Big Mart Sales Prediction

By Juhi Aggarwal

Content

- Understanding Problem Statement & Data
- Data Exploration
- Data Cleaning
- Model Training
- Optimizing

Problem Statement

BigMart is a chain of e-commerce outlets. Provided is the data with certain product and outlet attributes. We had to make a prediction model after analysing and cleaning the provided statement.

The Data

Variable
Item_Identifier
Item_Weight
Item_Fat_Content
Item_Visibility
Item_Type
Item_MRP
Outlet_Identifier
Outlet_Establishment_Year
Outlet_Size
Outlet_Location_Type
Outlet_Type
Item_Outlet_Sales

Understanding data

Numerical Features:

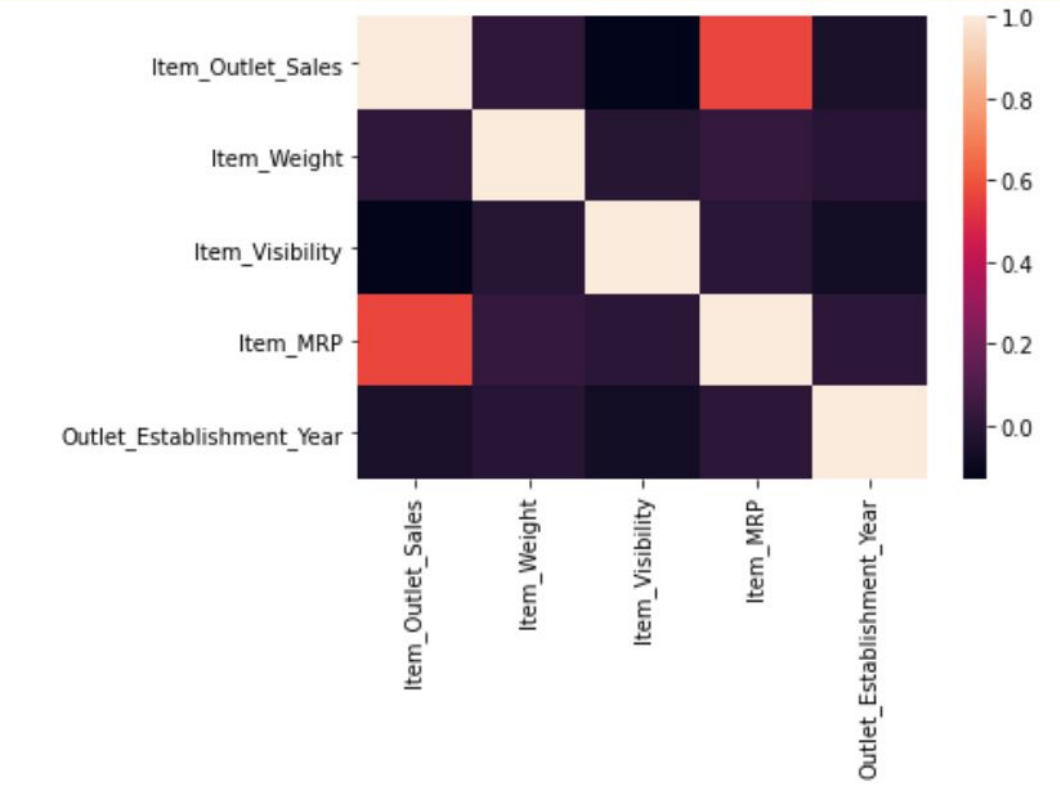
	Item_Weight	Item_Visibility	Item_MRP	Item_Outlet_Sales
count	7060.000000	8523.000000	8523.000000	8523.000000
mean	12.857645	0.066132	140.992782	2181.288914
std	4.643456	0.051598	62.275067	1706.499616
min	4.555000	0.000000	31.290000	33.290000
25%	8.773750	0.026989	93.826500	834.247400
50%	12.600000	0.053931	143.012800	1794.331000
75%	16.850000	0.094585	185.643700	3101.296400
max	21.350000	0.328391	266.888400	13086.964800

Understanding data

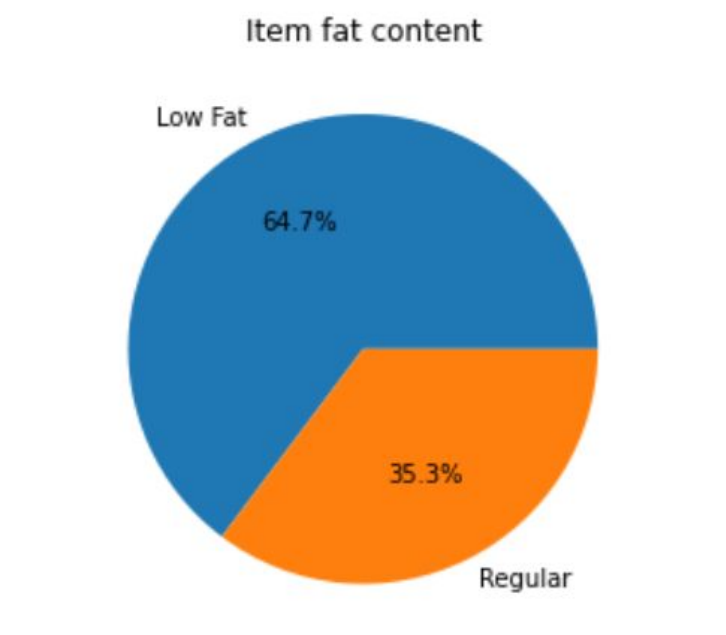
No. of Values of Categorical Features:

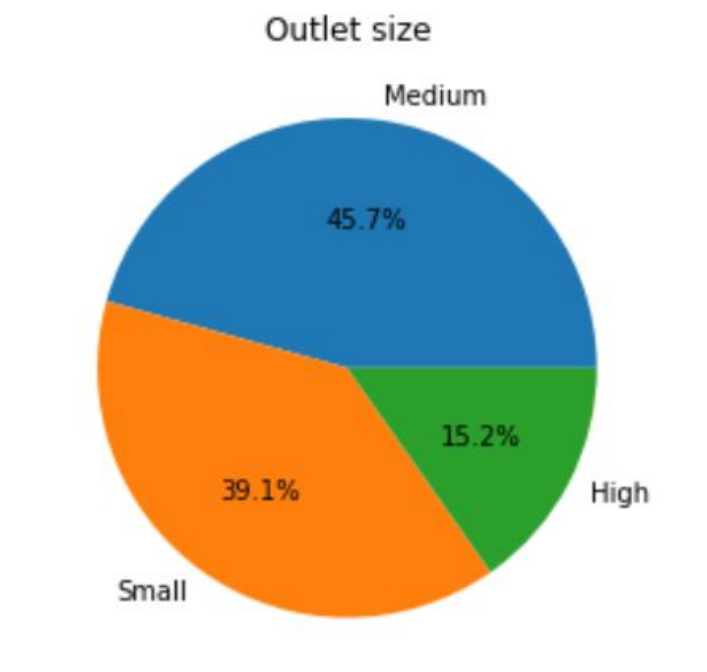
1	Item_Fat_Content	**2**
2	Item_Type	**16**
3	Outlet_Size	**3**
4	Outlet_Location_type	**3**
5	Outlet_Type	**4**
6	Outlet_Identifier	**10**
7	Outlet_Establishment_Year	**9**

Data Explorations



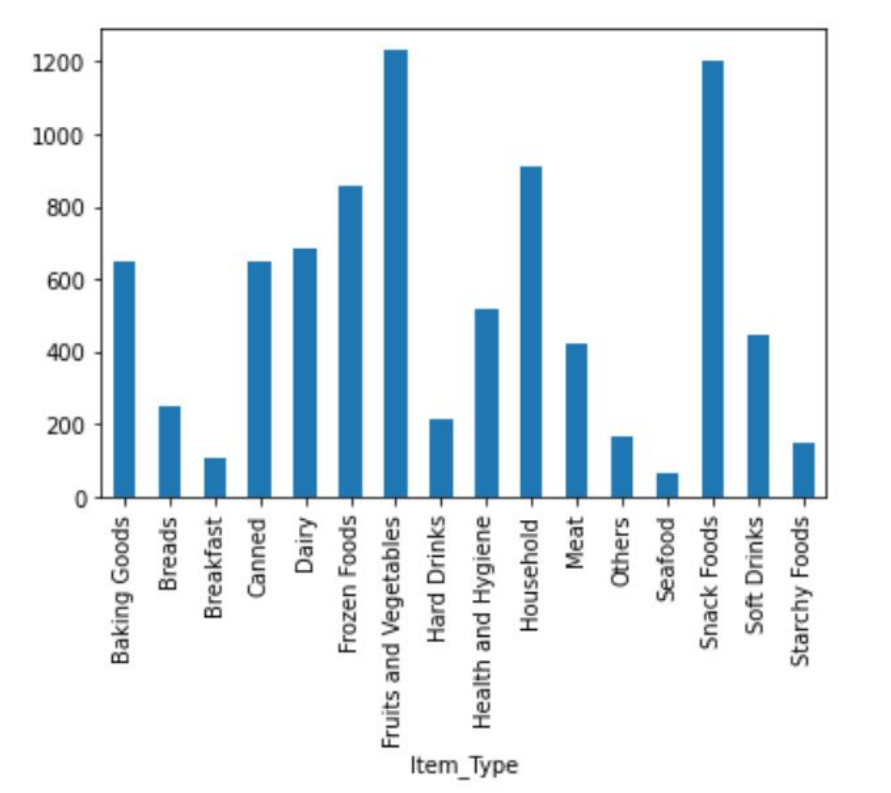
Data Explorations



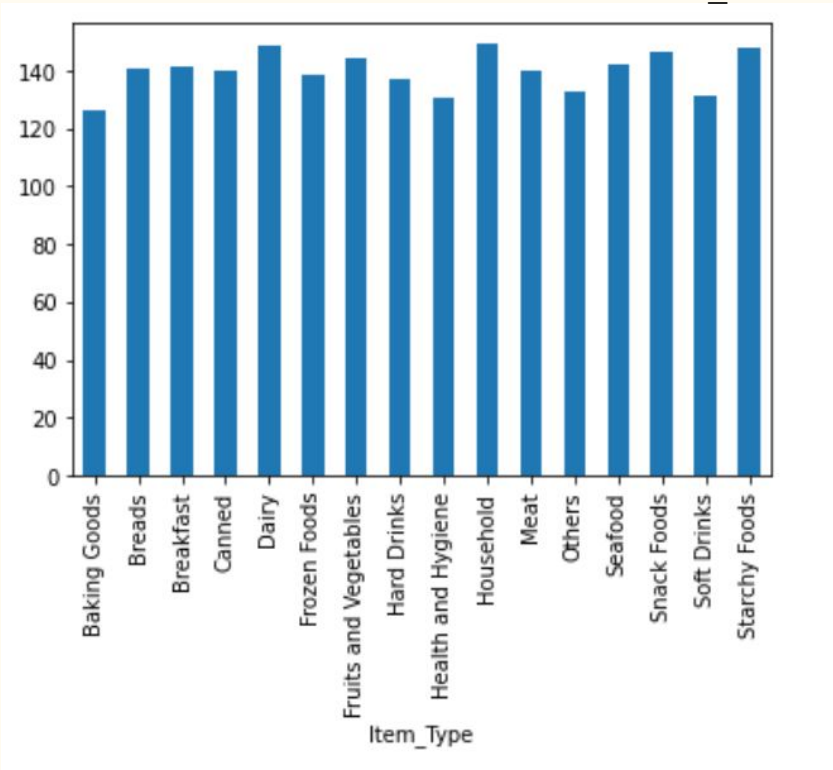


Data Explorations

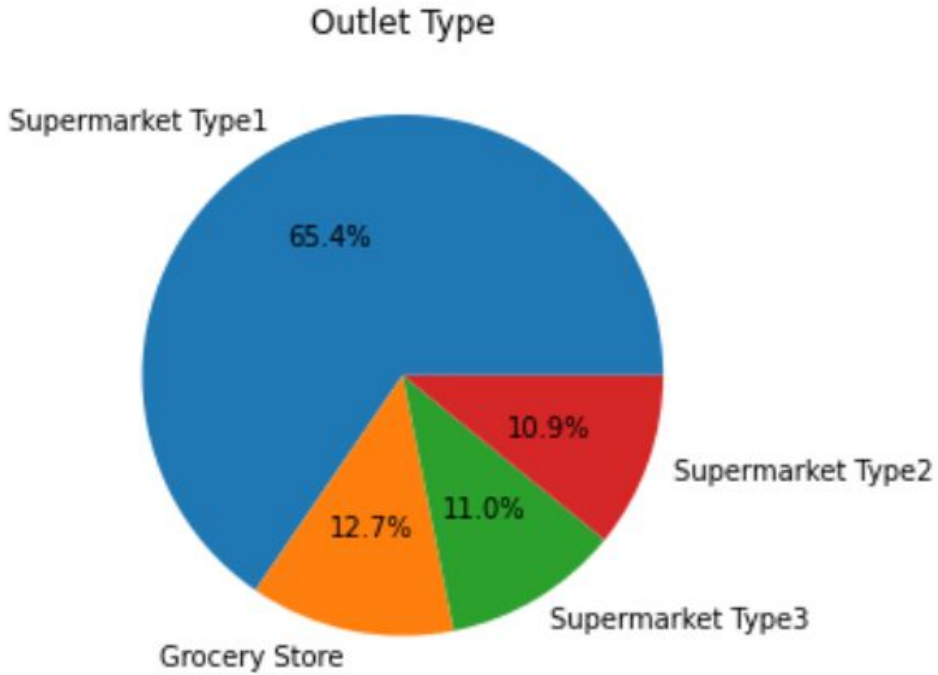
Count

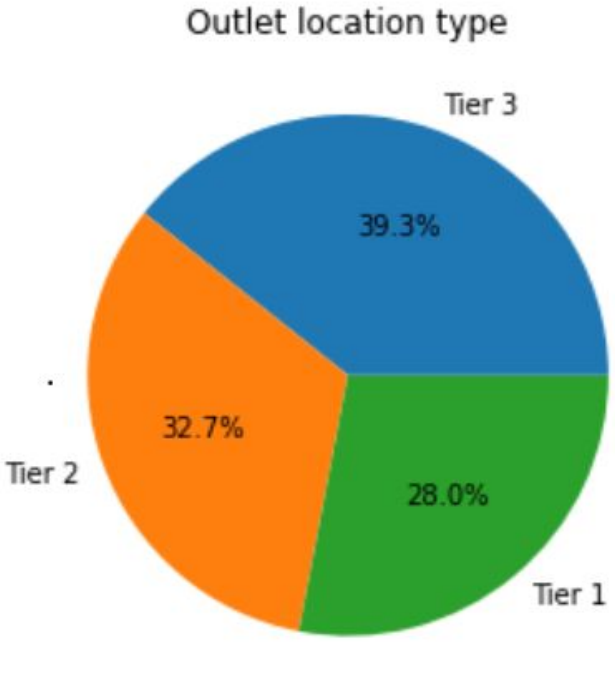


Mean_MRP

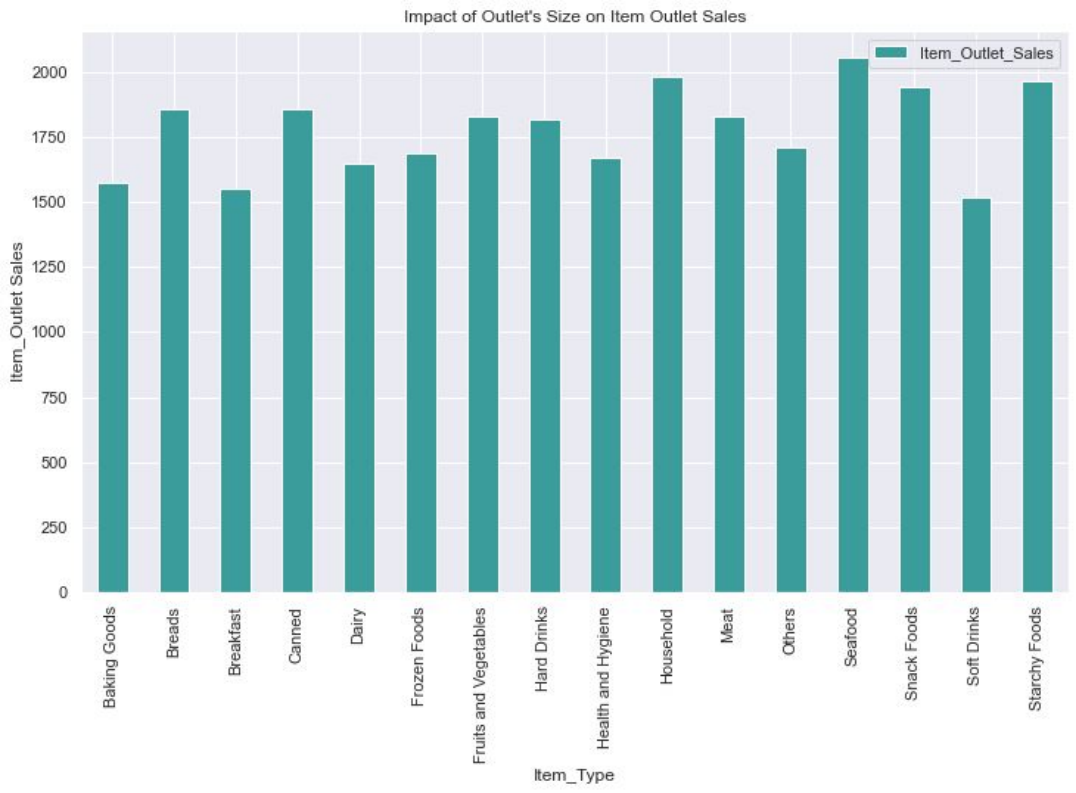


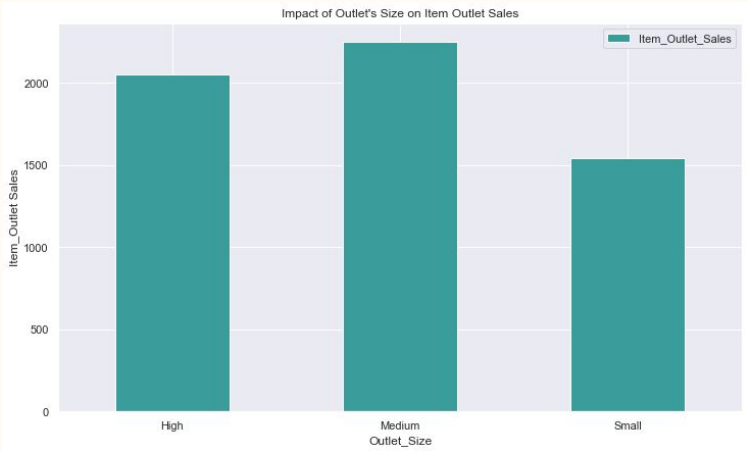
Data Explorations





Data Explorations





First Iteration

Data Cleaning:

The mess in Data:

A. **Uniform Fat Content;**
Replace the different upper and lower case values to two categories: Low Fat and Regular Fat

B. **Missing Item-Weights**
Filled by mean of their Item_types weights

C. **Dropping Item_Identifier, Outlet_Identifier columns.**

D. **Also Dropping Outlet_Size column due to many missing values.**

Model:

INDEX	MODEL	Mean_Squared_Error		r2_Score	
		Unscaled Data	Scaled Data	Unscaled Data	Scaled Data
1.	Linear Regression	1283941.064	0.422	0.561	0.550
2.	KNN	0.479	1599667.26	0.489	0.453
3.	Lasso Regression	1283899.75	0.420	0.561	0.552
4.	Random Forest Regression	1190071.579	0.391	0.593	0.583

Final Iteration

Data Cleaning

The mess in Data:

A. **Uniform Fat Content;**
Replace the different upper and lower case values to two categories: Low Fat and Regular Fat

B. **Missing Item-Weights**
Filled by mean of their Item_types weights

C. **Missing Outlet_Size**
By observing the data, Outlet_Type was mapped with Outlet_Size.

D. Retaining Outlet_Identifier

MODEL

INDEX	MODEL	Mean_Squared_Error		r2_Score	
		Unscaled Data	Scaled Data	Unscaled Data	Scaled Data
1.	Linear Regression	2.696e-26	1.753e-26	1.0	1.0
2.	KNN	11.466	0.134	0.837	0.998
3.	Lasso Regression	0.661	0.065	0.991	0.999
4.	Random Forest Regression	0.150	0.150	0.998	0.998

Conclusion with the main learnings

- DO NOT drop any column without proper reason!!
- Accuracy and evaluation metrics of the model is highly dependent on data we provide!!

Thank you!!