**Machine Learning 1**

**Big Mart Sales Prediction**

**D19018**

**Manish Kumar**

**Medium link:-** [**https://medium.com/@manish.kumar\_61520/machine-learning-1-cf77ecd69fd?sk=5f25636418f3172f0170fa122c5ca016**](https://medium.com/@manish.kumar_61520/machine-learning-1-cf77ecd69fd?sk=5f25636418f3172f0170fa122c5ca016)

**Abstract:-**

**The e-Commerce company Big mart have collected 2013 sales data from 1559 Products from 10 different cities. Which is of total 8723 products.**  **Each outlet is in different city location , supermarket type, outlet size, weight of each item, outlet establishment year etc. The aim is to build a predictive model and predict the sales of each product at a particular outlet. The Price of product is dependent on sales of company our target is to make a Machine learning algoritham which will Predict the sales of Supermarket.**

**Introduction:-**

**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 predict the sales of each product at a particular outlet.Using this model, BigMart will try to understand the properties of products and outlets which play a key role in increasing sales.

**Objective:-**

**To predict the sales of each outlet in will play a virtual role in increasing role.**

**Step 1:-**

**Dataset description:-**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Item\_Identifier | Unique product ID |
| Item\_Weight | Weight of product |
| Item\_Fat\_Content | Whether the product is low fat or not |
| Item\_Visibility | The % of total display area of all products in a store allocated to the particular product |
| Item\_Type | The category to which the product belongs |
| Item\_MRP | Maximum Retail Price (list price) of the product |
| Outlet\_Identifier | Unique store ID |
| Outlet\_Establishment\_Year | The year in which store was established |
| Outlet\_Size | The size of the store in terms of ground area covered |
| Outlet\_Location\_Type | The type of city in which the store is located |
| Outlet\_Type | Whether the outlet is just a grocery store or some sort of supermarket |
| Item\_Outlet\_Sales | Sales of the product in the particular store. This is the outcome variable to be predicted. |

**Approach :-**

**Step 2:-**

**Imputation Missing Value**

In this dataset The variable ‘item Weight’ has 17.16% missing value and variable ‘Outlet Size’ has 28.27% missing value. Item Weight is given the weights of item type which are item Goods, breads, breakfast, canned, dairy, Frozen foods, fruits and vegetables, hard drinks, health and hygine, household, etc. As mentioned above this is a numerical variable . here as we can observe each item has different group of weights , for better imputation I imputed with grouped mean of each of the classes as for example Frozen foods is imputed with mean of frozen foods, breakfast is imputed with mean of breakfast so on and so far. Outlet size is imputed with variable outlet type. As item size has classes like small, medium, high is imputed with supermarket type1, supermarket type2, supermarket type2 .Becuase distribution of outlet type and outlet size is same.

**Step 3:-**

**Feature engineering**

We have unique product in these variable that is mentioned below:-

Item\_Identifier 1559

Item\_Weight 416

Item\_Fat\_Content 5

Item\_Visibility 7880

Item\_Type 16

Item\_MRP 5938

Outlet\_Identifier 10

Outlet\_Establishment\_Year 9

Outlet\_Size 4

Outlet\_Location\_Type 3

Outlet\_Type 4

Item\_Outlet\_Sales 3495

source 1

This tells us that there are 1559 products and 10 outlets/stores (which was also mentioned in problem statement). Another thing that should catch attention is that Item\_Type has 16 unique values. Let’s explore further using the frequency of different categories in each nominal variable. I’ll exclude the ID and source variables for obvious reasons.In the below the frequency of products that is mentioned below.

Fruits and Vegetables 1232

Snack Foods 1200

Household 910

Frozen Foods 856

Dairy 682

Canned 649

Baking Goods 648

Health and Hygiene 520

Soft Drinks 445

Meat 425

Breads 251

Hard Drinks 214

Others 169

Starchy Foods 148

Breakfast 110

Seafood 64

**Frequency of Categories for varible Outlet\_Size**

Medium 2793

Small 2388

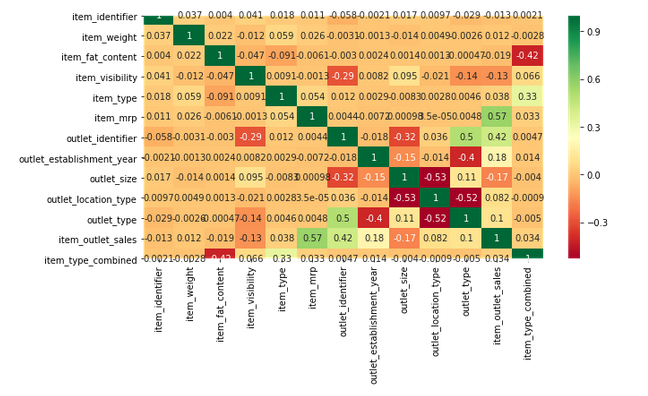
High 932

Distribution of these variables gives us good idea that we can create some new feature from these variables. Variable ‘Item Fat Content’ has classes like ‘Low Fat: low fat’, ‘Regular:reg’,’

So we created a new variable ‘Item\_Type\_Combined’.

Step 4:-

**Model Building:-**

After doing all the preprocessing steps we are now in model Building part as we dicussed above there is relationship between item\_mrp and item\_outlet\_sales so we use here linear regression to make this model. 

After Putting linear regression I got value of **R-square is 75% and root mean square error is .5176** that shows that our model is working fine.

Thank you