**15XDA7**– **MARKETING ANALYTICS**

|  |  |
| --- | --- |
| **Title** | **Big Mart Sales Prediction Datathon** |
| **Link to the contest** | <https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/> |
| **Domain** | **Business market sales prediction** |
| **Roll Number & Name** | **16PD05 – Aswath Rao & 16PD28 – Ridhanya** |
| **Tools used** | **PowerBI for visualization and Google Colab for scripts** |
| **References** | **Analytics Vidhya, Stackoverflow** |
| **Source code** | **github.com/Aswath98/MA** |

**Problem Statement**

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.

**Data**

We have train (8523) and test (5681) data set, train data set has both input and output variable(s). You need to predict the sales for test data set.

Data link - <https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/>

**Evaluation Metric**

Model performance will be evaluated on the basis of prediction of the sales for the test data (test.csv), which contains similar data-points as train except for the sales to be predicted. Root Mean Square Error value to judge response.

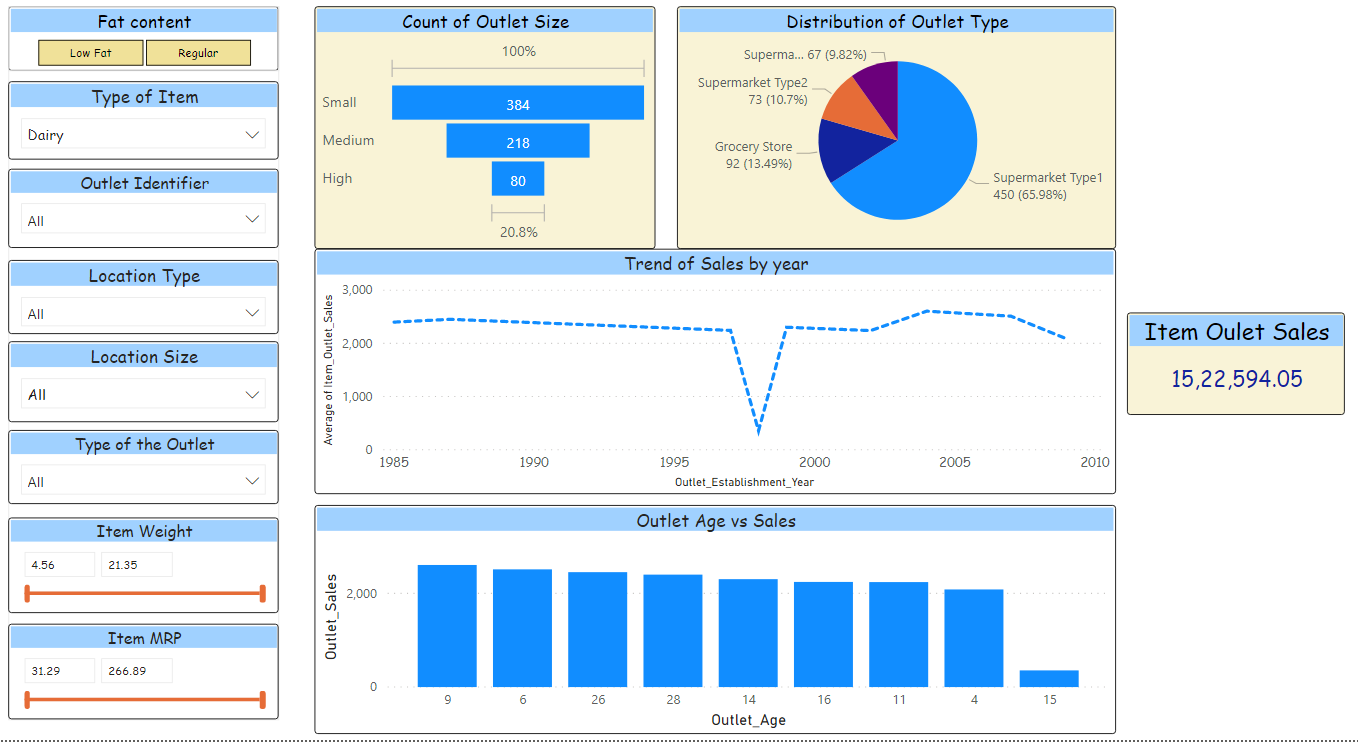
**We will explore the problem in following stages**

* Data Exploration – looking at categorical and continuous feature summaries and making inferences about the data.
* Data Cleaning – imputing missing values in the data and checking for outliers
* Feature Engineering – modifying existing variables and creating new ones for analysis
* Model Building – making predictive models on the data

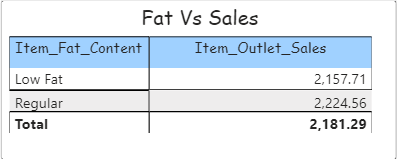
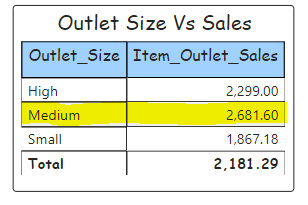
**Hypotheses**

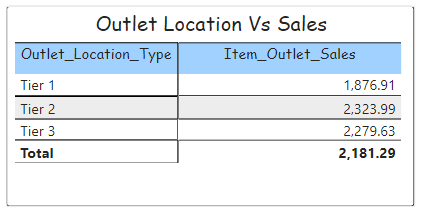
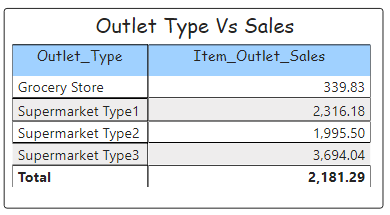
* Type of City - Stores located in Tier 1 cities should have higher sales because of the higher income levels of people.
* Store Capacity - Stores which are very big in size should have higher sales as people would prefer getting everything from one place.
* Location of Stores - Stores located within popular marketplaces should have higher sales because of better access to customers.
* Brand of products - Branded products should have higher sales because of higher trust in the customer.
* Visibility of Products - Products which are given bigger shelves in the store are likely to catch attention first and sell more.
* Item type – Daily usage products have more sales that products bought period wise
* Store establishment year – People will have a trust with stores that are running from long back, sales will be more from stores running for long time.

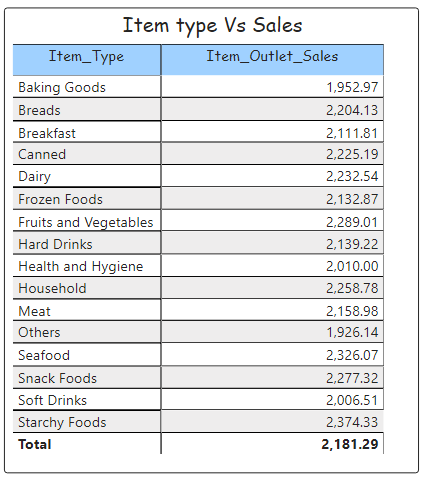
**Data Visualization**

****

**Marketing insights**

****

****

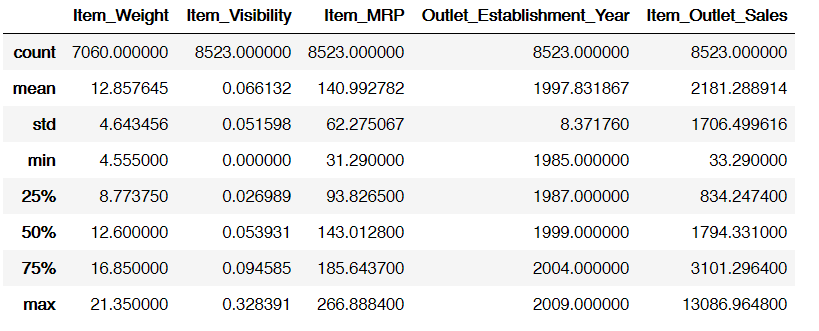
****

**Data Exploration**

The first step is to look at the data and try to identify the information about the parameters given

|  |  |
| --- | --- |
| **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 |  | | 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 particulat store. This is the outcome variable to be predicted. |

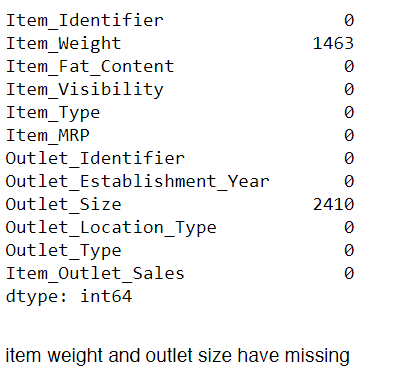
**Initial Data Description**



**Some observations**

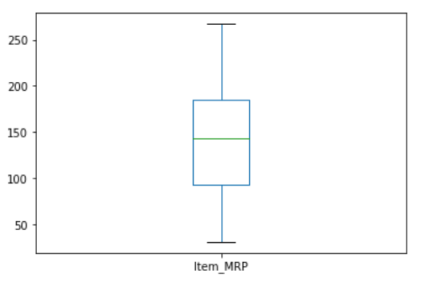
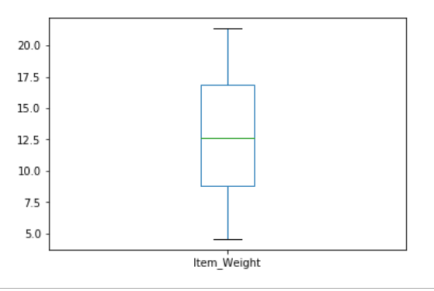
* Item\_Visibility has a min value of zero. This makes no practical sense because when a product is being sold in a store, the visibility cannot be 0.
* Outlet\_Establishment\_Years vary from 1985 to 2009. The values might not be apt in this form. Rather, if we can convert them to how old the particular store is, it should have a better impact on sales.

**Null values**

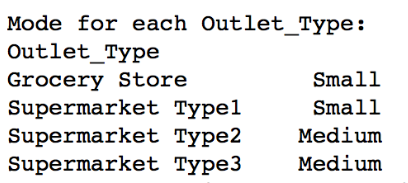


**Outlier Analysis**

Box plot confirms no outliers

* Item fat content is been modified into Regular and low fat categories
* The missing values of item weight is been replaced with mean of all weight of that item
* The missing values in Outlet type is filled accordingly



**Feature Engineering**

**Modify Item Visibility**

* Noticed that the minimum value here is 0, which makes no practical sense. Considering it like missing information and impute it with mean visibility of that product.

**Create a category of Item type**

Item Type variable has 16 categories. So it’s a good to combine them. One way could be to manually assign a new category to each. But there’s a catch here. If you look at the Item Identifier, i.e. the unique ID of each item, it starts with either FD, DR or NC. If you see the categories, these look like being Food, Drinks and Non-Consumables. So I’ve used the Item Identifier variable to create a new column.

**Determine the years of operation of a store**

We wanted to make a new column depicting the years of operation of a store

**Label encoding**

All the categorical columns are transformed into values using label encoding

**Modeling**

|  |  |  |
| --- | --- | --- |
| Model considered | RMSE Value | Analytics Vidhya Rank ( out of 3400 ) |
| Mean of all items | 1773.825 | 3200 |
| Mean group of items | 1598.6825 | 3175 |
| Mean group of Outlet | 1525.5149 | 3125 |
| Linear Regression | 1273.6146 | 2872 |
| Ridge Regression | 1276.1479 | 3000 |
| Decision Tree Regression | 1169.5697 | 1660 |
| Decision Tree with parameter tuning | 1155.6276 | 1200 |
| Random Forest Regression | 1153.7497 | 940 |
| LightGBM | 1167.3564 | 1500 |
| XGBoost Regression | 1152.5242 | 685 |
| Extra Tree Regression | 1228.1235 | 2000 |
| RandomForestRegression with RandomGrid | 1153.4839 | 804 |
| Tpot ( Auto ML ) | 1145.0937 | 103 |

**Conclusion**

Extra tree classifier with hyper parameter tuning is considered to be the best fit which gives a RMSE error of 1145.0937 and a rank of 103/3400

