Sales Prediction Concept.

Dataset.

Dataset Background and Source.

The dataset used was put together by a group of data scientists from **Bigmart** in the year 2013. It contains records of sales data for 1559 products from 10 over BigMart stores located in different cities.

The dataset downloaded contains structured data and was downloaded from the website, **AnalyticsVidhya.com**.

Insights and Objectives.

Basing on the available attributes(in the dataset), which include a set of properties of the products and of the stores in which they are being sold, an analysis is to be done so as to identify which of these properties affect the amount of sales of a specific product more. This will enable the manipulation of these properties so as to increase on the sales of a given product.

To achieve this goal, we will build a predictive model and find out the sales of each product at a particular store.

Below is a list and description of the each of the twelve attributes of the dataset to be used during the analysis of sales.

a. Item_Identifier: This is a unique identifier for each product sold.
b. Item_Weight: This is the weight of a given product.
c. Item_Fat_Content: A categorical attribute that holds information on whether a product contains either a “low” or “regular” fat composition. It has only two values, which are, “Low Fat” and “Regular”.
d. Item_Visibility: This contains the total amount of visibility that a product is given on the shelves of the store in which it is being sold expressed as a percentage.
e. Item_Type: This contains information on which food category a specific product falls.
f. Item_MRP: This contains the maximum retail price of a specific product in a store.
g. Outlet_Identifier: This is a unique identifier for each store in which products are sold.
h. Outlet_Establishment_Year: This is the year in which the specific store was established.
i. Outlet_Size: This is the size of a specific store in terms of the area it covers.
j. Outlet_Location_Type: A categorical attribute which holds information about the type of city in which a store is located. It has three values, which are, “Tier 1”, “Tier 2” and “Tier 3”.

k. Outlet_Type: A categorical attribute which determines which type of establishment a specific store is. It has four values, which are, “Supermarket Type1”, “Supermarket Type2”, “Supermarket Type3” and “Grocery Store”.
l. Item_Outlet_Sales: This attribute contains the sales of a product in a particular store. This is the value that will need to be predicted.

Data Analytics Approach (Data Pipeline).

The steps that are to be followed during the data analysis process are explained below:

1. **Data Collection:**

This step was already done and completed. The dataset used was downloaded as a whole from the website, **AnalyticsVidhya.com**. This dataset was put together by a group of **BigMart Data Scientists** who collected the sales data for over 1559 products from 10 BigMart stores located in different cities in the year 2013. This data was already in a structured format and, therefore, did not require to be organized in rows and columns.

2. **Data Cleaning.**

In the dataset, there is expected to be a set of errors, duplicates or incomplete areas that need to be corrected. These will be corrected in the following ways:

- In the column of **Item_Fat_content**, there are several values that are being used to represent one value, for example, “Low Fat” and “LF” are both used to represent **Low Fat Content**. All the values that are referring to the same thing will all be unified to be represented in the same way, for example, all products with a low fat content will have their value in the **Item_Fat_Content** column as **Low Fat** and the rest will have this value as **Regular**.
- The missing values in the column of **Item_Weight** will be handled by computing the mean value of all the instances of the weight of the specific product that is missing a value and assigning the calculated mean value as the new value for the weight of the product with a missing weight value.
- The missing values in the column of **Outlet_Size** will be handled by computing the mode of the outlet size of the outlet type to which the specific outlet which is missing a value for its size belongs. This mode value will be used replace its missing value for the **Outlet_Size** column.
- If a product has a value of **0** (**zero**) in the column of **Item_Visibility**, the mean value of the visibility of all the instances of that particular product will be computed. This computed mean value will be used to replace the zero value of visibility for the product.

3. **Feature Engineering.**

Because some of the relevant libraries only accept numerical values, all the nonnumerical variable, with the exception of **Item_Identifier** and **Outlet_Identifier** will be converted into a binary, numerical format by using a **Label Encoder.**

4. **Models.**

The models and data visualization methods that are expected to be used are mentioned below:

- Multiple Regression.
- Bar graphs.
- Scatter plots.
- Correlation Analysis.

Picture Representation of the Data Pipeline.