

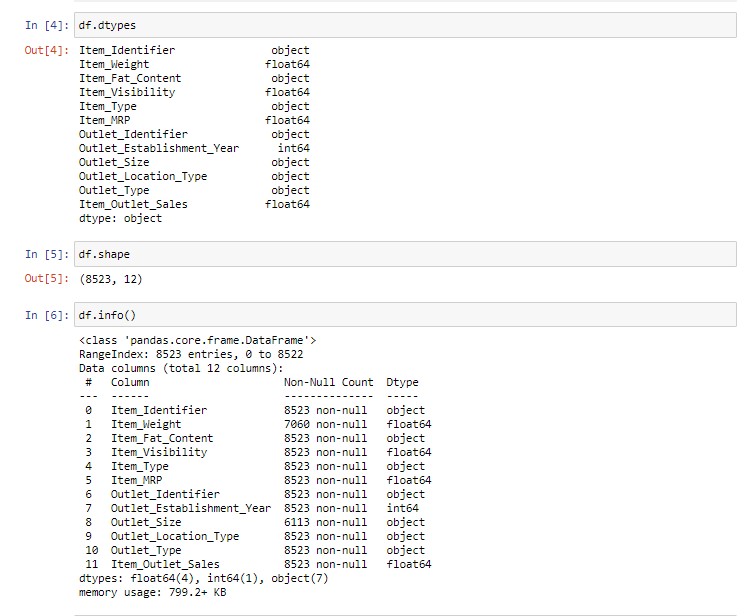
|  |
| --- |
| Big Data Sales Mart Problem |
| *Abstract*:  Nowadays Big Marts and Shopping Malls keep the track details of there sales data of each and every individual item for predicting future demand of the customer and update the inventory management as well.  These data stores basically contain a large number of customer data and  individual item attributes in a data warehouse. Further, anomalies and  frequent patterns are detected by mining the data store from the data  warehouse. The resultant data can be used for predicting future sales  volume with the help of diﬀerent machine learning techniques for the  retailers like Big Mart. In this paper, we propose a predictive model using  Xgboost technique for predicting the sales of a company like Big Mart  and found that the model produces better performance as compared to  existing models. A comparative analysis of the model with others in  terms performance metrics is also explained in details  These data stores basically contain a large number of customer data and  individual item attributes in a data warehouse. Further, anomalies and  frequent patterns are detected by mining the data store from the data  warehouse. The resultant data can be used for predicting future sales  volume with the help of diﬀerent machine learning techniques for the  retailers like Big Mart. In this paper, we propose a predictive model using  Xgboost technique for predicting the sales of a company like Big Mart  and found that the model produces better performance as compared to  existing models. A comparative analysis of the model with others in  terms performance metrics is also explained in details  These data stores basically contain a large number of customer data and  individual item attributes in a data warehouse. Further, anomalies and  frequent patterns are detected by mining the data store from the data  warehouse. The resultant data can be used for predicting future sales  volume with the help of diﬀerent machine learning techniques for the  retailers like Big Mart. In this paper, we propose a predictive model using  Xgboost technique for predicting the sales of a company like Big Mart  and found that the model produces better performance as compared to  existing models. A comparative analysis of the model with others in  terms performance metrics is also explained in details  These data stores basically contain a large number of customer data and individual item attributes in a data warehouse.  The resultant data can be used for predicting future sales volume with the help of different machine learning techniques for the retailers like Big Mart. In this ,we prapose a pedictive model using LinearRegression technique for predicting the sales of the company like Big Mart and sees that the model produces the better performance as compared to existing models. A comparative analysis of the model with others in terms of performance metrics is also explained in details.  *Keywords*: Machine Learning ,LinearRegression, RandomForestRegressor,r2\_score,Cross\_val\_score |
| April 184  *COMPANY NAME* : Data Trained Education  *Authored by*: Deepesh Singh |



# *Article On*

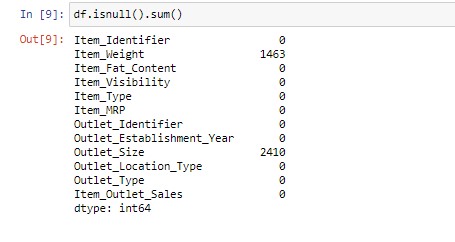
|  |
| --- |
| Predicting the sales on big data mart *The problem Statement*:  “*The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store.*  *Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.”*  *Importing Libraries*:  For proceeding further in the project, we have to import some necessary libraries. |
| *Loading Datasets****:***  We first load the Train csv dataset and then perform on the model. |

Now we check the shape dtypes and info of the datasets.

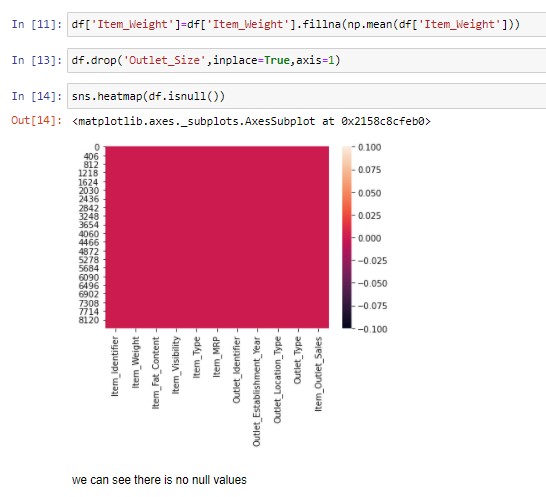


We can see in the above picture that we find the data types, shapes and info of the datasets. We see that total object type data is 7, and float is 4 and int is 1.

Let’s start by checking which columns has missing values in the dataset.



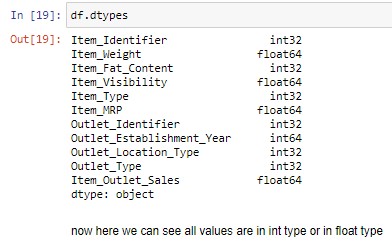
We can see there is missing values are present in dataset in two columns i.e. Item weight and Outlet size. And it should be removed, I removed the null values of the column Item weight by the mean operation, and drop the Outlet size by doing that there is no problem in our dataset and no problem for building the model.



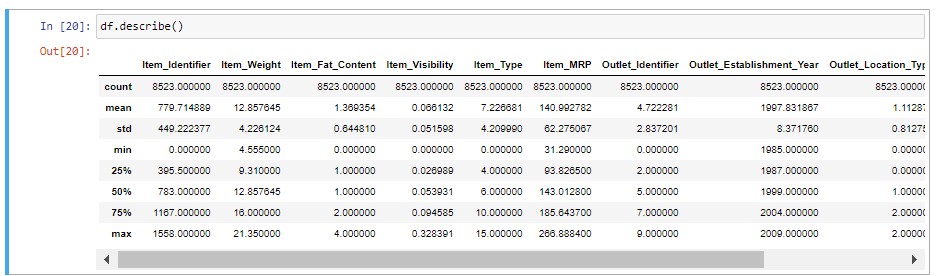
We do some Feature Engineering on the dataset and making our data ready for the analysis. In dataset there are string type categorical values that we have to convert into numerical value. By using Label Encoder, we can do that.



Now we can see that all values are in numerical form that showed below:



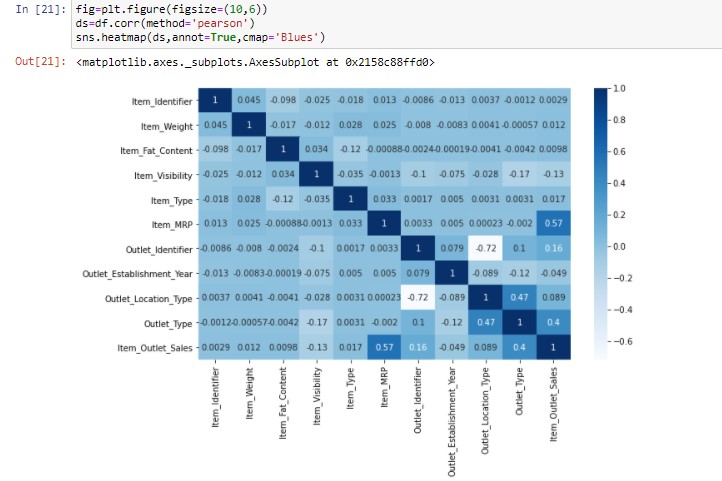
After this we describe the chart of the dataset that shows us some numerical matrices by showing mean, std,25th interquartile and etc.



*Some observations*:

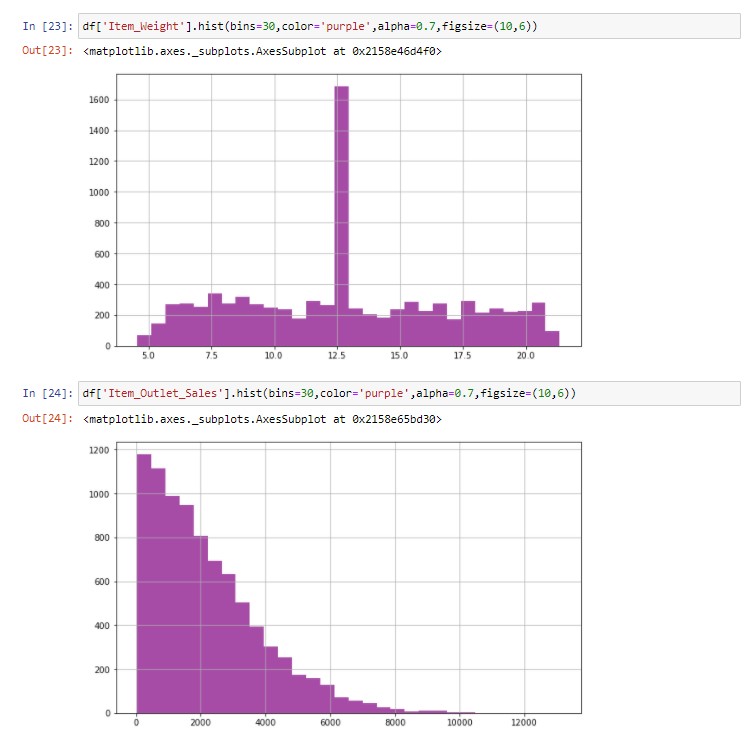
1. Item Viability has a min value of zero. This makes no practical sense because when a product is being sold in store, the visibility cannot be 0.
2. Outlet\_Establishment\_Years vary from 1985 to 2009. The values might not be apt in this form. Rather, if we can convert them to how old the particular store is, it should have better impact on sales.
3. The lower ‘count’ of item\_weight and Item\_Outlet\_sales confirm the findings of the missing values.

Plotting the correlation matrix of the dataset by using ‘Pearson’ method



In above picture we see there the correlation is shown by the blue color and its value is vary from 0 to 1 with lightest blue color to the darkest one.

After that we plot some hist plot of the columns like ‘Item-weight’ and ‘Item\_outlet\_sales that is shown below:



We see the hist plot of the ‘Item\_Outlet\_sales’ shows the negative relationship.

*Plotting pairplot of the data*:



In pairplot all the graph are plotted, whether it is histplot, countplot, scatter plot, barplot, density plot, all plots are plotted.

And all those are showing the relation

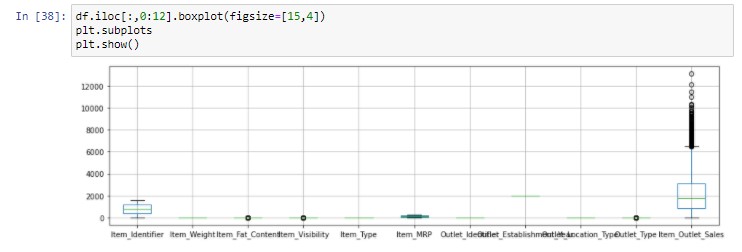
From density plot we see the skewed data, and from the scatter plot we see the relationship between two columns that it is positive or negative with respect to each other.

*Data Cleaning*:

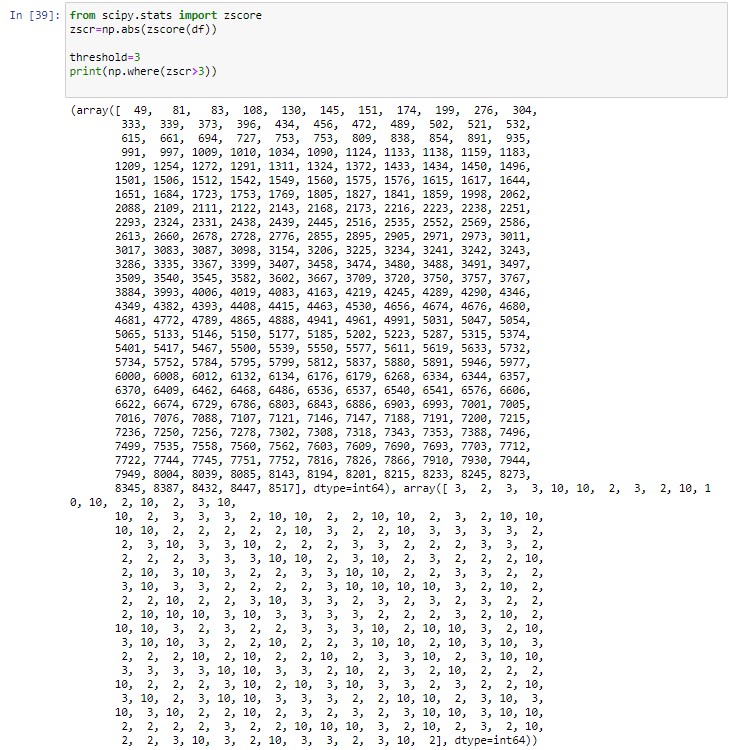
*Checking Outliers*-

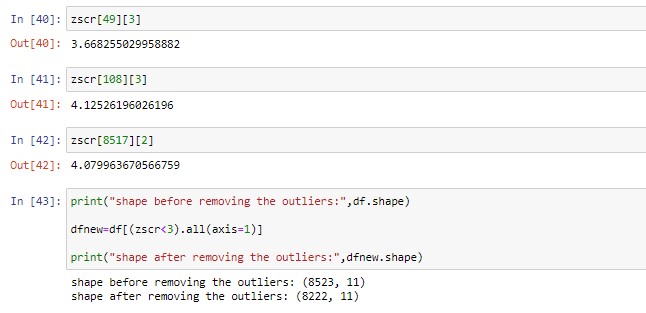
By plotting the box plot we can identify the outliers of the data, and we remove all those outliers which present is huge amount or more than 1 or 2.

For checking the outliers, we plot boxplot of the whole columns present in the data as below:



We can see there is outliers present in the data so we remove those outliers by importing zscore:

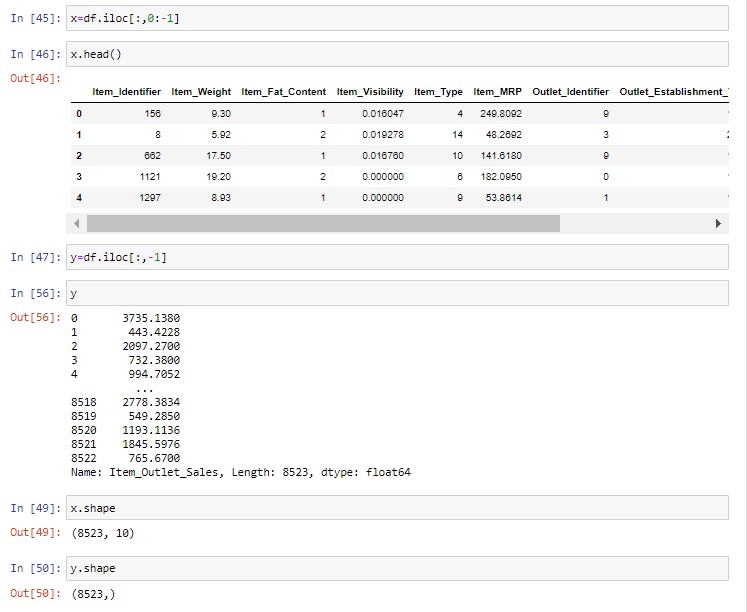




We can see all the outliers are removed and cleansing of the data is done.

*Train Test Split*

Now we separate the dataset into x and y variable for the training the data and the testing the data.

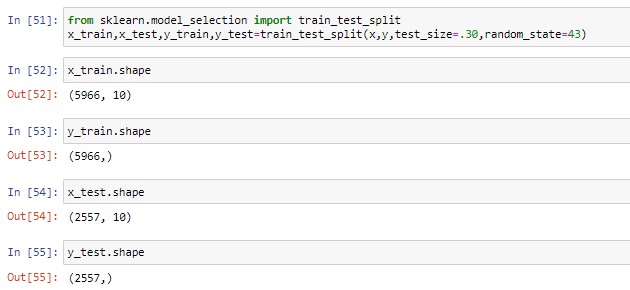


We can see above that our data is now separate into x and y variable.

Now we import train test split and send the data for the training and testing while giving how much data is for training and how much for testing.

And then check the shape of the both the x train and x test, y train and y test.

And also decide the test\_size and the random\_state.



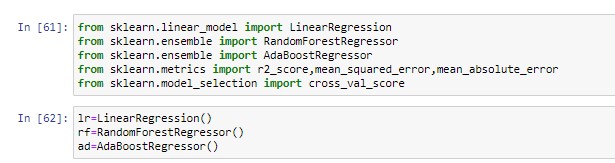
We can see the shape of the train and test data at the random\_state of 43.

*Model Building*

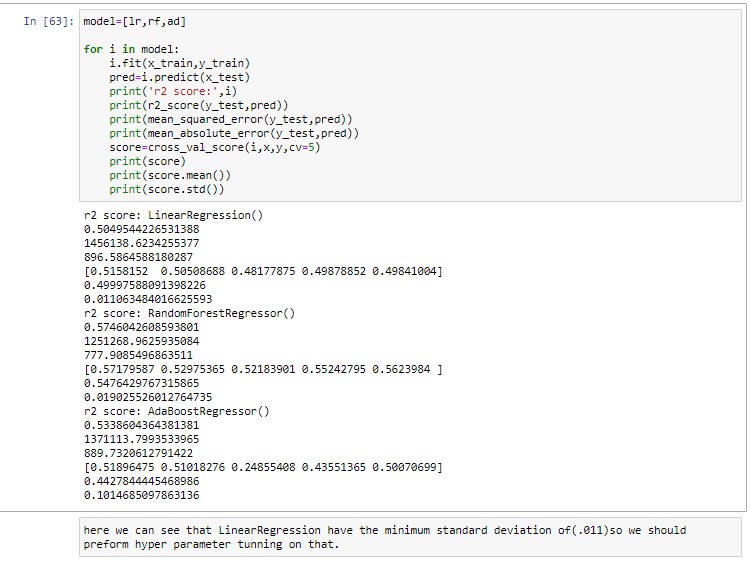
Now that we have the data ready, it’s time to start making predictive models. I will take you through 3 models including LinearRegression, RandomforestRegressor and AdaBoostRegressor.

Let’s start by making a baseline model. Baseline model is the one which requires no predictive model and it’s like an informed guess.

Now for this first importing the models and ensemble techniques.

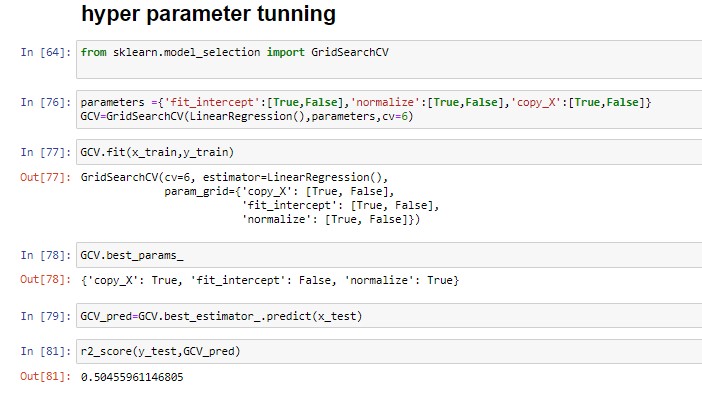


Now predicting all three models at one time by using ‘for’ loop.



We can see above that linear regression has the highest r2\_score with cross validation check. We also do the checking of overfitting/underfitting for exact and accurate score.

For enhancing the predictive score, we go for a hyper parameter tuning that will give our model better score by which we can predict better sales.



After Hyper parameter tuning, we get better score.

*Saving the Model*

Now we have to save the model

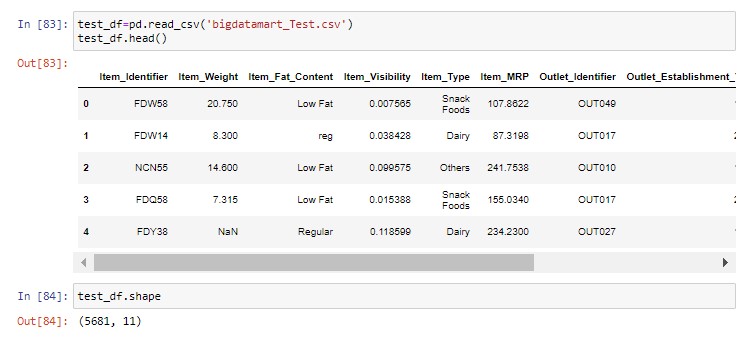
I do this with pickle and for this we have to import the pickle first



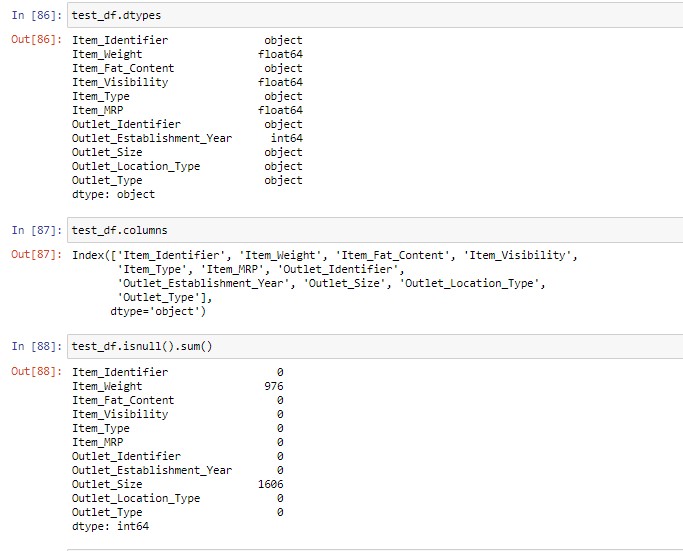
We save the model for the sales prediction.

Now we go for the Test\_csv data and perform all the process which we have done with Train\_csv data till EDA and data cleansing.

Loading the test dataset

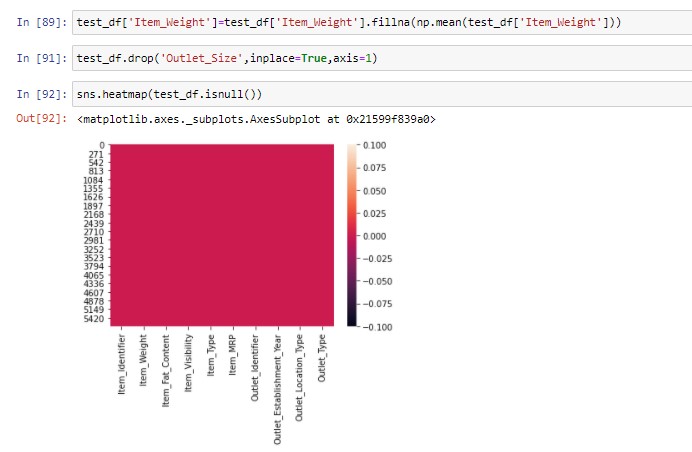


We loaded the dataset and checking for the shape, data types, info (), null values.



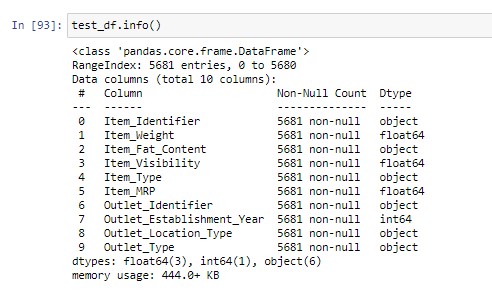
We can see that there are null values present in the dataset so we have to remove those null values.

We remove the null values as same as we done in train data

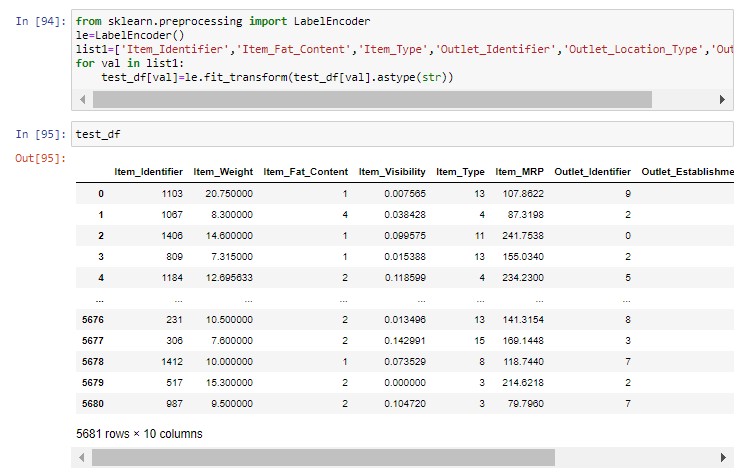


We can see above in heatmap that our null values of the test data are removed now.

Now we converting the categorical columns into the numerical columns, and first we check how much of our data is in the categorical form.

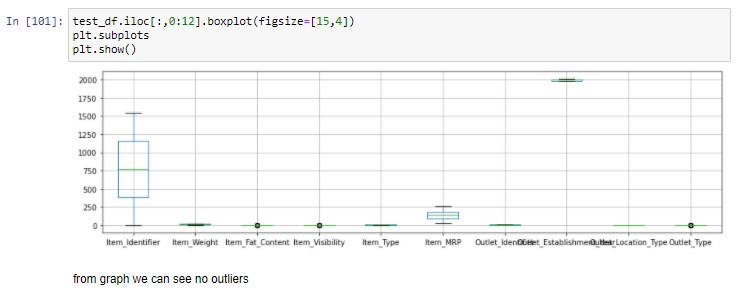


Now importing the Label Encoder and converting the data into numerical form



We can see that all the test data is in the form of float type or int type.

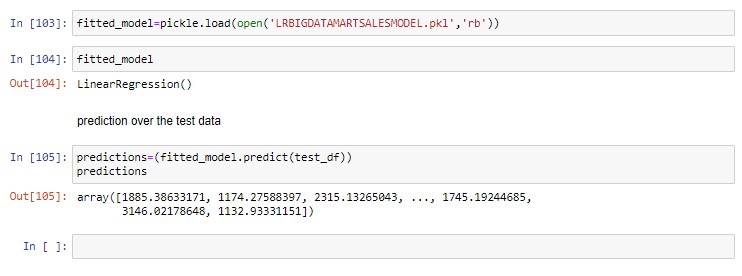
Checking outliers of the test data



We observe that there is no outliers present in the test dataset, some are present but they are too close to the data, so we don’t have to remove that.

Load The Model

Now we load the model which we saved earlier, which is of the train data, now we call that data for predicting the test data.



We can see above the predicted test data of the Big Mart.

*End Notes:*

This article took us through the entire process if solving a data science problem.

We started with making some hypothesis about the data. Then we moved on the data analysis where we found some unusual data, which has to be removed. And then we perform data cleaning and do some feature engineering. Where we imputed the missing value and solved the irregularities.

And converting the categorical data into numerical form by Label Encoder.

Finally, we made Linear regression, RandomForestRegressor, AdaBoostRegressor model and got a glimpse of how to tune them for better results.

And in last we predicted the sales problem of the big mart by which they can enhance the sales problem of the big mart.