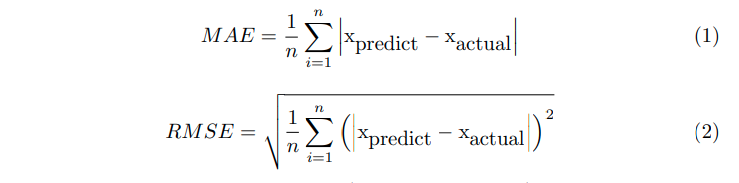
Problem Definition

*Day by day competition among different shopping malls as well as big marts is getting more serious and aggressive only due to the rapid growth of the global malls and on-line shopping. Every mall or mart is trying to provide personalized and short-time offers for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organization, logistics and transport service, etc. Present machine learning algorithm are very sophisticated and provide techniques to predict or forecast the future demand of sales for an organization, which also helps in overcoming the cheap availability of computing and storage systems. In this paper, we are addressing the problem of big mart sales prediction or forecasting of an item on customer’s future demand in different big mart stores across various locations and products based on the previous record. Different machine learning algorithms like linear regression analysis, random forest, etc are used for prediction or forecasting of sales volume. As good sales are the life of every organization so the forecasting of sales plays an important role in any shopping complex. Always a better prediction is helpful, to develop as well as to enhance the strategies of business about the marketplace which is also helpful to improve the knowledge of marketplace. A standard sales prediction study can help in deeply analyzing the situations or the conditions previously occurred and then, the inference can be applied about customer acquisition, funds inadequacy and strengths before setting a budget and marketing plans for the upcoming year. In other words, sales prediction is based on the available resources from the past. In depth knowledge of past is required for enhancing and improving the likelihood of marketplace irrespective of any circumstances especially the external circumstance, which allows to prepare the upcoming needs for the business. Extensive research is going on in retailers domain for forecasting the future sales demand. The basic and foremost technique used in predicting sale is the statistical methods, which is also known as the traditional method, but these methods take much more time for predicting a sales also these methods could not handle non linear data so to over these problems in traditional methods machine learning techniques are deployed. Machine learning techniques can not only handle non-linear data but also huge data-set efficiently. To measure the performance of the models, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used as an evaluation metric as mentioned in the Equation 1 and 2 respectively. Here both metrics are used as the parameter for accuracy measure of a continuous variable.*



*where n: total number of error and xpredict − xactual : Absolute error. The remaining part of this article is arranged as following: Section 1 briefly describes introduction of sales prediction of Big Mart and also elaborate about the evaluation metric used in the model. Previous related work has been pointed in Section 2. The detailed description and analysis of proposed model is given in Section 3. Where as implementations and results are demonstrated in Section 4 and the paper concludes with a conclusion in the last section.*

Proposed System

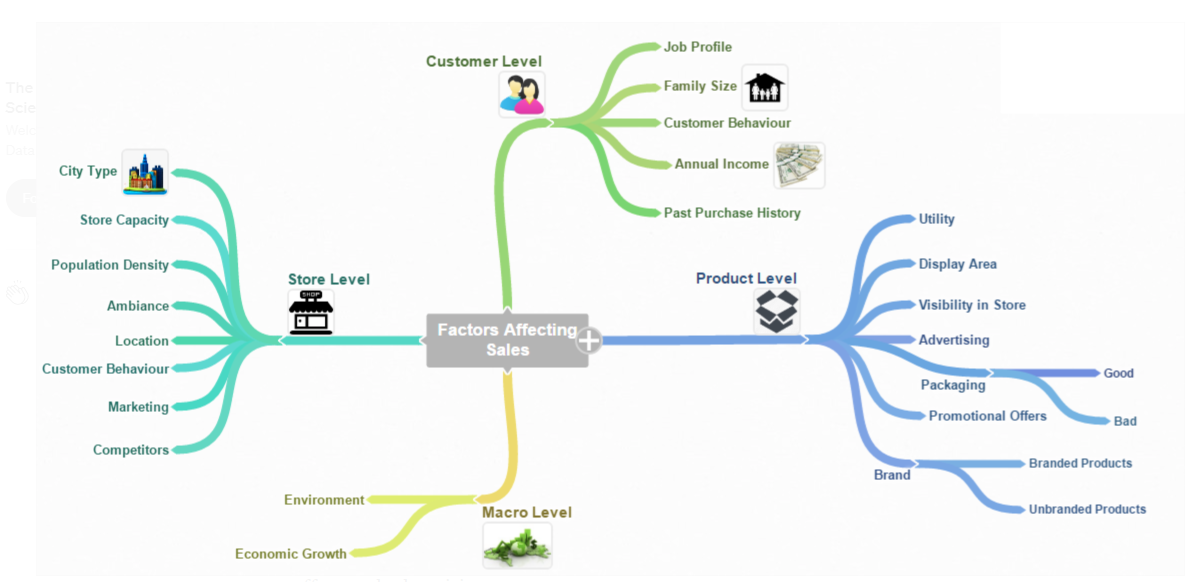
*For building a model to predict accurate results the dataset of Big Mart sales undergoes several sequence of steps as mentioned in Figure 1 and in this work we propose a model using Xgboost technique. Every step plays a vital role for building the proposed model. In our model we have used 2013 Big mart dataset [13]. After preprocessing and filling missing values, we used ensemble classifier using Decision trees, Linear regression, Ridge Lasso regression, Random forest and Xgboost. Both MAE and RSME are used as accuracy metrics for predicting the sales in Big Mart. From the accuracy metrics it was found that the model will predict best using minimum MAE and RSME. The details of the proposed method is explained in the following section.*

Select a Performance Measure

*Usually for regression problems the typical performance measure is the RootMean Square Error (RMSE). This function gives an idea of how much error the system makes in its predictions with higher weight for large errors. To measure the performance of the models, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used as an evaluation metric as mentioned in the Equation 1 and 2 respectively*

Assumptions:

*After framing our problem and deciding on the performance measure, it is good to make some assumptions on what possible outcomes we might expect from our analysis according to the available data. Therefore, by knowing the goal we should think which possible factors might affect the sales prediction outcome. We can start by diving the process into four* *levels:****Store level, Product level, Customer level and Macro level.***

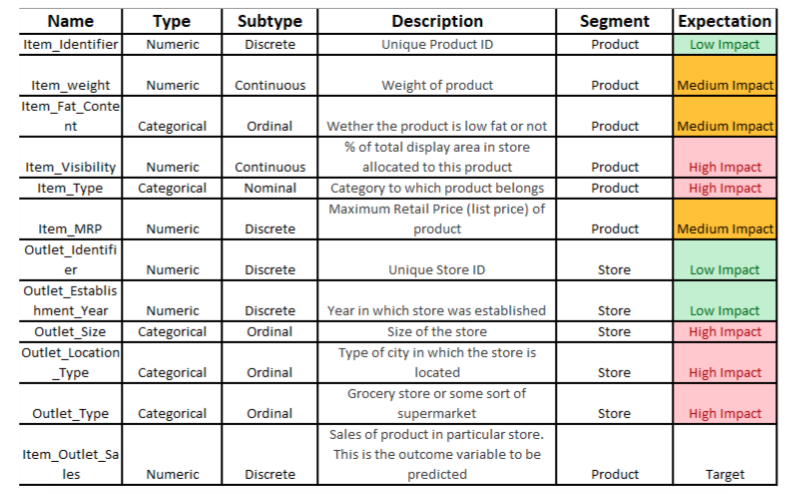


Store Level Hypotheses:

1. ***City type****: Stores located in urban or Tier 1 cities should have higher sales because of the higher income levels of people there.*
2. ***Population Density****: Stores located in densely populated areas should have higher sales because of more demand.*
3. ***Store Capacity****: Stores which are very big in size should have higher sales as they act like one-stop-shops and people would prefer getting everything from one place*
4. ***Competitors****: Stores having similar establishments nearby should* *have less sales because of more competition.*
5. ***Marketing****: Stores which have a good marketing division should have higher sales as it will be able to attract customers through the right offers and advertising.*
6. ***Location****: Stores located within popular marketplaces should have higher sales because of better access to customers.*
7. ***Customer Behavior****: Stores keeping the right set of products to meet the local needs of customers will have higher sales.*
8. ***Ambiance****: Stores which are well-maintained and managed by polite and humble people are expected to have higher footfall and thus higher sales.*

Product Level Hypotheses:

1. ***Brand:****Branded products should have higher sales because of higher trust in the customer.*
2. ***Packaging:****Products with good packaging can attract customers and sell more.*
3. ***Utility:****Daily use products should have a higher tendency to sell as compared to the specific use products****.***
4. ***Display Area:****Products which are given bigger shelves in the store are likely to catch attention first and sell more.*
5. ***Visibility in Store:****The location of product in a store will impact sales. Ones which are right at entrance will catch the eye of customer first rather than the ones in back.*
6. ***Advertising:****Better advertising of products in the store will should higher sales in most cases****.***
7. ***Promotional Offers:****Products accompanied with attractive offers and discounts will sell more.*

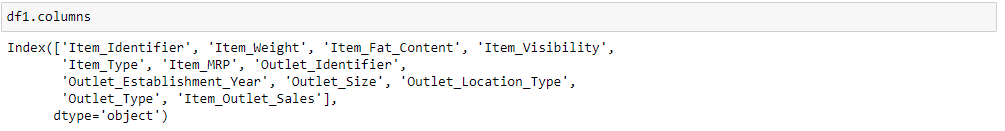


*As we can see from the image above most of the features follow into the segments “Product” and “Store”, therefore we can exclude our previous assumptions for the other two segments considered since we do not have data to support them.*

*It was my belief that from this first look at the data, the variables that will have higher impact on the product’s sale price are:*

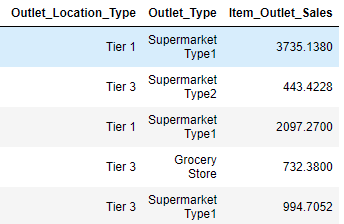
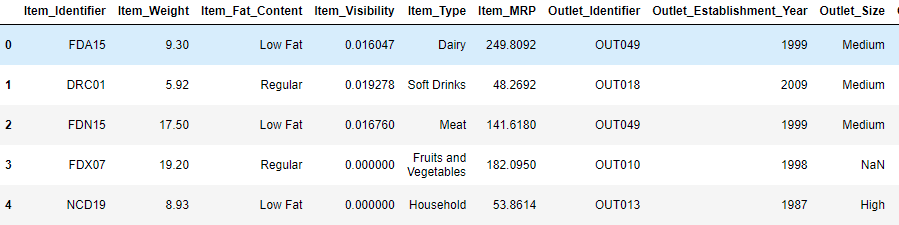
***Item\_Visibility , Item\_Type , OutletSize , Outlet\_Location\_Type , Outlet\_Type****.*

*The target variable is****Item\_Outlet\_Sales****.*



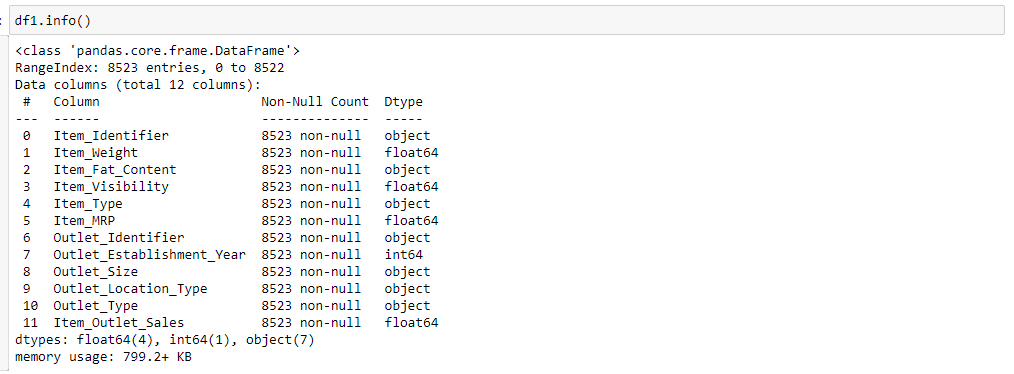
Data Sturcture:-





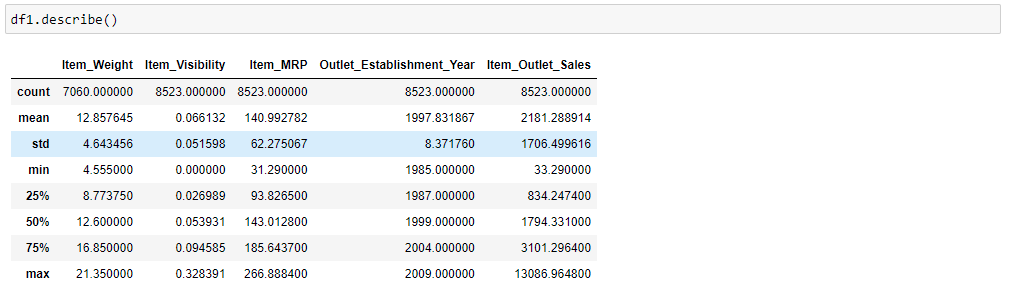
*If we look at variable Item\_Identifier , we can see different group of letters per each product such as ‘FD’ (Food), ‘DR’(Drinks) and ‘NC’ (Non-Consumable).*

*On the other hand, regarding Item\_Visibility there are items with the value zero . This does not make lot of sense, since this is indicating those items are not visible on the store.*



*Most of the items in the train dataset present 8523 non-null values. However, there are some cases such as Item\_Weight and Outlet\_Size which seem to present Null values. We always have to consider if this absence of values has a significant meaning. In this case it does not since all values should have weight higher than 0 and a stores cannot exist with zero size.*

*Moreover, from the 12 features, 5 are numeric and 7 categorical.*

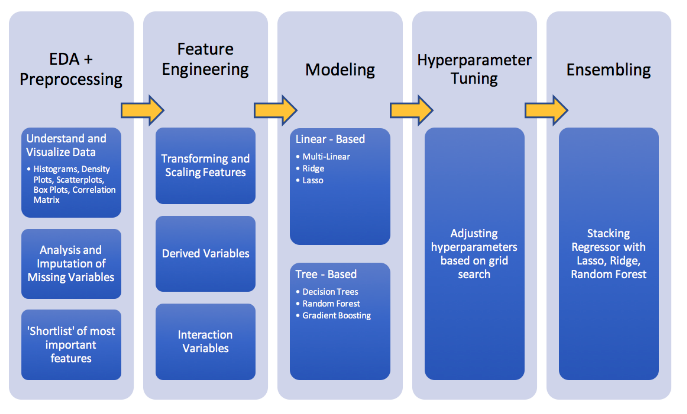


## Organization of our analysis

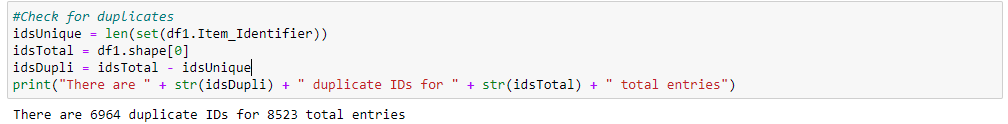
*Our goal as a Data Scientist is to identify the most important variables and to define the best regression model for predicting out target variable. Hence, this analysis will be divided into five stages:*

1. *Exploratory data analysis (EDA);*
2. *Data Pre-processing;*
3. *Feature engineering;*
4. *Feature Transformation;*
5. *Modeling;*
6. *Hyperparameter tuning*
7. *Ensembling.*

*Processing will follow steps shown in below picture.*

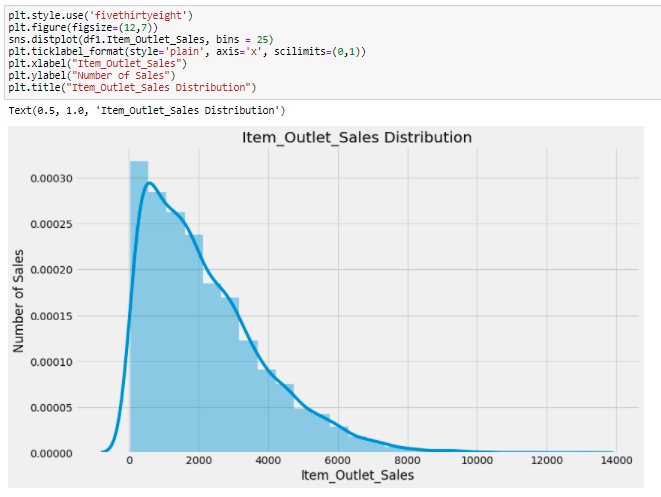


1. ***Exploratory Data Analysis (EDA):-*** *First of all we will check available data, its quality, will check if data contain duplicate values , Null/Missing values etc.*



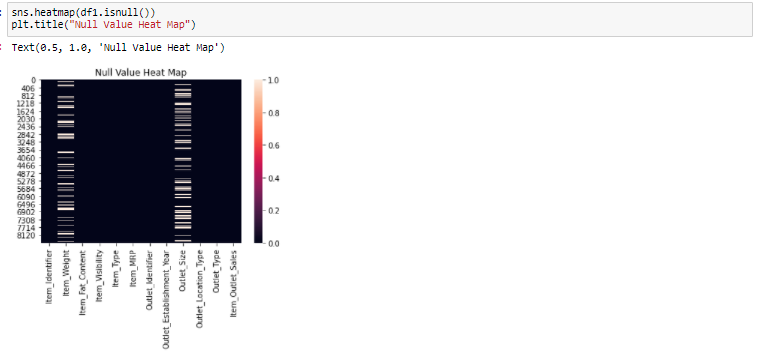
*Before starting the analysis it is interesting to check if the dataset suffers from duplicate values. In this case, the way to see if there are any duplicates is using the Item\_Identifier feature. Since a product can exist in more than one store it is expected for this repetition to exist. Curious fact is that there seems to be 1562 unique items only available in one single store.*

* 1. ***Univariate Analysis***
     1. ***Distribution of the target variable*:***Item\_Outlet\_Sales*

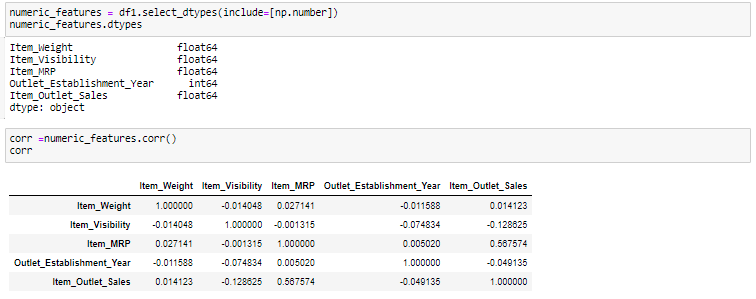


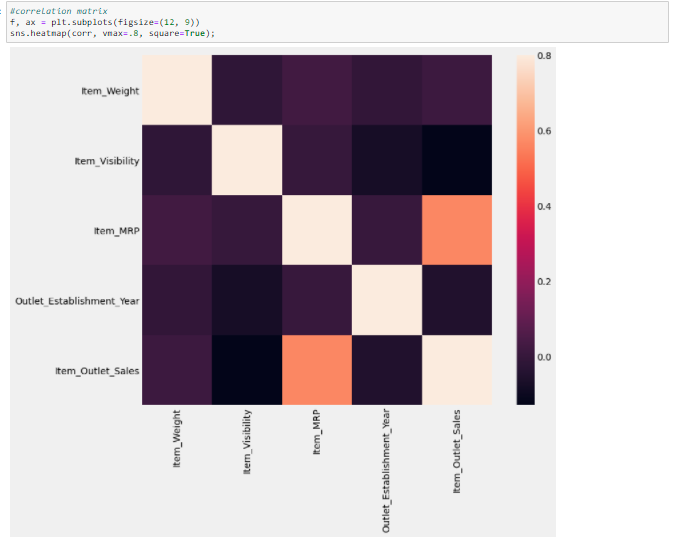
*We can see that our target variable is skewed to the right, towards the higher sales, with higher concentration on lower sales. To make this distribution more symmetrical, we could try taking its square root. Nevertheless, my results showed that the RMSE increased with this alteration.*

* + 1. ***Null Value Check:-***

 *We can see two variable Item\_weight and Outlet\_size has null value.*

* + 1. ***Correlation of Numerical value with target variable.***

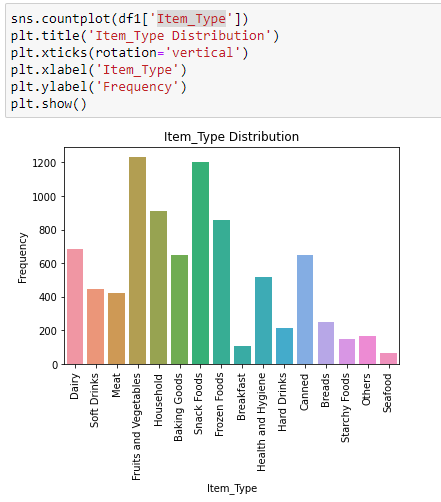




*From the current numeric variables we can observe that the Item\_Visibility is the feature with the lowest correlation with our target variable. Therefore, the less visible the product is in the store the higher the price will be. This is curious since from the initial assumptions this variables was expected to have high impact in the sales increase. Nevertheless, since this is not an expected behaviour and we should investigate. Moreover, this feature has a negative correlation with all of the other features. Furthermore, the most positive correlation belongs to Item\_MRP*

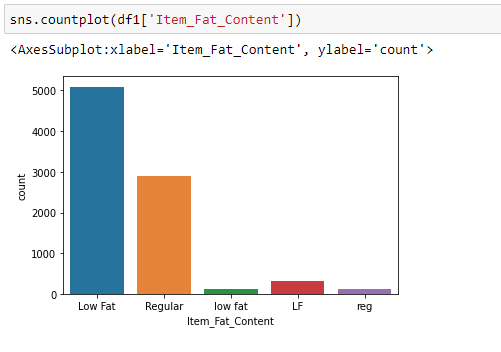
* + 1. ***Distribution of the variable****Item\_Type:-*

*Next we will check the distribution of Item\_type variable, there are 16 item\_type in this dataset, and we can see that maximum frequency is is coming for food items(Snack Food and Fruits and Vegetables.*

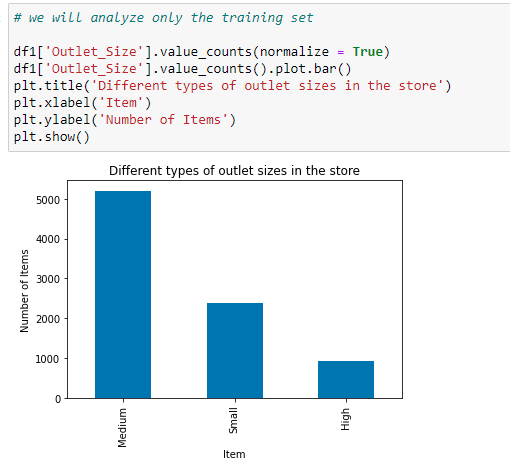


* + 1. ***Distribution of the variable****Item\_Fat\_Content:-*

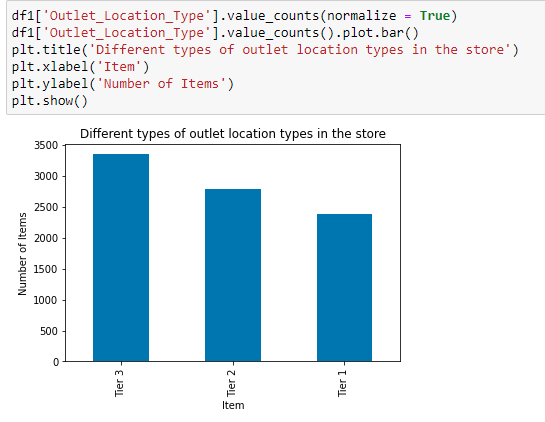
*For Item\_Fat\_Content there are two possible choices : “Low Fat” or “Regular”. However, in our data we have these two types of Fat writen in different manners. This must be corrected.*



* + 1. ***Distribution of the variable****Outlet\_Size*



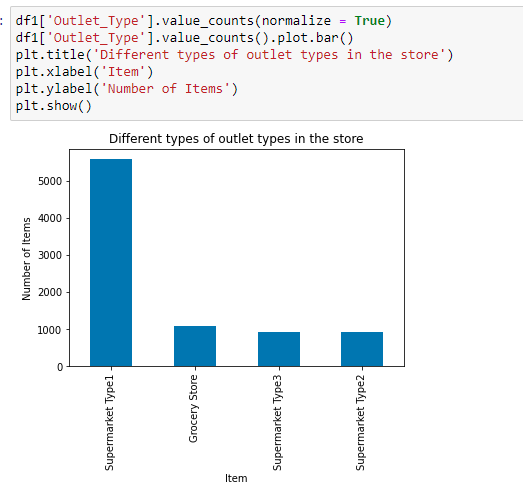
* + 1. ***Distribution of the variable****Outlet\_Location\_Type*



*Bigmart appears to be a supermarket brand that is more present in “Small” to “Medium” size cities than in more densily populated locations.*

* + 1. ***Distribution of the variable****Outlet\_Type:-*

*It looks like Supermarket Type2 , Grocery Store and Supermarket Type3 all have low expression in this distribution. Maybe we can create a single category with all of the three.* *Nevertheless, before doing this we must see their impact in the Item\_Outlet\_Sales .*



## *Bivariate Analysis*

## *Numerical Variable:-*

## *Firstly we individually analysed some of the existent features, now it is time to understand the relationship between our target variable and predictors as well as the relationship among predictors.*

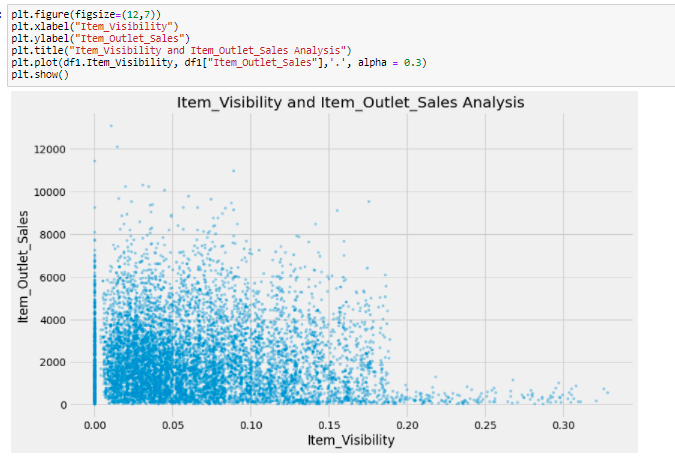
## *Item\_Weight and Item\_Outlet\_Sales analysis*

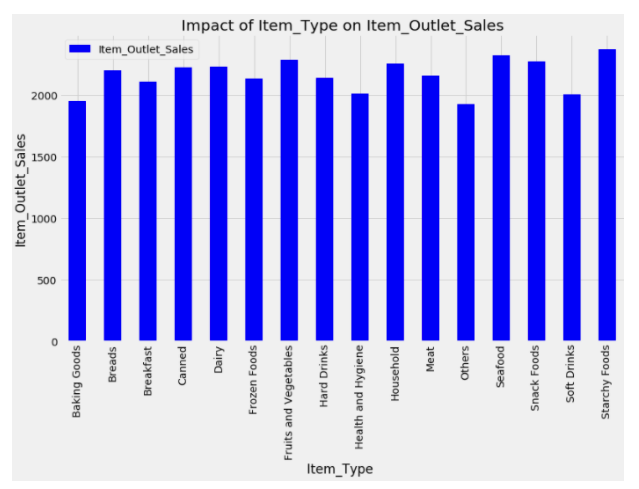
## *We saw previously that Item\_Weight had a low correlation with our target variable. If we plot both features we can see that relationship.*

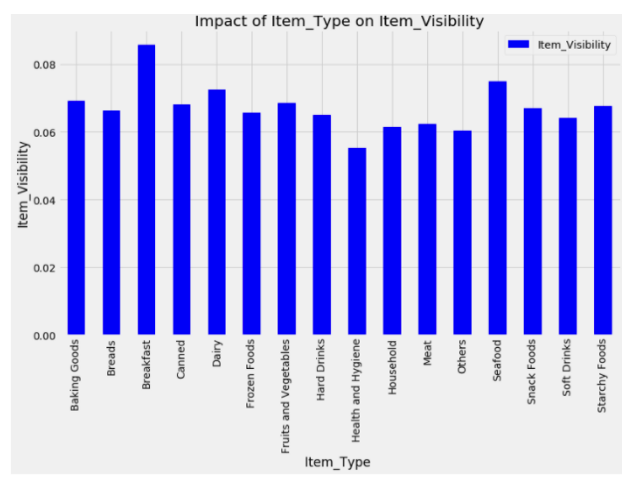
## 

## *Item\_Visibility and Item\_Outlet\_Sales analysis*

***Visibility in Store:****The location of product in a store will impact sales. Ones which are right at entrance will catch the eye of customer first rather than the ones in back. This was the assumption made… however, first the correlation and now this plot chart, indicate that the more visible a product is the less higher its sales will be. This might be due to the fact that a great number of daily use products, which do not need high visibility, control the top of the sales chart. As we can see from the bar charts below, most sold products have lower visibility. Furthermore, there is a concerning number of products with visibility zero.*







## *Outlet\_Establishment\_Year and Item\_Outlet\_Sales analysis*

*There seems to be no significant meaning between the year of store establishment and the sales for the items. 1998 has low values but thet might be due to the fact the few stores opened in that year.*

## 

## *Catagorical Variable:-*

## *Impact of Item\_Fat\_Content onItem\_Outlet\_Sales*

## *Daily use products should have a higher tendency to sell as compared to the specific use products. “Low Fat” products seem to have higher sales values than “Regular” products*

## 

## *Impact of Outlet\_Identifier on Item\_Outlet\_Sales:*

## 

## *From the ten stores, two are Groceries whereas six are Supermarket Type1, one Supermarket Type2 and one Supermarket Type3. You can get this information from the pivot\_tables below.*

## *From the above bar chart, we see that thr groceries (“OUT010”, “OUT019”) have the lowest sales results which is expected followed by the Supermarket Type 2 (“OUT018”). Curiously, most stores are of type Supermarket Type1 of size “High” and do not have the best results. The best results belong to “Out027” which is a “Medium” size Supermarket Type 3.*

## 

## 

## *Impact of Outlet\_Size on Item\_Outlet\_Sales*

## *In the beginning, our belief was that stores which are very big in size should have higher sales as they act like one-stop-shops and people would prefer getting everything from one place. According to the results, this is almost the case. Curiously, consumers tend to prefer medium size stores instead of big size. As we saw in the previous section, most stores have size “Medium” but still the “High” and “Small” stores which are clearly in an inferior number can beat or even come close to their numbers.*

## 

## *Impact of Outlet\_Type on Item\_Outlet\_Sales:- From this analysis possibly it would be a good idea to creat a new feature that shows the sales ratio according to the store size.*

## 

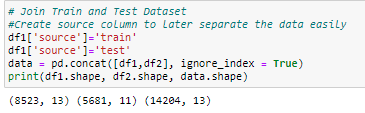
## *Impact of Outlet\_Location\_Type on Item\_Outlet\_Sales:- Do Tier 1 cities have higher sales? This was one of the premisses we made in the start of this study. However, if we look at our results we see that in fact it is stores from Tier 2 cities that present the highest results, followed by Tier 3 cities and with Tier 1 cities with the lowest results of the three type of locations. From the pivot\_table it is easy to see that Tier2 and Tier3 cities are those that have highest representation of stores.*

## 

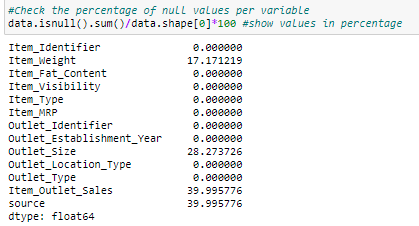
## 

1. Data Pre-Processing:- *During our EDA we were able to take some conclusions regarding our first assumptions and the available data:*

* *Regarding the variables which were thought to have high impact on the product’s sale price. Item\_Visibility does not have a high positive correlation as expected, quite the opposite. As well, there are no big variations in the sales due to theItem\_Type . On the other hand, it was possible to see that the size, location and type of store could have a positive impact on sales.*
* *If we look at variable Item\_Identifer , we can see different groups of letters per each product such as ‘FD’ (Food), ‘DR’(Drinks) and ‘NC’ (Non-Consumable). From this we can create a new variable.*
* *Regarding Item\_Visibility there are items with the value zero. This does not make lot of sense, since this is indicating those items are not visible on the store.*
* *Similar, Item\_Weight and Outlet\_Size seem to present NaN values.*
* *There seems to be 1562 unique items only available in a single store.*
* *Item\_Fat\_Content has vale “low fat” writen in different manners.*
* *For Item\_Type we try to create a new feature that does not have 16 unique values.*
* *Outlet\_Establishment\_Year besides being a hidden category, its values vary from 1985 to 2009 . It must be converted to how old the store is to better see the impact on sales.*
  1. Looking for missing values:-*Usually, datasets for every challenge such as those presented in Analytics Vidhya or Kaggle come seperated as a train.csv and a test.csv. It is generally a good idea to combine both sets into one, in order to perform data cleaning and feature engineering and later divide them again. With this step we do not have to go through the trouble of repeting twice the same code, for both datasets. Let’ s combine them into a dataframe datawith a source column specifying where each observation belongs.*

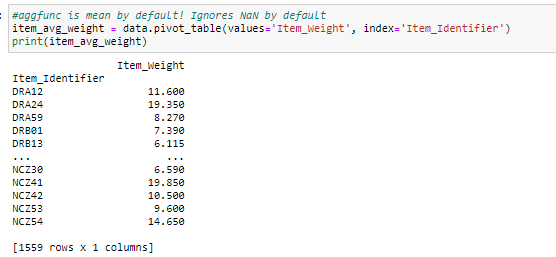


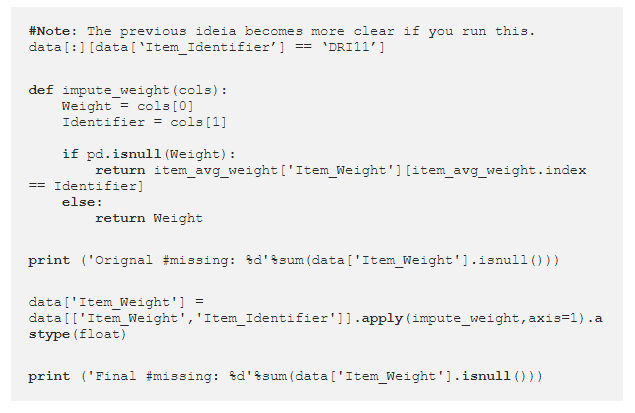
*The previous image shows that our data has some NaNvalues. Numpy’s NaNvalues are good because Pandas is able to recognize this object and count it. However, let’s image that instead of Nulla NaNthere’s, for example, “Not available”. This is counted as a string and not NaN which masks our missing values.*



*Note that the Item\_Outlet\_Salesis the target variable and 39% of its values are NaN. For this case only, the missing values are the ones which belong to the test set. Remember we added the train and test datasets. So we need not worry about it. Neverthless, we’ll impute the missing values in Item\_Weight and Outlet\_Size.*

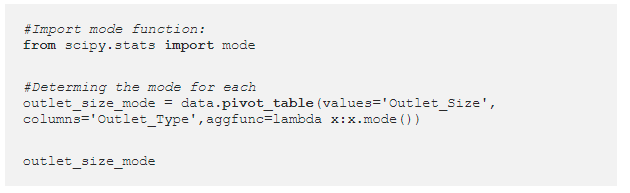
* + 1. Imputing Missing Values:- *We’ve seen previously on the EDA section that the Item\_Weight and the Outlet\_Size had missing values. Hence, for now we will impute for this missing values the mean for each corresponding variable.*
    2. ***Imputing the mean for Item\_Weight missing values***



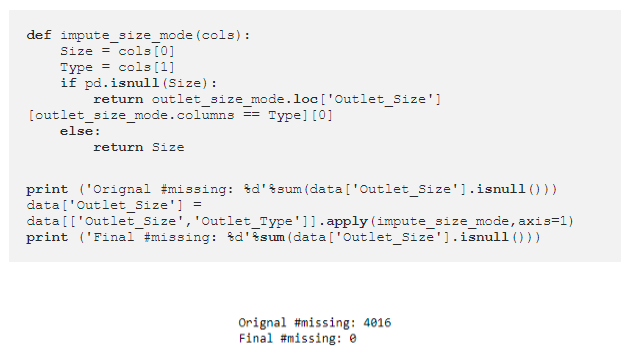
*After running the above code, in which we use function impute\_weight to you get the following message showing there are no more missing values.*

https://miro.medium.com/max/172/1*-04aeDCiJB_G0UCKuDoFIQ.png

* + 1. ***Imputing Outlet\_Size missing values with the mode:***  *we will apply the same logic. In this case, instead of using the default codeaggfunc = mean() for the pivot\_table()we will use the mode.*

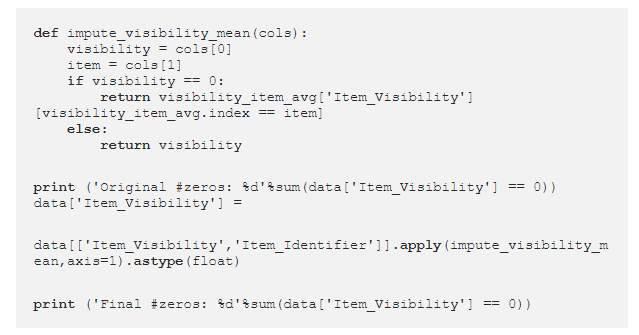






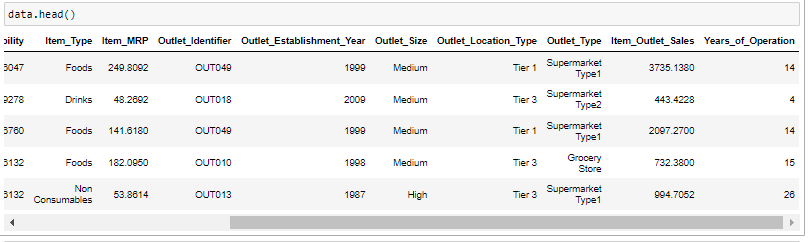
1. Feature Engineering:- *So we now got rid of all the missing values in our data that might negatively influence our analysis. If you remember, during our analysis we saw some nuances in the data and now is time to fix them and make our data ready for analysis.*

*3.1* ***Item\_Visibility minimum value is 0:-***  *In our data exploration we saw that Item\_Visibility had the minimum value 0, which makes no sense since every product must be visible to all clients. Let’s consider it as missing value and impute it with mean visibility of that product.*

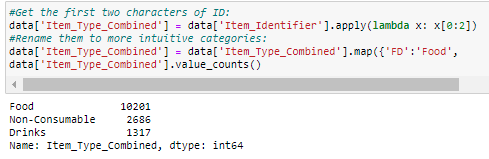


* 1. ***Determine the years of operation of a store:-*** *We talked about using how long has been working instead of the date of start. Remember that the data we have is from 2013. Thus we must consider this year into our calculations-*

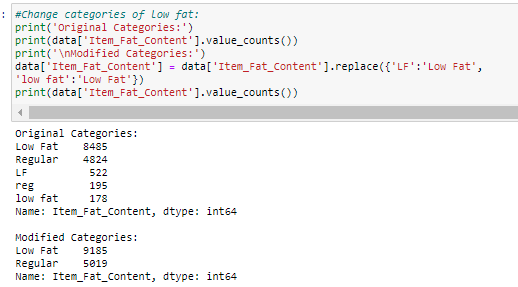




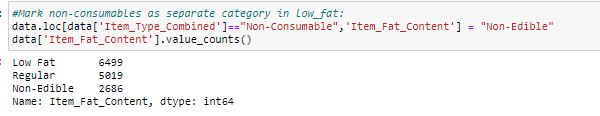
*3.3 Earlier we saw that the Item\_Type variable has 16 categories which might not prove to be very useful in our analysis. So it’s a good idea to combine them. If we look closely to the Item\_Identifier of each item we see that each one starts with either “FD” (Food), “DR” (Drinks) or “NC” (Non-Consumables). Therefore, we can group the items within these 3 categories*



* 1. *Modify categories of Item\_Fat\_Content:- We found typos and difference in representation in categories of Item\_Fat\_Content variable. This can be corrected as:*

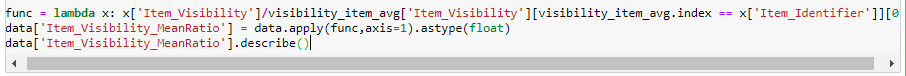


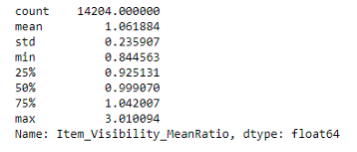
*Now it makes more sense. But hang on, we also saw in section 3.4. there were some non-consumables as well and a fat-content should not be specified for them. So we can also create a separate category for such kind of observations.*



1. Feature Transformations:-

*4.1* ***Creating variable Item\_Visibility\_Mean\_Ratio:-*** *In the beginning of this article we hypothesized that products that are more visible are likely to have higher sales. For example, we can create a new variable that show us the importance given to a product in a given store according to the mean of significance given to the same product in all other stores.*

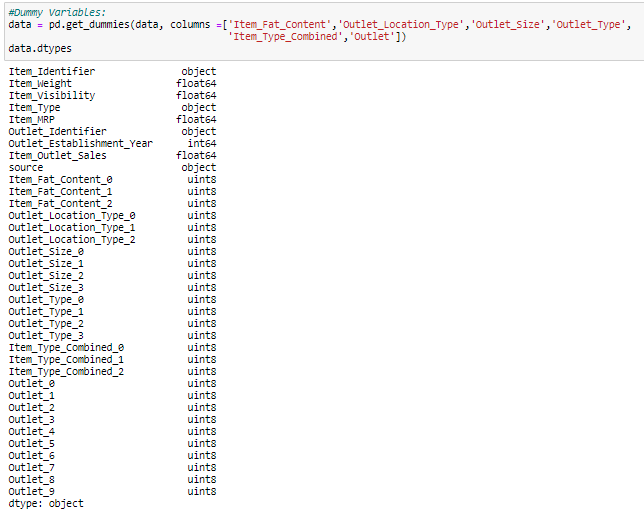




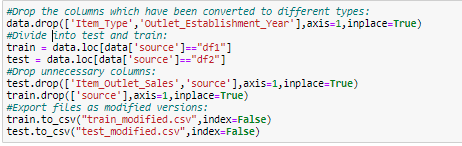
* 1. ***Categorical Variables — One Hot Encoding:-*** *Since scikit-learn only accepts numerical variables, we need to convert all categories of nominal variables into numeric types. Let’s start with turning all categorical variables into numerical values using LabelEncoder() (Encode labels with value between 0 and n\_classes-1). After that, we can use get\_dummies to generate dummy variables from these numerical categorical variables*



*One-Hot-Coding refers to creating dummy variables, one for each category of a categorical variable. For example, the Item\_Fat\_Content has 3 categories — LowFat,Regular,Non-Edible. One hot coding will remove this variable and generate 3 new variables. Each will have binary numbers — 0 (if the category is not present) and 1(if category is present). This can be done using get\_dummies function of Pandas.*



* 1. ***Exporting Data:-*** *Final step is to convert data back into train and test data sets. Its generally a good idea to export both of these as modified data sets so that they can be re-used for multiple sessions. This can be achieved using following code*

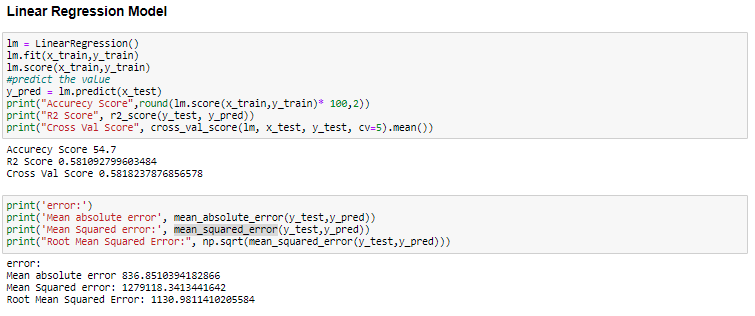


1. Model Building:-*. Create Train test data frame.*

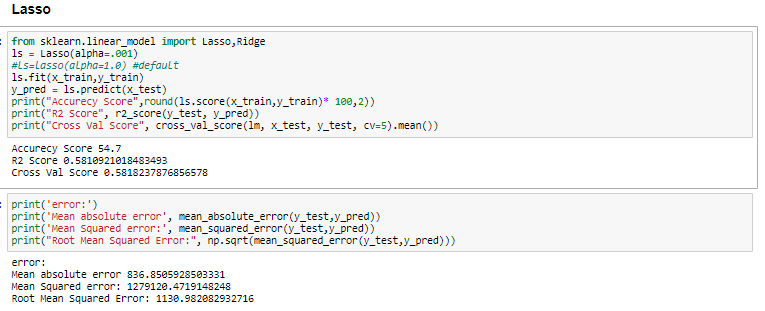


*Train data set contain feature and target value both, we will divide train dataset into x and y for model training and collection of model accuracy matrices, for this I will use below code.*

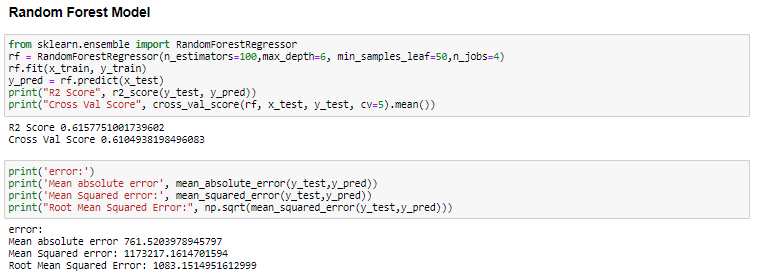


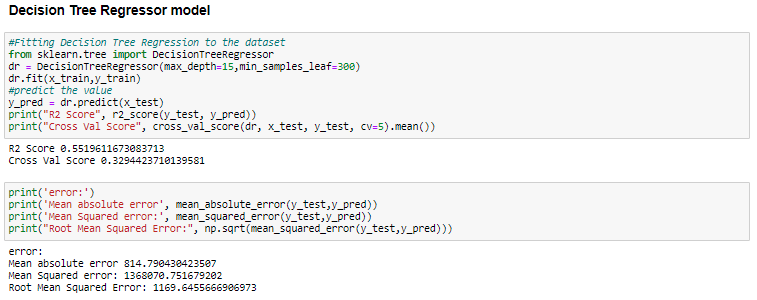
*5.1* 

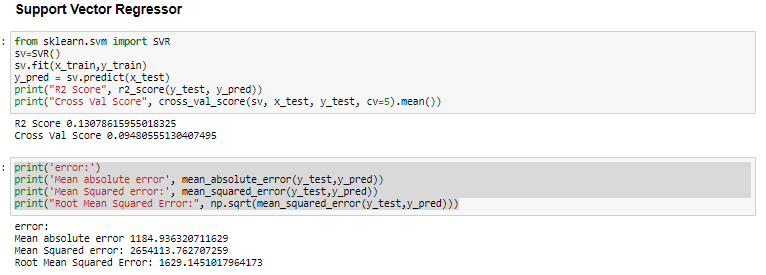
*Root Mean Squared Error for Linear regression is 1130.98 ~ 1131.*





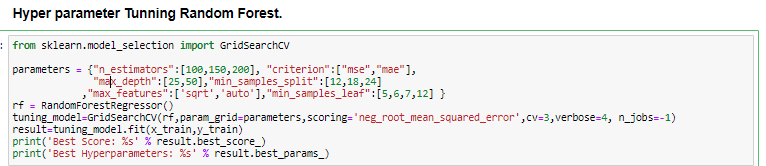






*Lowest RMSE is coming for Random forest regression model, We can say this is the best model to predict the price for different big mart sales items.*

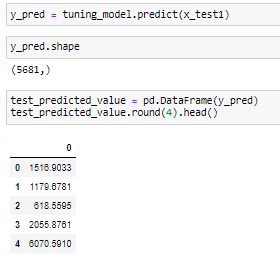
*Best Model is Random Forest Ression model, we will tune model to get more accuracy , for this we will us Grid search hyperparameter tuning.*



Best RMSE Score: -1091.0130320220771.

Best Hyperparameters: {'criterion': 'mse', 'max\_depth': 25, 'max\_features': 'auto', 'min\_samples\_leaf': 12, 'min\_samples\_split': 18, 'n\_estimators': 150}

***Predict Item\_Outlet\_Sales for Test Data.***



***Saving Best Model for further use.***

