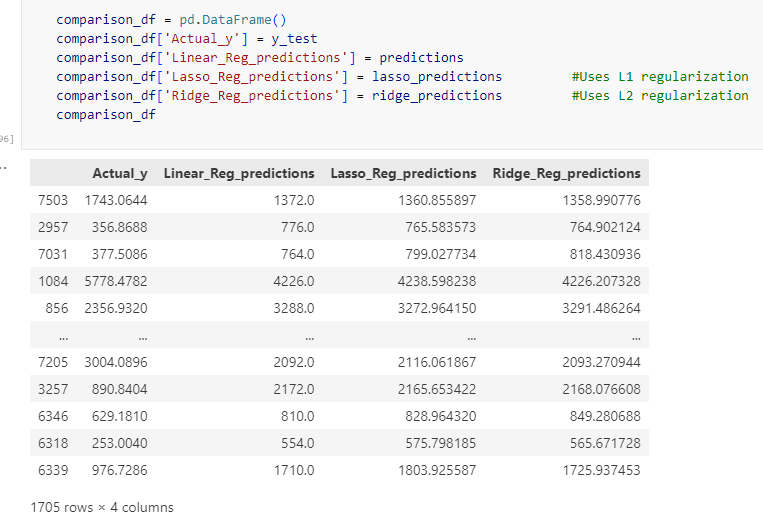
1. Lasso Regression:

In [statistics](https://en.wikipedia.org/wiki/Statistics) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning), lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) method that performs both [variable selection](https://en.wikipedia.org/wiki/Variable_selection) and [regularization](https://en.wikipedia.org/wiki/Regularization_(mathematics)) in order to enhance the prediction accuracy and interpretability of the resulting [statistical model](https://en.wikipedia.org/wiki/Statistical_model). Lasso was originally formulated for [linear regression](https://en.wikipedia.org/wiki/Linear_regression) models. This simple case reveals a substantial amount about the estimator. These include its relationship to [ridge regression](https://en.wikipedia.org/wiki/Ridge_regression) and [best subset selection](https://en.wikipedia.org/wiki/Best_subset_selection) and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that (like standard linear regression) the coefficient estimates do not need to be unique if [covariates](https://en.wikipedia.org/wiki/Covariate) are [collinear](https://en.wikipedia.org/wiki/Collinear).



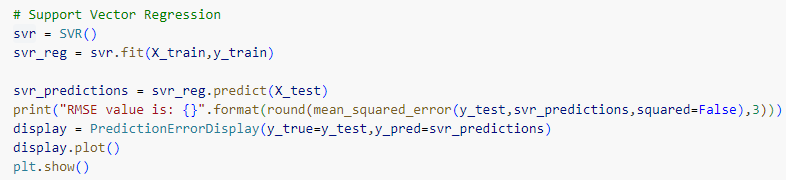


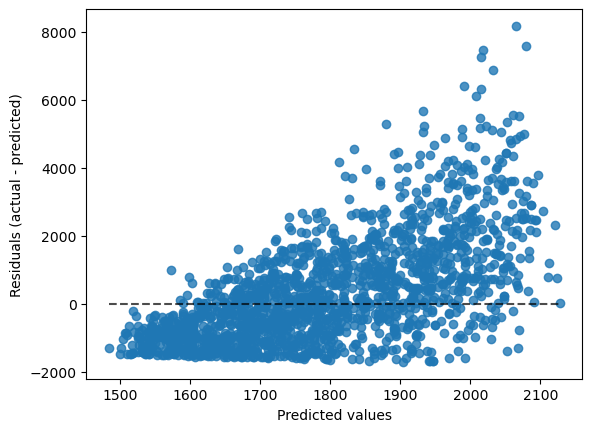
Comparison:

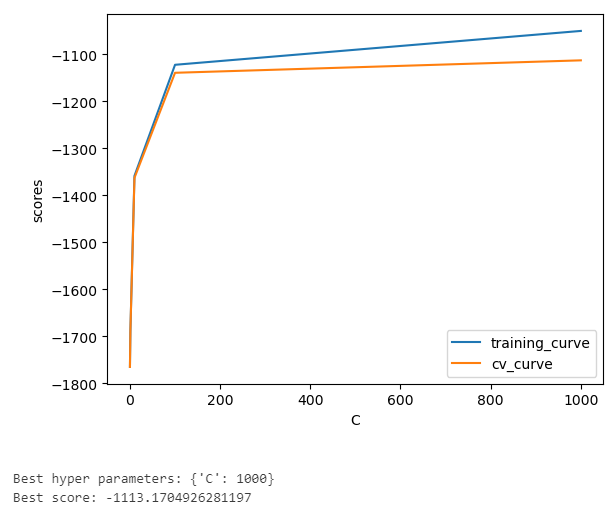
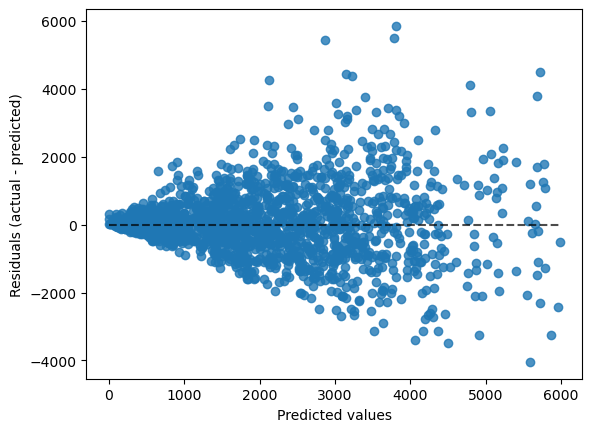
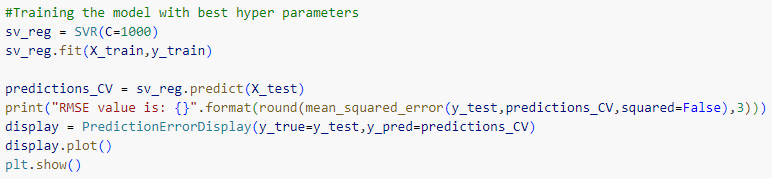


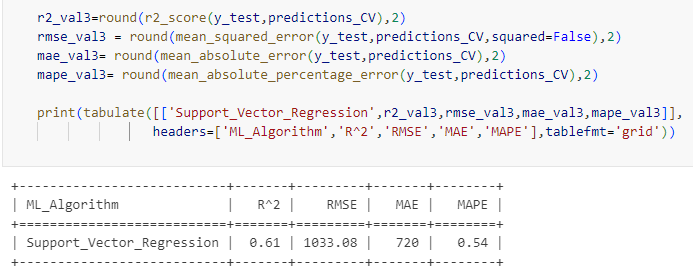
Support Vector Regression (SVR):

Support Vector Regression (SVR) is a type of machine learning algorithm used for regression analysis. The goal of SVR is to find a function that approximates the relationship between the input variables and a continuous target variable, while minimizing the prediction error. Unlike Support Vector Machines (SVMs) used for classification tasks, SVR seeks to find a hyperplane that best fits the data points in a continuous space. This is achieved by mapping the input variables to a high-dimensional feature space and finding the hyperplane that maximizes the margin (distance) between the hyperplane and the closest data points, while also minimizing the prediction error. SVR can handle non-linear relationships between the input variables and the target variable by using a kernel function to map the data to a higher-dimensional space. This makes it a powerful tool for regression tasks where there may be complex relationships between the input variables and the target variable.





Random Forest Regression:

Random forests or random decision forests is an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set). Random forests generally outperform [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning), but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance. 

XG Boost Regression:

[XGBoost regressionis**a method of using XGBoost, a library that provides an efficient and effective implementation of the gradient boosting algorithm, for predicting a numerical value**](https://www.bing.com/ck/a?!&&p=6dae4d8f7c75dbceJmltdHM9MTY5NDMwNDAwMCZpZ3VpZD0wY2UxNTlhOS1hOTU1LTYzMjAtMGFlOC00YmU5YTg4NzYyODUmaW5zaWQ9NTg4Mg&ptn=3&hsh=3&fclid=0ce159a9-a955-6320-0ae8-4be9a8876285&psq=XG+Boost+Regression%3a&u=a1aHR0cDovL3d3dy5yZWdyZXNzaW9uaXN0LmNvbS8yMDIyLzA4L3hnYm9vc3Qv&ntb=1)**.**[XGBoost uses a combination of many smaller decision trees as base learners and an objective function that contains a loss function and a regularization term](https://www.bing.com/ck/a?!&&p=6811f39d66e1c766JmltdHM9MTY5NDMwNDAwMCZpZ3VpZD0wY2UxNTlhOS1hOTU1LTYzMjAtMGFlOC00YmU5YTg4NzYyODUmaW5zaWQ9NTg4Ng&ptn=3&hsh=3&fclid=0ce159a9-a955-6320-0ae8-4be9a8876285&psq=XG+Boost+Regression%3a&u=a1aHR0cDovL3d3dy5yZWdyZXNzaW9uaXN0LmNvbS8yMDIyLzA4L3hnYm9vc3Qv&ntb=1). [The loss function measures the difference between the actual and predicted values, and the most common one for regression problems is: linear](https://www.bing.com/ck/a?!&&p=bef1158922951957JmltdHM9MTY5NDMwNDAwMCZpZ3VpZD0wY2UxNTlhOS1hOTU1LTYzMjAtMGFlOC00YmU5YTg4NzYyODUmaW5zaWQ9NTg5MA&ptn=3&hsh=3&fclid=0ce159a9-a955-6320-0ae8-4be9a8876285&psq=XG+Boost+Regression%3a&u=a1aHR0cHM6Ly9mZWRlcmFscHJpc20uY29tL2lzLXhnYm9vc3QtZ29vZC1mb3ItcmVncmVzc2lvbi8&ntb=1). [XGBoost is a popular and powerful approach for building supervised regression models that can be evaluated using repeated k-fold cross-validation and used to make predictions on new data](https://www.bing.com/ck/a?!&&p=5c6efbc1c14cfea6JmltdHM9MTY5NDMwNDAwMCZpZ3VpZD0wY2UxNTlhOS1hOTU1LTYzMjAtMGFlOC00YmU5YTg4NzYyODUmaW5zaWQ9NTg5Mg&ptn=3&hsh=3&fclid=0ce159a9-a955-6320-0ae8-4be9a8876285&psq=XG+Boost+Regression%3a&u=a1aHR0cHM6Ly9tYWNoaW5lbGVhcm5pbmdtYXN0ZXJ5LmNvbS94Z2Jvb3N0LWZvci1yZWdyZXNzaW9uLw&ntb=1). [XGBoost is also a scalable and distributed library that supports parallel tree boosting and ensemble learning](https://www.bing.com/ck/a?!&&p=e58e92011523ffc7JmltdHM9MTY5NDMwNDAwMCZpZ3VpZD0wY2UxNTlhOS1hOTU1LTYzMjAtMGFlOC00YmU5YTg4NzYyODUmaW5zaWQ9NTg5NQ&ptn=3&hsh=3&fclid=0ce159a9-a955-6320-0ae8-4be9a8876285&psq=XG+Boost+Regression%3a&u=a1aHR0cDovL3d3dy5yZWdyZXNzaW9uaXN0LmNvbS8yMDIyLzA4L3hnYm9vc3Qv&ntb=1).



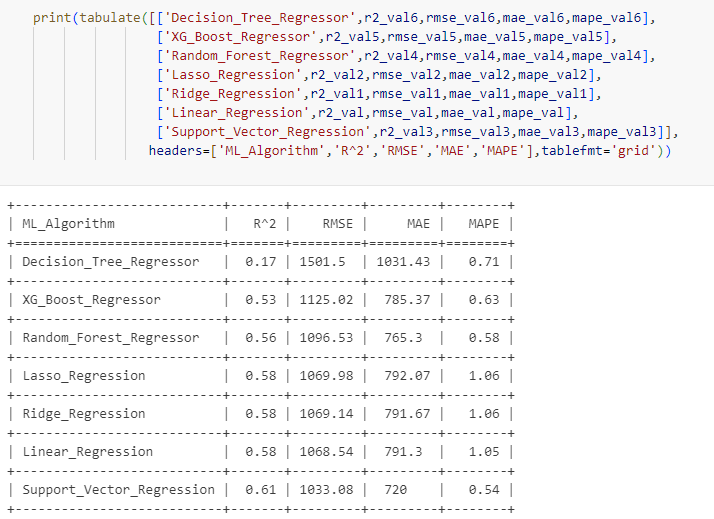
Decision Tree Regression:

Decision Tree is a supervised learning technique that can be used for both classification and regression problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision Tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a Decision Tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.



Comparison between the models

Different models are compared based on coefficient of variation (R^2), RMSE, MAE, and MAPE as follows:



Conclusion

From the above table it can be said that support vector regression model is best among the others considered for predicting the sales of the product.

Thus, using the support vector regression model, the sales on the test data is predicted as follows:

