**Big Data Mart Sales Problem**

**Introduction:**

The data scientists at BigMart have collected sales data for 1559 products across 10 stores in different cities for the year 2013. Now each product has certain attributes that sets it apart from other products.

# ATTRIBUTES OF VARIABLES

1. Item\_Identifier - A set of data elements permanently marked on an item that is globally unique and unambiguous and never changes in order to provide traceability of the item throughout its total life cycle.
2. Item\_Weight - This is the actual weight of the item
3. Item\_Fat\_Content - The amount of fat contained in item.
4. Item\_Visibility - The ability for shoppers to discover, identify, and engage with products.
5. Item\_Type - This define the contents of the item and the information that is stored about it.
6. Item\_MRP - Material Requirements Planning (MRP) is the system for calculating the materials and components needed to manufacture a particular item in an outlet/branch of a store.
7. Outlet\_Identifier - The Unique identifier for each outlet/branch of various stores.
8. Outlet\_Establishment\_Year - The year on which that an outlet of a store chooses to claim as its starting point.
9. Outlet\_Size - The space and capacity of an outlet of a store.
10. Outlet\_Location\_Type - The particular place or position an outlet is positioned.
11. Outlet\_Type - This refers to the nature and level of the store.
12. Item\_Outlet\_Sales - This is the amount of the exchange of a commodity for money generated by the outlet.

**Problem Statement:**

The objective is to create a model that can predict the sales per product for each store. Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.

* Supervised machine learning problem.
* The target value is be Item\_Outlet\_Sales.

**Data Analysis:**

The dataset has 8523 rows and 12 columns. We tried to understand the 12 columns of the dataset. The dataset has 12 columns out of which 7 are categorical data type, 4 columns are float data type & 1 is integer type data. The target column is Item\_Outlet\_Sales. This column is made of floating data and that is why it’s a regression problem.

There are 2 column which have missing values. We used isnull and sum function to identify the null values

They are

* Item Weight
* Outlet Size

The Outlet\_Size has 2140 missing data and Item\_Weight has 1463 missing values. Now, we have to work on filling these missing values so that these features are still useful in the prediction stage. There are many reasons why data can end up with missing values. For example:

1. The product wasn't weighed.
2. The data provider didn't include the outlet size of some products.

The traditional approach of imputing (the process of filling missing values) for columns is to use statistical measures such as mean, median, or mode. Numerical columns are filled with mean or median and categorical columns are filled with mode.

**Encoding:**

Also there were 7 categorical columns which have to be encoded for further analysis. So we used Ordinal Encoder and encoded these categorical columns. The 7 categorical columns are Item\_Identifier, Item\_Fat\_Content, Item\_Type, Outlet\_Identifier, Outlet\_Size, Outlet\_Location\_Type and Outlet Type.

**EDA (Exploratory Data Analysis)**-

The objectives of the EDA is to ensure:

1. Independent Features will have a normal distribution using Data Transformation to remove skewness
2. Datasets will have the same scale using Normalization and Standardization
3. Column mean will be zero
4. Standard Deviation of the data should be 1

Individual analysis of each column was done. We used two dimensional plots to understand the relationship between different variables.

The observations of EDA was:

* The Item\_identifiers have a lot of factors(1559) which has affected the visibility of its plot
* Majority of the sample items are low in fat.
* Majority of the sample items were fruits and vegetables
* Majority of the samples were selected from Outlet number 27 under Supermarket 3
* Majority of the samples were selected from Outlets with Medium Sizes
* Majority of the samples were selected from the Tier 3 Outlets
* Majority of the samples were selected from Supermarket Type1

We can see that all the features do not follow a normal distribution.

The normal distribution of these features; Item\_Identifier, Item\_Fat\_Content,, Item\_Type, Outlet\_Identifier, Outlet\_Size, Outlet\_Location\_Type, Outlet\_Type has no contribution to our Model Building since they are categorical data.

The normal distribution of the Item\_Outlet\_sales columns also has no contribution to our Model Building since it’s the Target variable

The scatter plot revealed there is a correlation between Item\_MRP & Item Outlet \_Sales. Increase in the item\_visibility can decrease the item outlet sales because it is having negative correlation. Item\_weight & Outlet\_Establishment\_Year have no correlation with Item\_Outllet\_Sales.

Correlation matrix showed: Item\_MRP is highly correlated with Item\_Outlet\_sales. Item\_visiblity is negatively correlated with Item\_Outlet\_sales. Item\_MRP has 57% correlation with Item Outlet sales. Item\_weight & Outlet\_Establishment\_Year have least correlation with Item\_Outllet\_Sales. Feature with Maximum correlation = '57%' Feature with Minimum correlation = '0.2%'

Multicollinearity: Multicollinnearity means two variables are explaining the same thing, meaning so one of them is not useful and therefore we have to drop one of them. From the heat map plotted we can see that the almost all pairs of features do not have noticeable correlation between them. We proceeded with VIF (Variance inflation factor) to figure out exactly if there is any Multicollinearity. After doing VIF, we found that there was no Multicollinearity between the columns.

Outliers: Outliers checking through Box plot was done and after visualisation we found that the item\_visibility feature possess outliers. We use zscore to confirm and to remove the outliers. If any Zscore value is above 3 we will treat it as an outlier and remove it so that we get a cleaned data set. After doing Zscore we found that 301 rows had outliers. So we dropped those rows and now the cleaned dataset has 8222 and 12 columns.

Skewness of the columns were also checked and it was found that the below features are skewed:

* Item visibility
* Outlet\_Establishment\_Year
* Outlet\_Type

**Power transformation** was applied which helped in bringing down the skewness of the above mentioned columns.

As all the columns have different measuring system we used Standard scalar for scaling the entire data. Standard scaler will bring mean to zero and standard deviation to 1

**Building Machine Learning Models.**

The dataset was split between independent variables which was named- x and dependent variable which was named -y. The dataset was then split into Training & testing data. Here we used the 80: 20 rule were 80% of the data was for training & 20% was for testing.

The Models which were used:

* Linear Regression
* KNeighbors Regressor
* SGD Regressor
* Support Vector Regressor(SVR)

For each of these models Cross validation was carried out. After this the best Model chosen was KNeighbors Regressor.

Once the best model was chosen based on the highest accuracy score of 63% we carried out the Hyper Parameter tuning for KNeighbors Regressor. For Hyper Parameter tuning we used Grid Search CV. The best parameter was n=10 and leaf\_size=10.

Finally we saved this model.