Report on

BigMart-Analysis-and-Prediction-on-Sales-Data

Objective

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 of this data science project 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.

1. Importing Dependencies:

- The necessary libraries for data manipulation, analysis, and machine learning were imported.
- These include pandas, numpy, matplotlib, seaborn, LabelEncoder for encoding categorical features, and XGBRegressor for the machine learning model.

2. Data Collection and Analysis:
The dataset was loaded using pandas, and an initial exploration was performed.

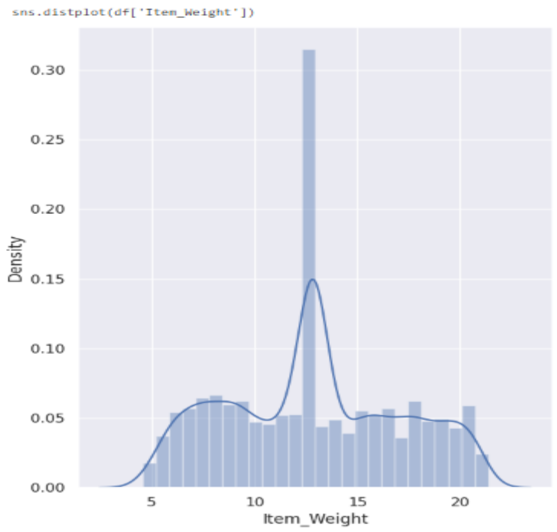
- The dataset contains 8523 entries and 12 columns.
- Data types include float64, int64, and object (categorical).
- Two features, 'Item_Weight' and 'Outlet_Size', have missing values.

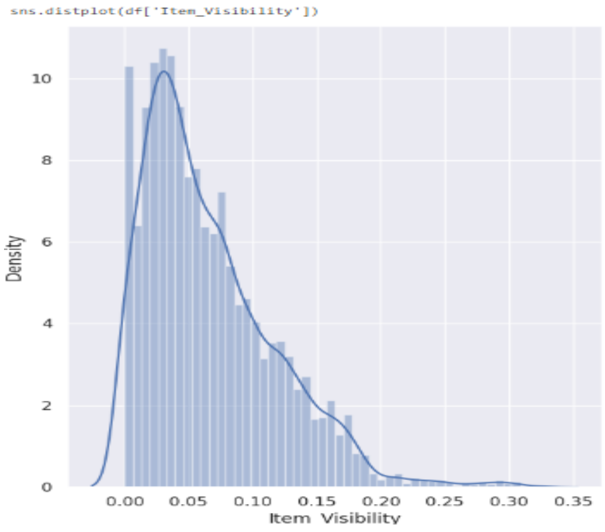
3. Handling Missing Values:

- Missing values in 'Item_Weight' were filled with the mean value of the column.
- Missing values in 'Outlet_Size' were filled with the mode (most frequent value).

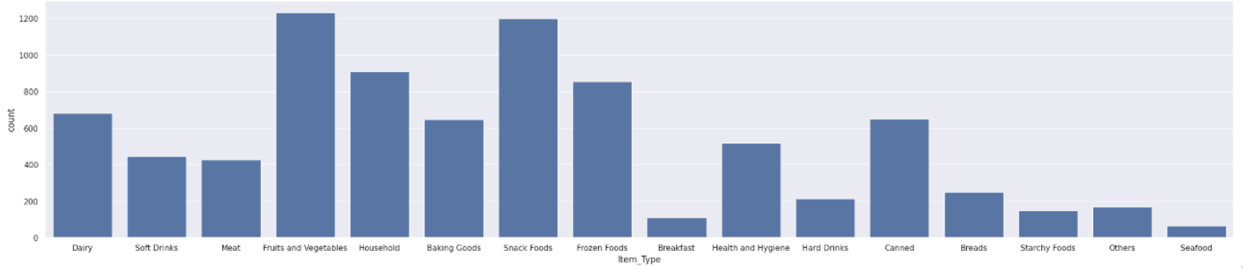
4. Data Analysis and Visualization:
Exploratory Data Analysis (EDA) was performed to understand the distribution and characteristics of various features.

- Distributions of 'Item_Weight', 'Item_Visibility', 'Item_MRP', and 'Item_Outlet_Sales' were visualized using seaborn and matplotlib.





Count plots were created to visualize the distribution of 'Outlet_Establishment_Year', 'Item_Fat_Content', 'Item_Type', and 'Outlet_Size'.



5. Label Encoding:
Categorical features were encoded using ‘encoder.fit_transform’ to convert them into numerical values for machine learning model training.

6. Splitting Features and Targets:
The dataset was divided into features (X) and the target variable (y). The target variable is 'Item_Outlet_Sales'.

7. Splitting the Data:
The data was split into training and testing sets using the ‘train_test_split’ function from sklearn. The training set comprises 70% of the data, while the testing set comprises 30%.

8. Machine Learning Model Training:
An XGBoost Regressor model was used for training. XGBoost is an ensemble learning algorithm known for its high performance.

- The model was trained on the training data.
- Predictions were made on the training data.
- The obtained R-squared value is approximately 0.888, indicating a good fit of the model to the training data.

Conclusion:
The machine learning model has been successfully trained on the BigMart sales dataset. The model can now be further evaluated on the testing dataset to assess its generalization performance. Additionally, further model tuning and feature engineering can be explored to improve predictive accuracy.

This report provides an overview of the key steps taken in the BigMart sales prediction project, from data exploration to model training.

Data Visualisation Using Tableau:

Key Metrics:

Total Sales: The total sales of the product amounts to **18.59** million.

Average Sales: The average sales of each product are within the range of **2.18K**.

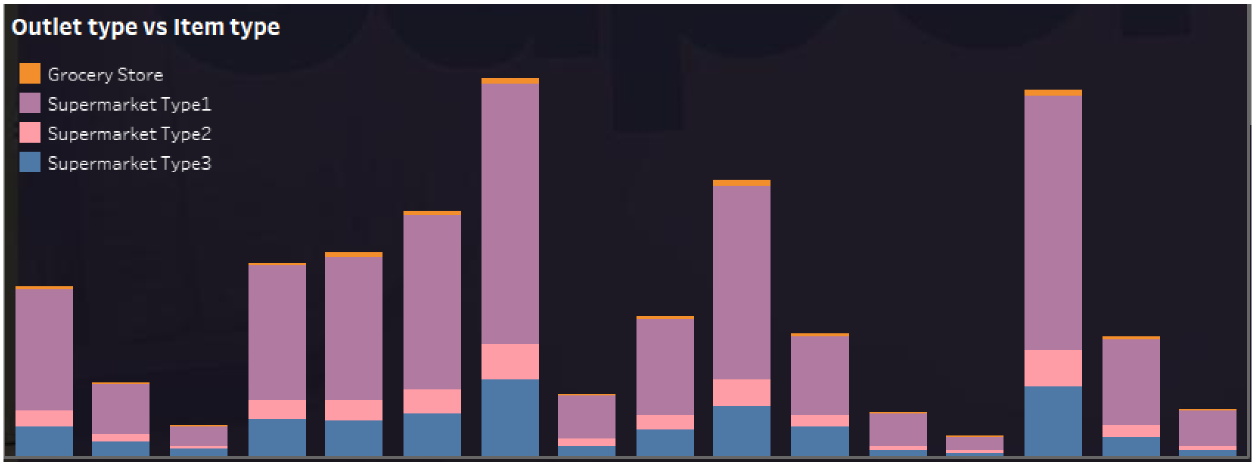
Key Performance Indicators:

Outlet type vs Item type: This indicator evaluates the sales of different items in comparison to the Outlet type.
The outlet type id categorised into 4 segments namely

1. Grocery Store.
2. Supermarket Type 1.
3. Supermarket Type 2.
4. Supermarket Type 3.

We can understand the following:

- Among **Grocery stores** the highest sales is recorded for **Snack foods** reaching a total of **51,596** units
- **Supermarket Type1** positions itself as a specialized provider in the retail of high-quality **Fruits and Vegetables**, with sales totaling **1,931,958** units in this market segment.
- **Supermarket Type 2** experienced peak sales in the realm of **Snack Foods**, marking an impressive figure of **278,715** units.
- **Supermarket Type3** leads the market in the retail of top-quality **Fruits and Vegetables**, achieving sales of **576,028** units.

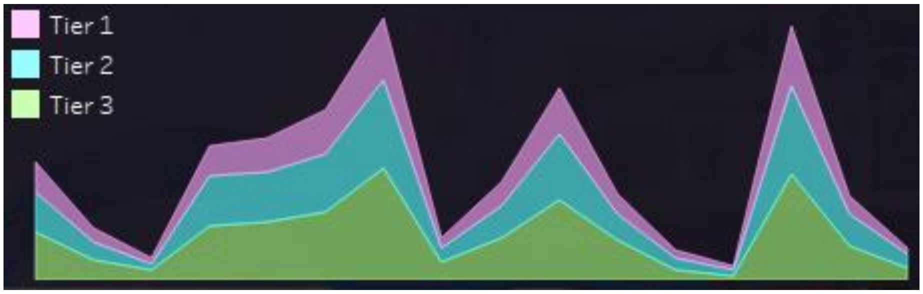


Outlet Location Type vs Item Type: This Parameter evaluates the sales of different items based on the location the outlets are located.

The location types are categorised as:

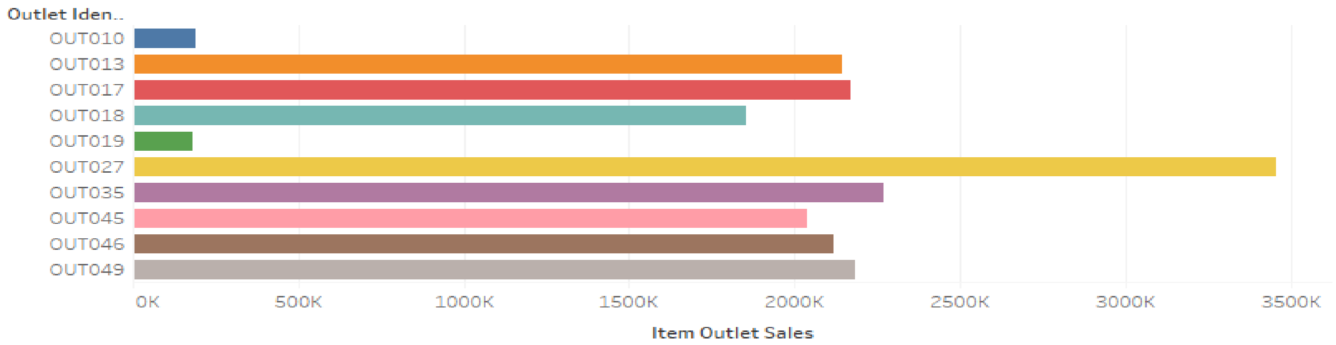
1. Tier 1
2. Tier 2
3. Tier 3

The study shows the following findings:



- *Fruits and Vegetables, Household products* and *Snack Foods* are the highest performing products across the all Tier 1,2 and 3 cities.
- *Breakfast* items, *Hard Drinks and seafoods* are lowest performing items across all the cities.

<u>Outlet Identifier vs Item Outlet Sales</u>: This parameter indicated sales of items for each outlet having outlet id. There is a total of 10 unique outlet id.

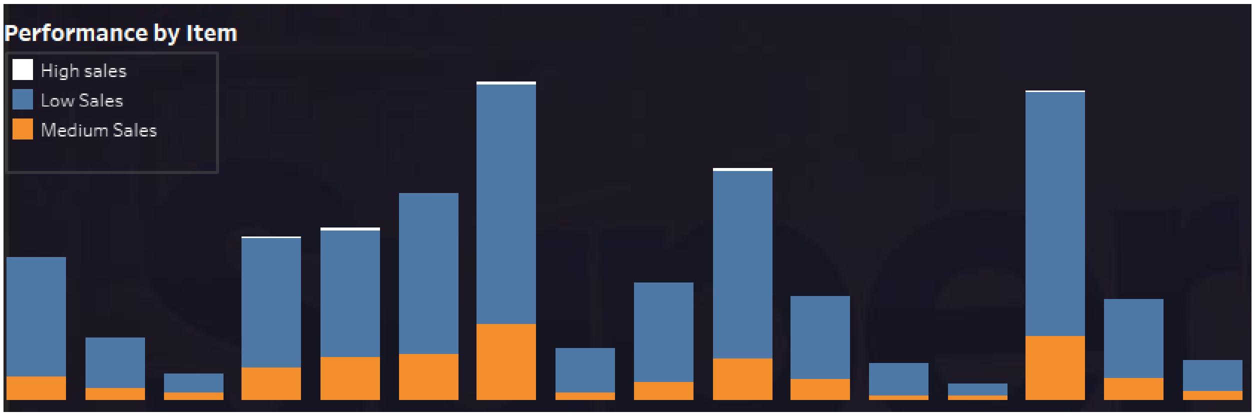


- OUT027 emerges as the standout outlet, boasting an impressive sales value of 3,543,926 units.
- OUT035, securing a commendable sales figure of 2,268,123 units.
- OUT017, maintaining a strong position with a sales value of 2,167,415 units.

<u>Item Performance</u>: This performance indicator evaluates the margin of sales for different items. There are 16 unique items in the dataset.

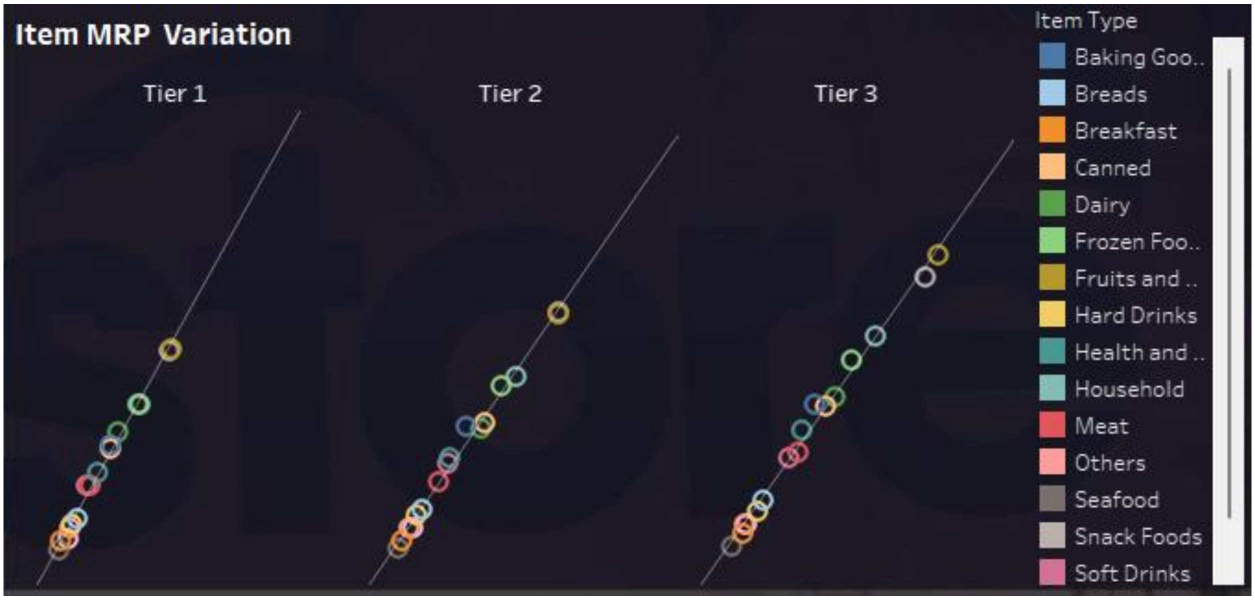
We have used ‘calculated Field’ to apply a condition for High, Low and Medium sales

```
IF [Item Outlet Sales]>=10000 then "High sales"
elseif [Item Outlet Sales]>=5000 then "Medium Sales"
else "Low Sales"
END
```



- Among all the products **Fruits and Vegetables, Household, Snack, Dairy, Canned** categories recorded high sales.
- Breakfast, Seafood, Hard drinks recording low sales.
- Other categories revealing medium to low sales.

__Item MRP Variation:__ This illustrates the range of MRP for various items over Tier1, Tier2, Tier3 cities.



Among products **Fruits Vegetables, Snack Foods and Household** items emerge as the costliest items across all cities

For Tier 1 cities, the Maximum Retail Price (MRP) values for these products are as follows:
Fruits: $332
Vegetables: $328
Snack Foods: $245

In Tier 2 cities, there is a noticeable increase in MRP values:
Fruits: $386
Vegetables: $388
Snack Foods: $287

Tier 3 cities witness a further increment in MRP:
Fruits: $477
Vegetables: $443
Snack Foods: $352

Navigating the Future:

Promotion of High-Performing Items:

- Focus on promoting high-performing items such as Fruits and Vegetables, Household products, and Snack Foods.
- Develop targeted marketing campaigns and promotions to boost sales for these categories.

Outlet-Specific Strategies:

- Tailor strategies for each outlet type based on their strengths. For instance, highlight the specialization of Supermarket Type 1 in high-quality Fruits and Vegetables.

Optimize Inventory Management:

- Analyse sales performance across outlet identifiers to ensure optimal inventory management.
- Stock popular items more abundantly and consider seasonal variations in demand.

Collaboration with Outlets:

- Collaborate closely with outlets that have shown strong performance, such as OUT027 and OUT035.
- Explore joint promotional activities to mutually benefit both the company and the outlets.

Investment in Low-Performing Categories:

- Evaluate the reasons behind the lower sales in categories like Breakfast items, Hard Drinks, and Seafoods.
- Consider investing in marketing, product improvement, or repositioning strategies for these categories.

Product Diversification:

- Explore opportunities for product diversification, especially in categories with medium to low sales.
- Introduce new items or variations to attract a broader customer base.

Conclusion:

In conclusion, the analysis of key metrics and performance indicators provides valuable insights for the company to enhance its business strategies and drive future growth. The company has achieved a commendable total sales figure of $18.59 million, with an average sales per product of approximately $2.18 thousand.

- Tailoring strategies for different outlet types, leveraging location-based marketing, and optimizing inventory management are key areas for improvement.
- The detailed analysis of outlet identifiers highlights the standout performance of OUT027, OUT035, and OUT017, suggesting potential areas for increased focus or collaboration.

By understanding customer preferences, strategic pricing, product diversification, and data-driven decision-making and continuously adapting strategies the company can stay competitive and capitalize on emerging trends. Additionally, addressing the performance of lower-performing categories and exploring opportunities for improvement will contribute to a more balanced product portfolio.

Overall, the company is well-positioned to drive future business growth by implementing targeted marketing campaigns, optimizing inventory, and adapting strategies to meet the unique characteristics of different outlets and locations. Regular monitoring of performance metrics and flexibility in decision-making will be essential in navigating the dynamic retail landscape and ensuring sustained success in the market.