**BIG MART SALES PREDICTION REPORT**

**Abstract:**

Nowadays shopping malls and Big Marts keep the track of their sales data of each and every individual item for predicting future demand of the customer and update the inventory management as well. These data stores contain a large number of customer data and individual item attributes in a data warehouse. Further, anomalies and frequent patterns are detected by mining the datastore from the data warehouse. The resultant data can be used for predicting future sales volume with the help of different machine learning techniques for retailers like Big Mart. In this paper, we propose a predictive model using the Random Forest technique for predicting the sales of a company like Big Mart and found that the model produces better performance as compared to existing models. A comparative analysis of the model with others in terms of performance metrics is also explained in detail.

**Introduction:**

Day by day competition among different shopping malls as well as big marts is getting more serious and aggressive only due to the rapid growth of the global malls and on-line shopping. Every mall or mart is trying to provide personalized and short-time offers for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organization, logistics and transport service, etc. Present machine learning algorithm are very sophisticated and provide techniques to predict or forecast the future demand of sales for an organization, which also helps in overcoming the cheap availability of computing and storage systems. In this paper, we are addressing the problem of big mart sales prediction or forecasting of an item on customer’s future demand in different big mart stores across various locations and products based on the previous record. Different machine learning algorithms like linear regression analysis, random forest, etc are used for prediction or forecasting of sales volume. As good sales are the life of every organization so the forecasting of sales plays an important role in any shopping complex. Always a better prediction is helpful, to develop as well as to enhance the strategies of business about the marketplace which is also helpfulto improve the knowledge of marketplace. A standard sales prediction study can help in deeply analyzing the situations or the conditions previously occurred and then, the inference can be applied about customer acquisition, funds inadequacy and strengths before setting a budget and marketing plans for the upcoming year. In other words, sales prediction is based on the available resources from the past. In-depth knowledge of the past is required for enhancing and improving the likelihood of marketplace irrespective of any circumstances especially the external circumstance, which allows preparing the upcoming needs for the business. Extensive research is going on in retailers' domain for forecasting the future sales demand. The basic and foremost technique used in predicting sales is the statistical methods, which is also known as the traditional method, but these methods take much more time for predicting sales also these methods could not handle nonlinear data so to over these problems in traditional methods machine learning techniques are deployed. Machine learning techniques can not only handle non-linear data but also huge data-set efficiently. To measure the performance of the models, Root Mean Square Error (RMSE), cross-validation is used

**Proposed System:**

For building a model to predict accurate results the dataset of Big Mart sales undergoes several sequences of steps as mentioned in this work, we propose a model using the Random Forest technique. In our model we have used 2013 Big mart dataset After preprocessing and filling missing values, we used Decision trees, Linear regression, Ridge regression, Lasso Regression , Random forest, Extra tree and Xgboost.RMSE is used as accuracy metrics for predicting the sales in Big Mart. From the accuracy metrics it was found that the model will predict best using minimum RMSE The details of the proposed method is explained in the following section.

**Dataset Description:**

In our work we have used 2013 Sales data of Big Mart as the dataset. Where the dataset consists of 12 attributes like Item Fat, Item Type, Item MRP, Outlet Type, Item Visibility, Item Weight, Outlet Identifier, Outlet Size, Outlet Establishment Year, Outlet Location Type, Item Identifier and Item Outlet Sales. Out of these attributes response variable is the Item Outlet Sales attribute and remaining attributes are used as the predictor variables. The data-set consists of 8523 products across different cities and locations. The data-set is also based on hypotheses of store level and product level. Where store level involves attributes like: city, population density, store capacity, location, etc and the product level

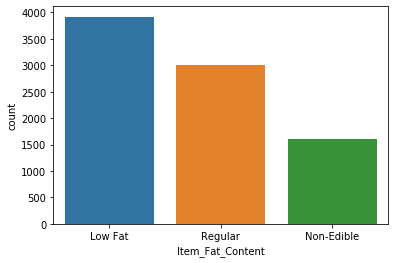
hypotheses involves attributes like: brand, advertisement, promotional offer, etc. After considering all, a dataset is formed and finally the data-set was divided into two parts, training set and test set in the ratio 80 : 20.

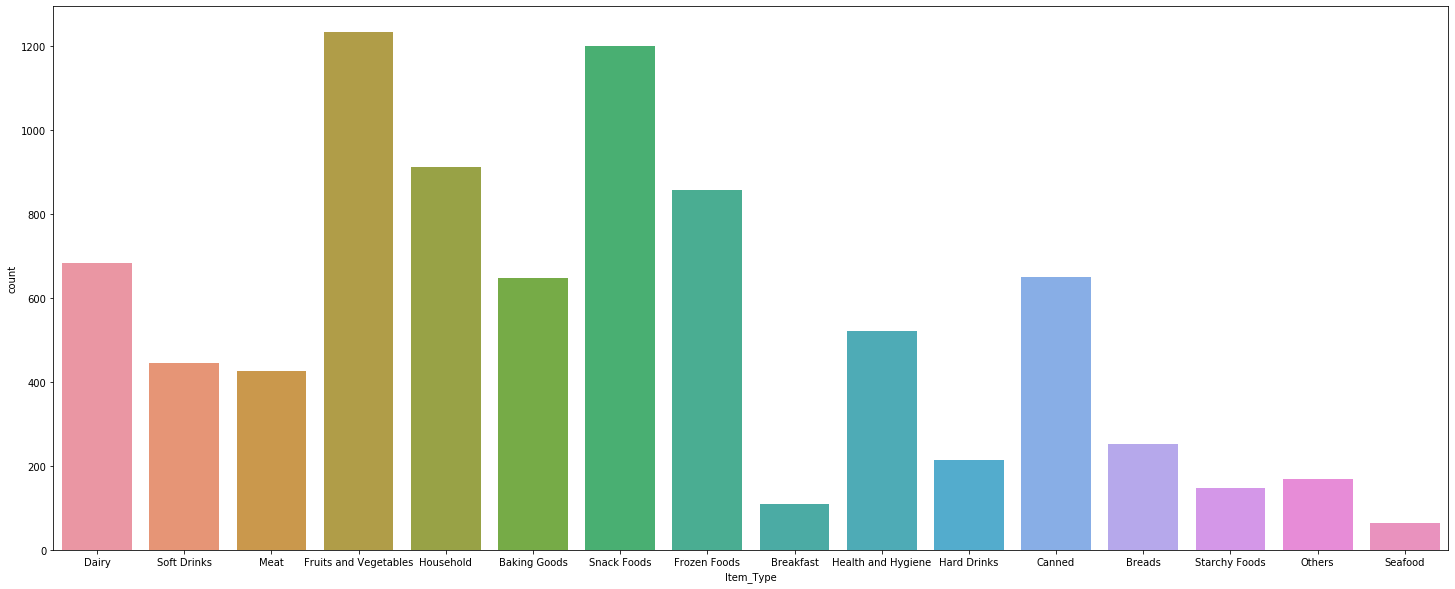
**Data Wrangling:**

It was observed from the previous section that the attributes Outlet Size and Item Weight has missing values. In our work in case of Outlet Size missing value we replace it by the mode of that attribute and for the Item Weight missing values we replace by mean of that particular attribute. The missing attributes are numerical where the replacement by mean and mode diminishes the correlation among imputed attributes. For our model we are assuming that there is no relationship between the measured attribute and imputed attribute.

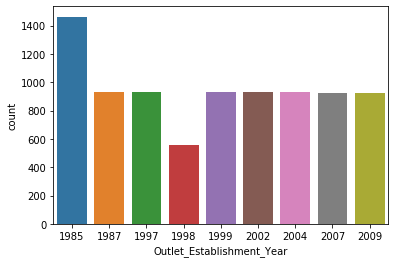
**EDA:**

During EDA and data visualization following inferences were drawn:

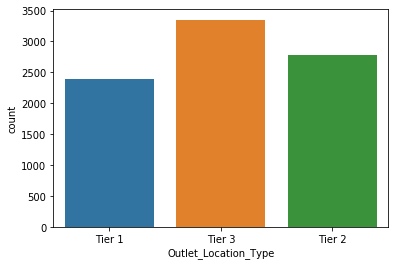
1. The count of low fat products are more
2. Distribution of Item Type:



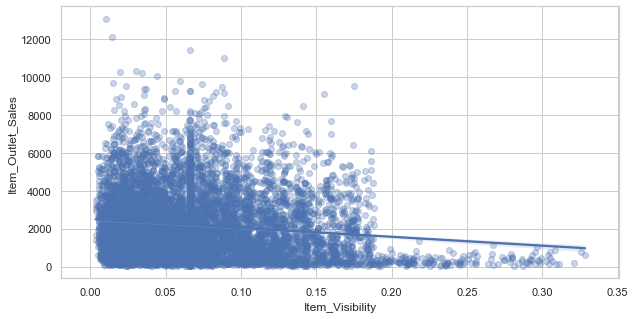
1. In almost all years , no. Of outlets established was almost equal but 1985 has more and 1998 has the least



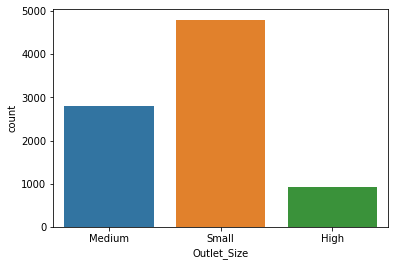
1. Impact of outlet location type on target variable item outlet sale



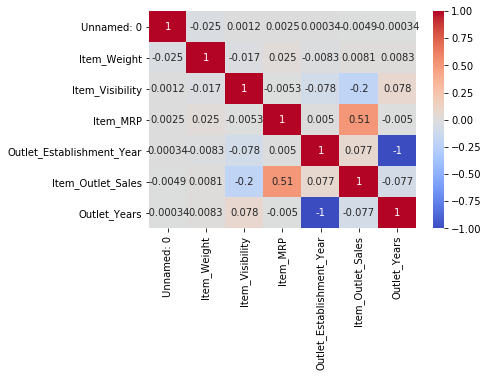
5. Impact of item visibility on target variable item outlet sale. Less visible items are sold more compared to more visibility items as outlet contains daily used items which contradicts the null hypothesis.



1. Distribution of outlet size. The number of outlet size are available in the dataset.



1. Correlation among features of a dataset



**Preprocessing of Training Data**:

Some nuances were observed in the data-set during data exploration phase. So this phase is used in resolving all nuances found from the dataset and make them ready for building the appropriate model. During this phase it was noticed that the Item visibility attribute had a zero value, practically which has no sense. So the mean value item visibility of that product will be used for zero values attribute. This makes all products likely to sell. All categorical attributes discrepancies are resolved by modifying all categorical attributes into appropriate ones. In some cases, it was noticed that non-consumables and fat content property are not specified. To avoid this we create a third category of Item fat content i.e. none. In the Item Identifier attribute, it was found that the unique ID starts with either DR or FD or NC. So, we create a new attribute Item Type New with three categories like Foods, Drinks and Non-consumables. Finally, for determining how old a particular outlet is, we add an additional attribute outlet to the dataset. We applied Label encoding to convert categorical into numerical values

**Modelling:**

**Different models were applied. Details can be found in the table:**

|  |  |  |
| --- | --- | --- |
| **Model Name** | **RMSE** | **CV score** |
| Linear Regression | 1199.66 | 0.5004507421344684 |
| Lasso Regression | 1199.71 | 0.5004430288096436 |
| Ridge Regression | 1199.71 | 0.5004430288096436 |
| Decision Tree | 1554.42 | 0.16184553504542393 |
| Random Forest | 1141.91 | 0.5477542898079568 |
| Extra Tree | 1160.39 | 0.5189858238613315 |
| XG boost | 1179.62 | 0.5277455120663482 |

Random Forest has the least RMSE. So we chose this model for prediction .

Item MRP and Outlet Type affects the most sales .