BIG MART SALES PREDICTION SYSTEM

Name : J. Thusyanthan

Index No : 18/ENG/109

Course Code : CO3302

Course Name : Computer Engineering Project

Background

Bigmart is a supermarket chain which is located nearly at every megacity. The sales of Bigmart are very crucial, and data scientists study those patterns per product and per store to decide about the new centers. Using machine learning to predict Bigmart sales enables the data scientist to do so, as it studies the various patterns per store and per product to give accurate results[8].
Here our problem is many Big mart sales businesses earning money at start and after some years some of their branches have net loss. Because of they haven't much knowledge in predicting which product will sale high in future demanded products and which branch having high demand.

But this solution is to help Big marts to increase their sales by finding the best demanded products and demanded areas. This helps the big mart to run the business continuously and in future can open branches in particular areas and increase the production of demanded products.

To find out what role certain properties of an item play and how they affect their sales by understanding Big Mart sales." To help Big Mart achieve this goal, a predictive model can be built to find out for every store, the key factors that can increase their sales and what changes could be made to the product or store's characteristics.

This is an easily scalable model to provide detailed info and accurate predictions for sales volume for different type of products and this solution can be used for start-ups and sales forecasting.

The assumptions or factors affecting the sales of a store are mentioned below[3],

- City type: Generally, stores situated in Tier 1 cities or urban areas have higher sales as the income level of people are high.
- Population Density: Similarly, densely populated area stores have higher sales because of more demand.
- Store Capacity: Big size stores have higher sales as they act like one-in-all.
- Shops: customer would prefer getting everything from one place
- Competitors: Establishment year affects the sale volume due to legacy effect competition.
- Marketing: Marketing and advertising immensely affects sales as it increases its visibility and catchy slogans stay with customers for a long period effecting the sales.

- Location: Similarly, popular marketplaces or better located stores have higher sales due to easy access.
- Customer Behavior: Stores having right set of products to meet the local needs will have higher sales.
- Brand: Customer always trust branded products, so they have a higher sales.
- Utility: Products used daily have a higher tendency to be sold as compared to the specific use products.
- Display Area: Products kept in bigger shelves inside the store are likely to make attention first and sell more.
- Visibility in Store: The location of product in a store will impact sales. One which are right at entrance will catch the eye of customer first rather than the ones in back.
- Display: Better display of products in the store makes higher sales in most cases.
- Promotional Offers: Products accompanied with attractive offers and discounts sell more.

Objectives

The aim of this project is to conduct a complete study of the sales prediction for big mart by analyzing the data set and develop models using machine learning algorithms for sales prediction and predict the sales of each product at a particular outlet.

Sales prediction includes aims such as demand of the products, demanded areas for high sales, reduce the wastages, increase the profit to run the business continuously and avoid the losses.

Sales Prediction is important for predicting the future demand. Business enterprises focus on sales prediction. Prediction helps the business to work according to plan and cut down wasteful expenditure. As a result goods can be offered at a fair price. It is easy to control the performance. It enables the business to decide whether to add a new product or to drop unsuccessful one that is not demand in the market.

Our goal is to help Big marts to increase their sales by finding the best demanded products and demanded areas. This helps the big mart to run the business continuously and in future can open branches in particular areas and increase the production of demanded products

By using this sales prediction model Big marts can move their business by minimizing the risk in the business and move forward to the future by success journey.

Milestones

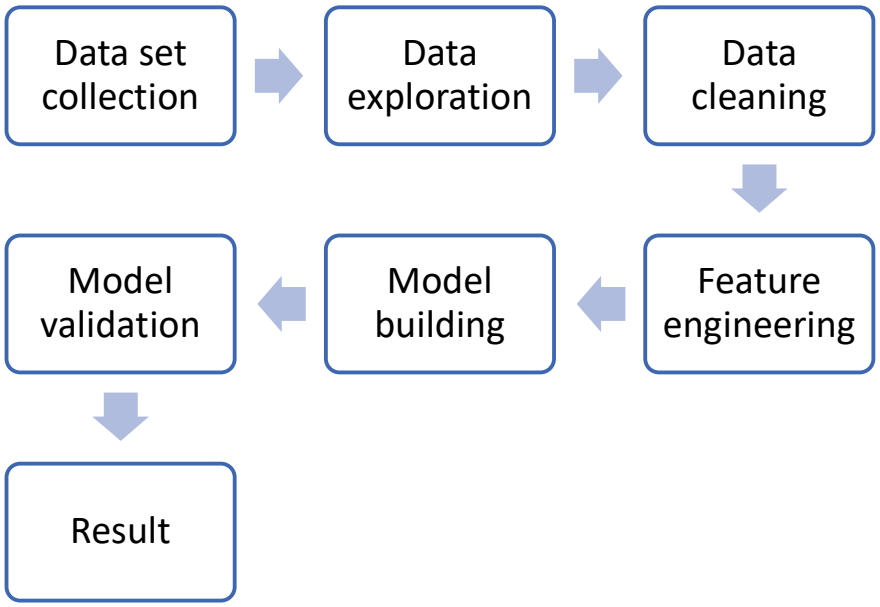
Milestones of the project are shown below,

1. Collection of data.
2. Analyzing, preprocessing and feature extraction of the data set by handling the missing data and outliers.
3. Model building using various algorithms such as Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, XGB Regressor, Ada Boost Regressor and Decision tree regressor.
4. Checking the accuracy of each algorithms to compare the accuracy of the model.
5. Design and implementation of predictive interface .

Table 1- Timeline of the project

Activity	Week									
	1	2	3	4	5	6	7	8	9	10
Case study										
Data collection										
Pre processing										
Algorithm design										
Model development										
Prediction phase										
Interface development										
System testing										

Methodology



Description

Data set collection

Big Mart's data scientists collected sales data of their 10 stores situated at different locations with each store having 1559 different products as per data collection. Using all the observations it is inferred what role certain properties of an item play and how they affect their sales.

Attribute	Description
Item identifier	Product Unique id
Item weight	Product Weight
Item fat content	Fat content of the product
Item visibility	Total percentage of product visible to customer from the store
Item type category	Product belong to which category
Item MRP	Price of the product
Outlet identifier	Store unique id
Outlet Establishment year	Established year of the store
Outlet size	Store size
Outlet location type	Type of located cities
Outlet type	Supermarket sort
Item outlet sales	Product sales in particular area

Data exploration

Here we classify the data from the hypothesis vs available evidence which indicates that the size of the outlet attribute and the weight of the object faces the question of missing values, as well as the least value of Object view is Zero which is not feasible. The Item type attribute contains 16 specific values. Variable outlet sales have been skewed positively. So, a log is applied on Answer Variable to skewedness.

Data cleaning

Data set should be cleaned and pruned to handle the missing values and to avoid data duplications. Here item weight and outlet size have missing values so we fill those with median and mode of them respectively. Also remove the outliers.

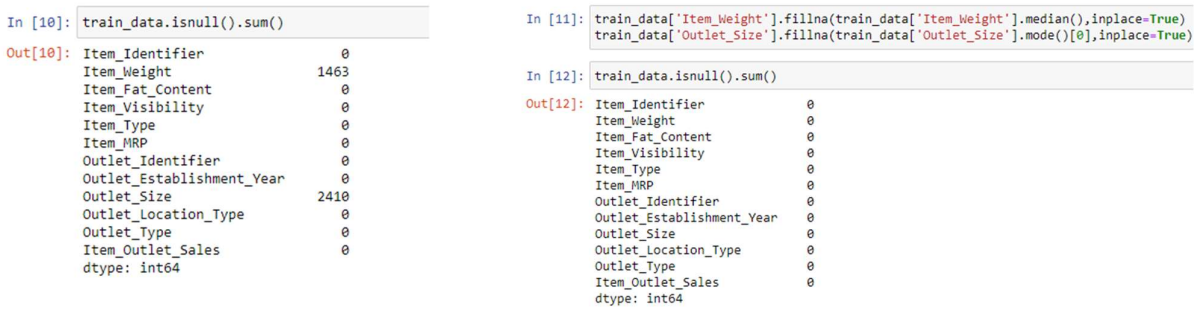


Figure 1- Removing the null values

Feature Engineering

Feature engineering is all about converting cleaned data into predictive models to present the available problem in a better way. During data exploration, some noise was observed. A few created features can be combined for the model to work better. Feature engineering phase converts data into a form understandable by the algorithms.

Figure 2- Combining the noise data

Model building

After Feature engineering, the processed data is used to give accurate results by applying multiple algorithms. A model is a set of algorithms that facilitate the process of finding relation between multiple datasets. An effective model can predict accurate results by finding exact insights of data.

- Linear regression- Linear regression algorithm tries to predict the results by plotting the graph between an independent variable and a dependent variable that are derived from the dataset. It is a general statistical analysis mechanism used to build machine learning models. The general equation for linear regression is

  $$Z = a + bE$$

  Where, Z is the dependent variable and E is independent variable.
- Random Forest- Random Forest Algorithm is used to incorporate predictions from multiple decision trees into a single model. This algorithm uses bagging mechanism to create a forest of decision trees. It the incorporates the predictions from multiple decision trees to give very accurate predictions.
- XG Boost algorithm- The XG Boost algorithm is developed using Decision trees and Gradient boosting. This algorithm stands on the principle of boosting other weaker algorithms placed in a gradient decent boosting framework. This approach works very accurately beating almost all other algorithms in providing accurate prediction. It can be defined as an extension to Gradient Boosting algorithm

In addition we use decision tree, ada boost regressor, lasso, ridge and gradient boosting regressor also.

Model validation

After training and testing the data sets, creating a predictive interface using higher accuracy model validation phase. All models received features as input, which are then segregated into training and test set. The test dataset is used for sales prediction.

Tools and resources that will be used throughout the project are as follows,

- Jupyter notebook
- Visual studio code

- Flask framework
- GitHub
- YouTube

The following programming Languages will be used,

- Python
- Html
- CSS
- Java script

Results

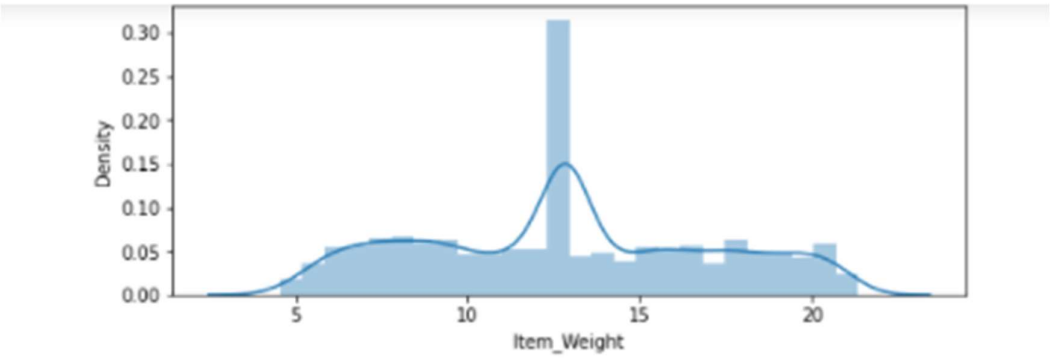


Figure 3- Density of item weight

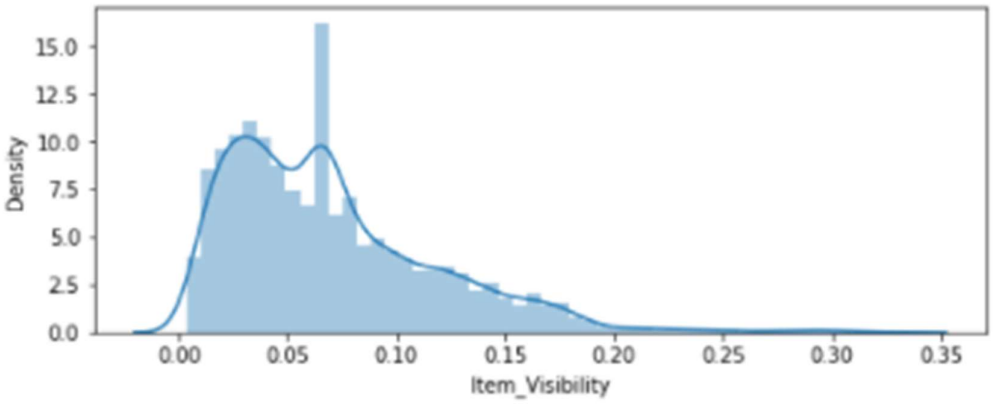


Figure 4- Density of item visibility

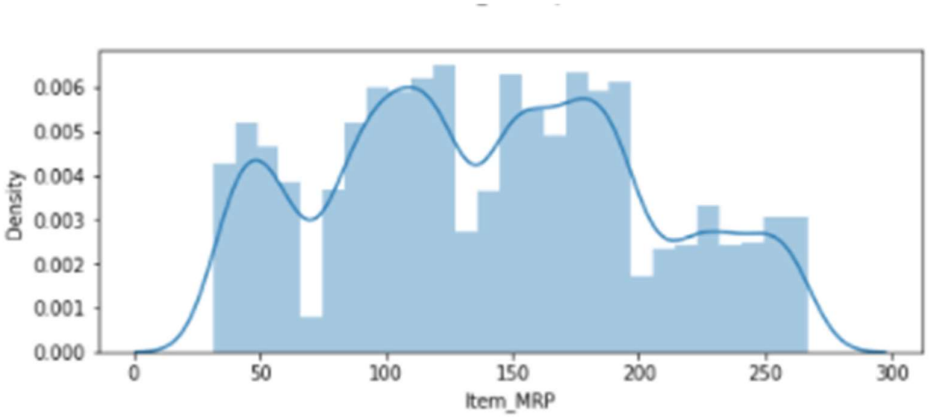


Figure 5- Density of item MRP

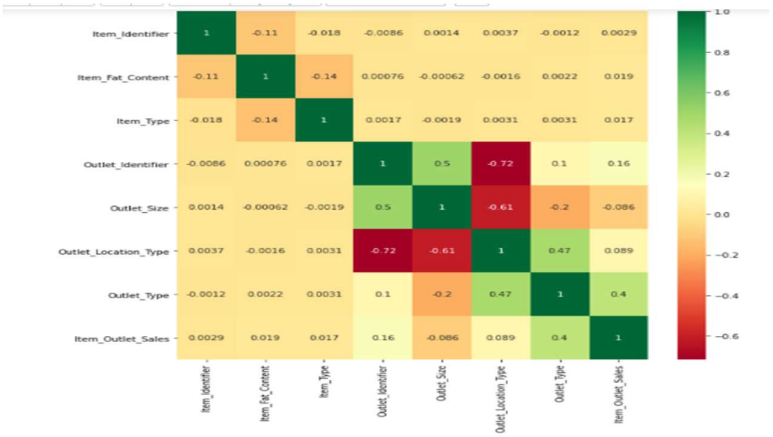


Figure 6- Heatmap for categorical data with target variable item outlet sales

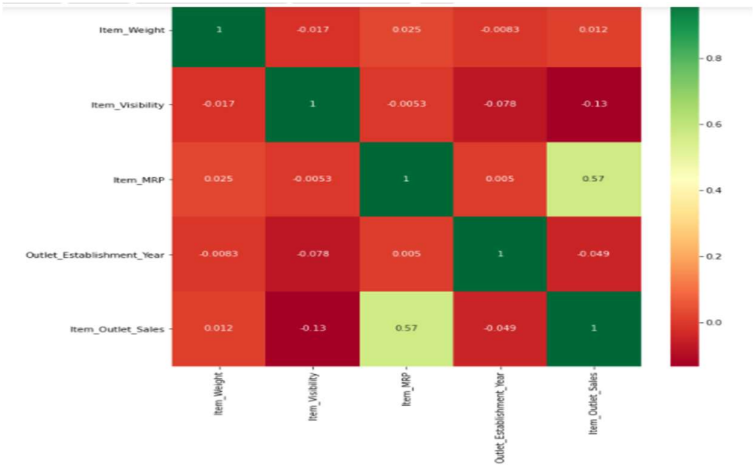


Figure 7- Heatmap for numerical data with target variable item outlet sales

From the above results we can visualize our dataset and see how they have relationship with each. Also above heatmaps can give an idea about how many percentage they have correlation with the target variable item outlet sales.

Among several predictive algorithms, data mining algorithms are considered for prediction. It includes Decision Tree, Gradient Boosting Regressor, Linear regression, XG Boost algorithm, and Random forest. These algorithms helped to predict how the sales can be done for develop the business economically in future.

In this project we used following algorithms and got the different values of accuracy.

Table 2-Algorithms and accuracy of them

Algorithm	Accuracy(%)
Linear Regression	56.44
Decision Tree Regressor	100
Random Forest Regressor	93.74
Ridge	56.44
Gradient Boosting Regressor	65.56
Ada Boost Regressor	43.69
XGB Regressor	85.76

Here decision tree regressor gave accuracy 100% but in reality all predictive modelling problems have prediction error. So we neglect that.

So we select the Random Forest Regressor for our prediction model with accuracy of 93.74%

Then after doing some parameter tuning by using grid search cv we got the accuracy of 93.76%

So by using parameter tuning here the accuracy of the model increased by 0.02%.

Then finally take this model to flask framework and developed a web application interface to predict the

In addition to this, study about the existing models and there by defining a new model to predict the sales in the bigmart with high accuracy can be done.By using this sales prediction model Big marts can move their business by minimizing the risk in the business and move forward to the future by success journey.

Future Developments

- Can use both classification and clustering algorithms also.
- Also can user hybrid of classification and clustering algorithms.
- Design the user interface more user friendly.

References

[1]"A Two-Level Statistical Model for Big Mart Sales Prediction", *Ieeexplore.ieee.org*, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8675060/authors#authors. [Accessed: 28- Aug - 2021].

[2]G. Behera and N. Nain, "A Comparative Study of Big Mart Sales Prediction", *Communications in Computer and Information Science*, pp. 421-432, 2020. Available: 10.1007/978-981-15-4015-8_37 [Accessed 28 Aug 2021].

[3]G. Behera and N. Nain, "Grid Search Optimization (GSO) Based Future Sales Prediction for Big Mart," 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Sorrento, Italy, 2019, pp. 172-178, doi: 10.1109/SITIS.2019.00038.

[4]K. S.Mohanapriya, "Big Mart Sales Forecast using Linear Regression", *Solidstatetechnology.us*, 2021. [Online]. Available: http://www.solidstatetechnology.us/index.php/JSST/article/view/4367. [Accessed: 28- Aug - 2021].

[5]N. Saravana Kumar, K. Hariprasath, N. Kaviyavarshini and A. Kavinya, "A study on forecasting bigmart sales using optimized machine learning techniques", *Science in Information Technology Letters*, vol. 1, no. 2, pp. 52-59, 2020. Available: 10.31763/sitech.v1i2.167 [Accessed 28 Aug 2021].

[6]"Search | Kaggle", *Kaggle.com*, 2021. [Online]. Available: https://www.kaggle.com/search?q=big+mart+sales+prediction. [Accessed: 28- Aug - 2021].

[7]*Xajzkjdx.cn*, 2021. [Online]. Available: http://xajzkjdx.cn/gallery/423-april2020.pdf. [Accessed: 28- Aug - 2021].

[8]B. Learning, A. Kumar, S. Labs and 0. Join 250, "Bigmart Sales prediction using Machine Learning - from Skyfi Labs", *Skyfilabs.com*, 2021. [Online]. Available: https://www.skyfilabs.com/project-ideas/bigmart-sales-prediction-using-machine-learning. [Accessed: 28- Aug- 2021].