Regression Project – Big mart sales prediction

Introduction

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.

Please note that the data may have missing values as some stores might not report all the data due to technical glitches. Hence, it will be required to treat them accordingly.

Data:

* Item\_Identifier - It is a unique product ID assigned to every distinct item. It consists of an alphanumeric string of length 5
* Item\_Weight - This field includes the wieght of the product
* Item\_Fat\_Content - This attribute is categorical and describes whether the product is low fat or not. There are 2 categories of this attribute: ['Low Fat', 'Regular']. However, it is important to note that 'Low Fat' has also been written as 'low fat' and 'LF' in dataset, whereas, 'Regular' has been referred as 'reg' as well
* Item\_Visibility - This field mentions the percentage of total display area of all products in a store allocated to the particular product
* Item\_Type - This is a categorical attribute and describes the food category to which the item belongs. There are 16 different categories listed as follows: ['Dairy', 'Soft Drinks', 'Meat', 'Fruits and Vegetables', 'Household', 'Baking Goods', 'Snack Foods', 'Frozen Foods', 'Breakfast', 'Health and Hygiene', 'Hard Drinks', 'Canned', 'Breads', 'Starchy Foods', 'Others', 'Seafood']
* Item\_MRP - This is the Maximum Retail Price (list price) of the product
* Outlet\_Identifier - It is a unique store ID assigned. It consists of an alphanumeric string of length 6
* Outlet\_Establishment\_Year - This attribute mentions the year in which store was established
* Outlet\_Size - The attribute tells the size of the store in terms of ground area covered. It is a categorical value and described in 3 categories: ['High', 'Medium', 'Small']
* Outlet\_Location\_Type - This field has categorical data and tells about the size of the city in which the store is located through 3 categories: ['Tier 1', 'Tier 2', 'Tier 3']
* Outlet\_Type - This field contains categorical value and tells whether the outlet is just a grocery store or some sort of supermarket. Following are the 4 categories in which the data is divided: ['Supermarket Type1', 'Supermarket Type2', 'Grocery Store','Supermarket Type3']
* Item\_Outlet\_Sales - This is the outcome variable to be predicted. It contains the sales of the product in the particulat store

Goal:

Our Goal is to predict Item\_outlet\_sales with regression model of choice.