

* Title: Bigmart sales Analysis.

* Problem statement:

For data comprising of transaction records of a sales store the data has 8523 rows of 12 variables. Predict the sales of store.

* Objective: Learn Sales Analysis.

* Outcomes: Student will be able to analyse sales dataset and make predictions.

* Theory:

The data scientist 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.

The dataset consists of following attributes

Variable	Description
Item-identifier	Unique product ID.
Item-weight	weight of product.
Item-Fat-Content	whether product is low fat or not.
Item visibility	The % of total display area of all products in a store allocated to particular product. The category to which product belongs.
Item-MRP	Maximum retail price of an item.
Outlet-Identifier	unique store ID.
Outlet-Establishment-year	The year in which store was established.
Outlet-Size	The size of store in terms of ground area covered.
Outlet-Location-Type	The type of city in which the store is allocated.
Outlet-Type	Whether the outlet is just a grocery store or some supermarket.
Item-Outlet-sales	Sales of the product in particular store. This is outcome variable predicted.

* Valuation Metric:

Root mean Square Error Value:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\text{Prediction}_i - \text{Actual}_i)^2}{N}}$$

where,

N = Total no. of observations.

predicted: The response covered by user.

Actual: Actual values of sales.

The problem can be explored in following steps.

(1) Hypothesis generation:

Understanding the problem better by brain-storming possible factors that can impact outcome.

(2) Data Exploration: Looking at Categorical and Continuous feature summaries & making inferences about data.

(3) Data Cleaning: Imputing missing values in data and checking for outliers.

(4) Feature Engineering:-

Modifying existing Variable and Creating new ones for analysis.

(5) Model Building

Making predictive model on data

Different sets of hypothesis are created for stores and products:-

Store level hypothesis include:-

- (1) City type. (2) Population Density.
- (3) store Capacity (4) Competitors (5) marketing
- (6) Location.
- (7) Customer behavior.
- (8) Ambiance.

Product level hypothesis include:-

- (1) Brand
- (2) Packaging
- (3) Utility
- (4) Display area
- (5) Visibility in store.
- (6) Advertising
- (7) promotional offer.

All missing values are treating outliers are removed in data cleansing process. though outlier removal is very important in regression techniques.

Advanced tree based algorithms are impervious to outliers.

Predictive models can be built using linear regression, decision tree or random forest etc can be used.

* Conclusion:

In this way big mart sales prediction is done using linear regression, random forest and decision tree classifier.