Project: Big Mart Sales Prediction

Introduction

The Big Mart is an e-commerce and a super market chain, has 10 stores spread across different cities in a country. The company wants to improve the sales of the company and thus, wants to build a predictive model that can predict the sales of each product available in each store. Based on the attributes of the products and outlets, the predictive model would help the company determine the expected sales of each product at a particular store and would also help identify feature importance of each variable. The company may use this analysis to take constructive measure to ensure a growth in their business (Analytics Vidhya, 2016).

Problem Statement

The Data Science team at Big Mart has the responsibility of collecting and analyzing the Sales data, which includes product attributes and different properties of the outlets. The essential job of the team is to analyze various products' attributes as well as the properties of the stores which can impact the sales and use this analysis to build a predictive model which can successfully predict the estimate sales of each product at different stores. This predictive model would be helpful for the company in determining the feature importance and key role of different products and stores in improving the sales of the company.

Data Description - Test and Train Data Set

1. Purpose of Data set – Record the information of the available products and store properties.
2. Source of Data set – The data set is available on Analytics Vidhya Hackathon, downloaded from https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/?utm_source=auto-email

3. Time window – Year 2013
4. Cost of Data – Downloadable for free.
5. Collection technique – Update the database as the new products gets available on the store and even update the database as soon as any attribute of the product gets change.
6. Collection tools – Big Mart database and buyer's transactional history.
7. Quality – The data is genuine since it has been provided by the internal source and thus there is a transparency in the data been available.
8. Completeness – The data isn't complete, there are some missing fields in the Test data set.

Variables that are used in the Big Mart Data set

Name	**Definition**	**Variable Classification**
Item_Identifier	Unique product ID	Character
Item_weight	Weight of product	Numeric
Item_Fat_content	Whether product is low fat or not	Character
Item_visibility	The percentage of total display area of all products in a store allocated to the particular product.	Numeric
Item_Type	The category to which the product belongs.	Character
Item_MRP	Maximum retail price of the product.	Numeric
Outlet_Identifier	Unique store ID	Character
Outlet_Establishment_Year	The year in which store was established	Character
Outlet_Size	The size of the store in terms of the ground area covered.	Character
Outlet_Location_Type	The type of the city in which the store is located.	Character
Outlet_Type	Whether the outlet is a grocery store or a supermarket	Character
Item_Outlet_Sales	Sales of the product in the particular	Numeric

	store. This is the outcome variable to be predicted.	

Techniques and methodologies involved in Analysis and building model

1. **Hypothesis Generation**

   This step involved understanding the business problem of the company and hypothesizing different variables in order to understand their impact on the target variable (Sales). This helped in performing EDA and focusing on important variables. Following are the hypothesis been made at Product level and Outlet level:

   Product Level hypothesis

   - **Product Visibility:** Items kept at the upper level of the shelves or are given larger space in the store are likely to grab more customer attention.
   - **Product Type:** Daily use products such as vegetables and dairy products are likely to have higher sales in comparison to any specific product.

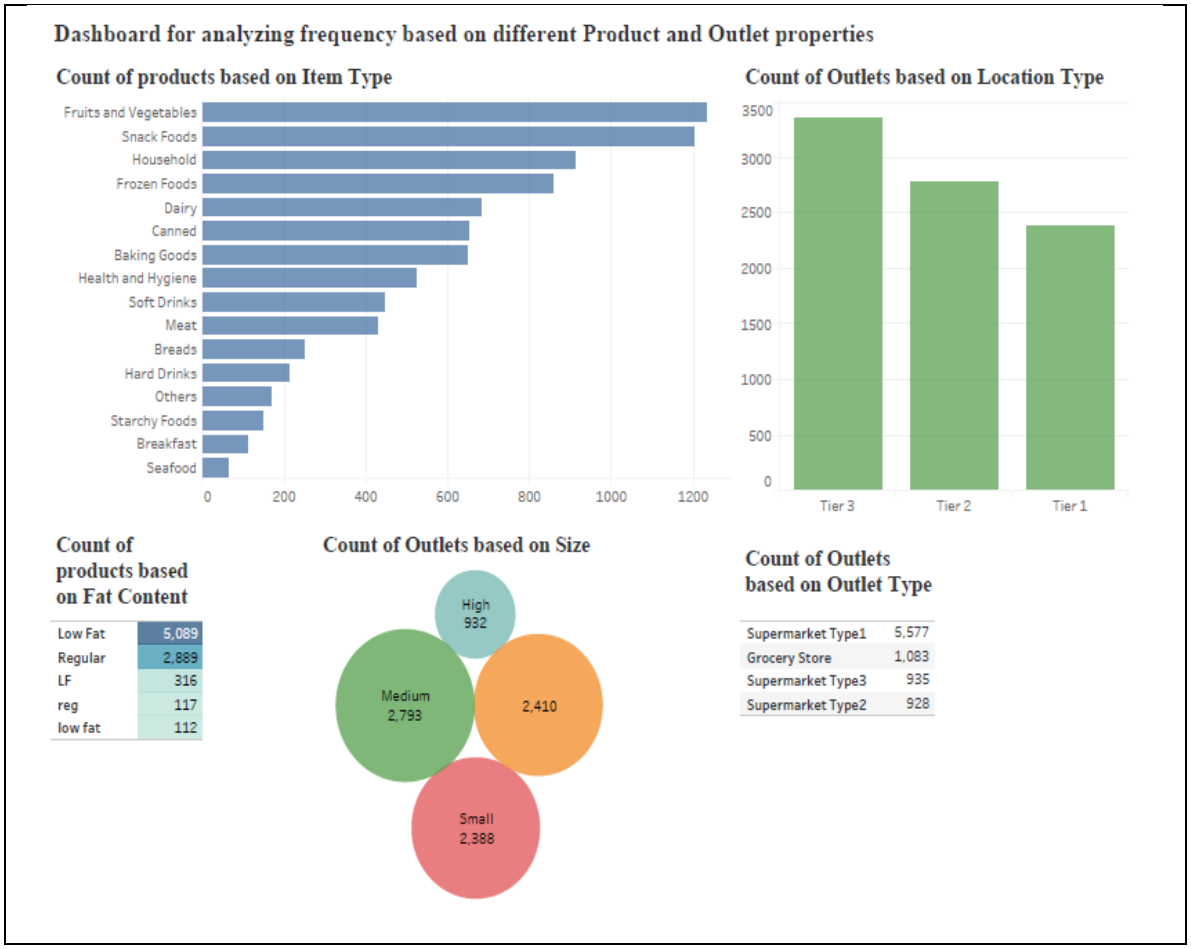
   Store Level hypothesis

   - **Outlet Location:** Outlets situated at the Tier 1 cities are likely to have higher sales as compared to outlets present at Tier 2 or Tier 3 cities because of the higher income of the people living in Tier 1 cities.
   - **Outlet Type:** Supermarkets having good ambience and managed by sophisticated staff are expected to make higher business than other types of stores.

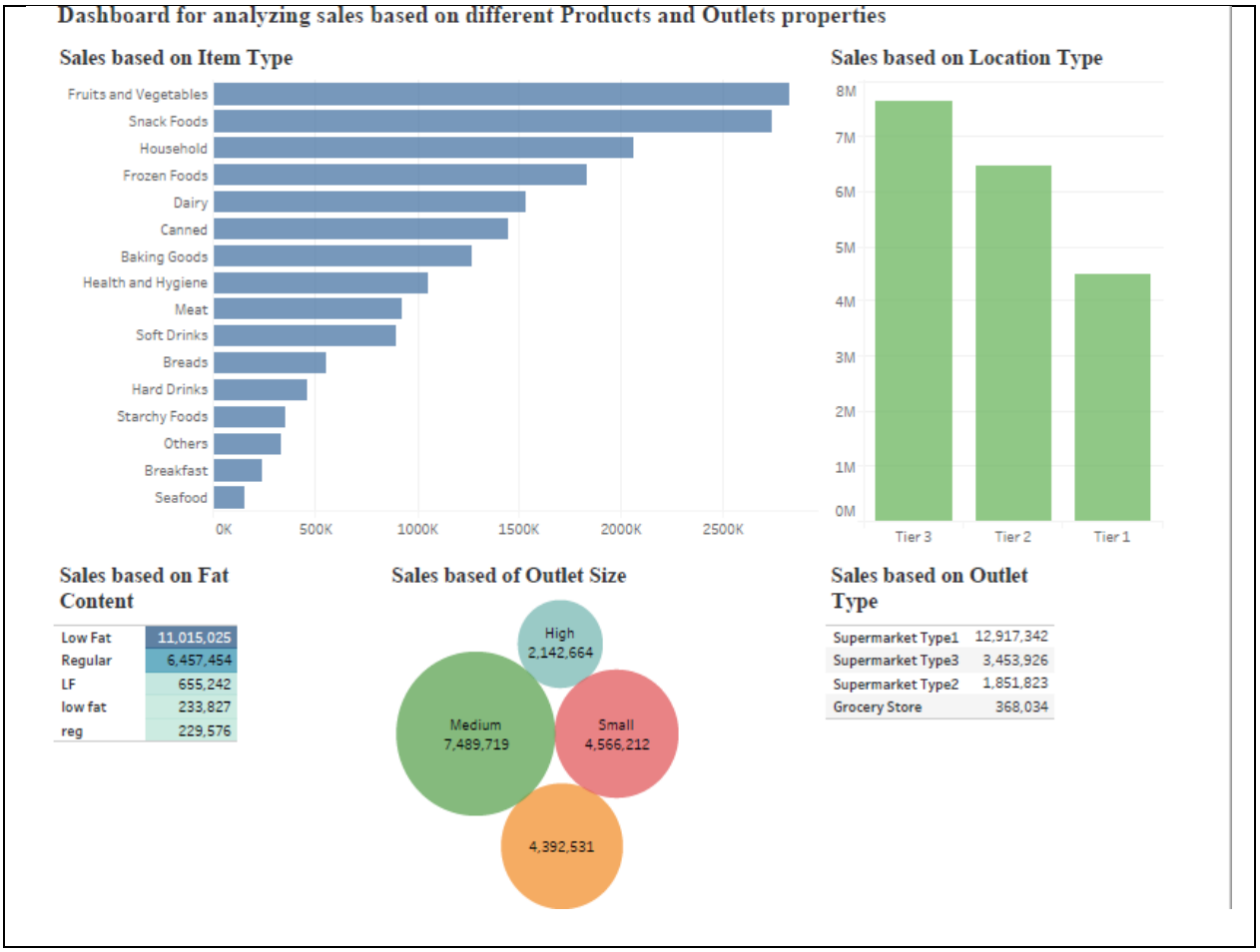
2. **Exploratory Data Analysis**

   Based on the hypothesis, analyzed the data statistically and visually to find the key insights in the data as well as any anomalous behavior present which can be rectified

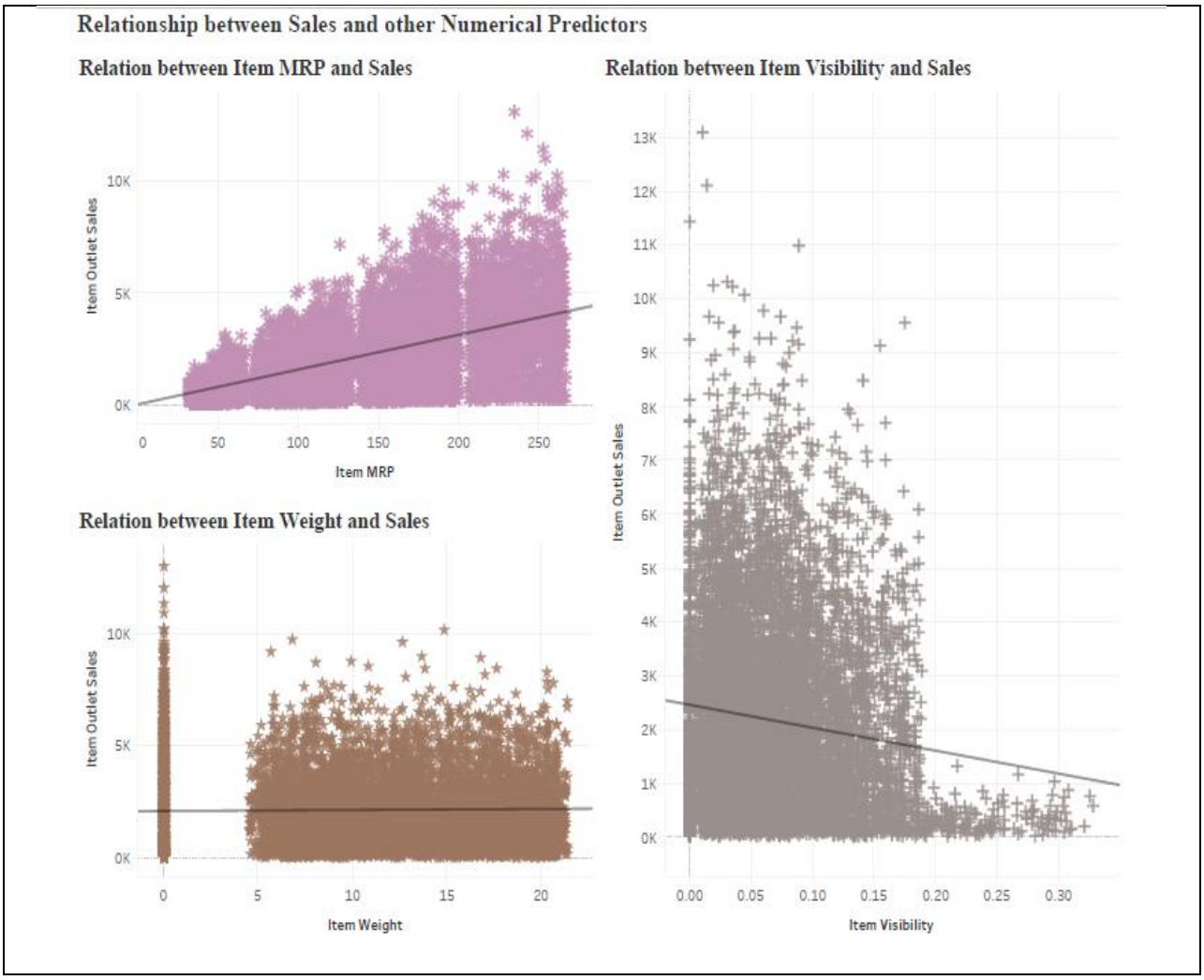
during Data Processing. Following are the key findings obtained upon analyzing the dataset on Tableau:



The above dashboard represents the frequency of products available in different unique categories within various variables. It was observed that there are 16 unique Item_Type, 5 unique categories in Item_Fat_Content, 3 Outlet_Location_Type, 4 Outlet_Size and 4 Outlet_Type. However, on analyzing the results carefully it was found that some variables have either missing values or misspelled entries.



The above dashboard highlights the impact on sales based on different product as well as outlet attributes. The analysis shows that the products under Food and vegetables category has highest sales and the outlets situated in Tier 3 cities have highest sales in comparison to the outlets present in other cities.



The above dashboard showcases the relation between numerical variables. The Item_MRP has relatively strong positive relation with the Sales, however Item_Visibility has a negative relation with the Sales which seems to be logically incorrect.

3. Data Pre-Processing

Both the training as well as test dataset were combined before applying any modifications. Once, the datasets were combined, statistical summary of the combined dataset was obtained. It was observed that Item_weight and Outlet_Size have missing values. Thus, a **mean weight** of products categorized by their identifier number was

determined to impute the missing values and to impute the Outlet_Types, a **mode of Outlet_Size** categorized by Outlet_Type was computed. Earlier, it was found that Item_Visbility has anomalous behavior, thus to get more insights about the variable, a summary of the variable was determined. Here, it was found that the minimum visibility is 0. However, if a product got sold, it had some visibility in the store, thus, a **mean visibility** based on Item_Identifier was computed to replace entries consisting 0 visibility.

4. **Feature Engineering**

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

   There may be a possibility that the number of years an outlet has been running its business would have an impact on the sales, if an outlet has a good history and is popular in business would have higher sales, or a newly launched outlet with modern amenities and ambience would have a higher sale. Thus, determining the year of operation of a store may have an impact on the predictive model.

   b) **Create a broad category of Item_Type**

   Upon closely analyzing the Item_Identifier it was found that the first two characters of the ID can be used to create another variable. In this category the products are divided into 3 categories; Food (FD), Drinks (DR) and Non-Consumable (NC).

   c) **Modify Item_Fat_Content**

   The variable Item_Fat_Content has some misspelled entries which were rectified. In this, LF and low fat were changed to “Low Fat” and reg to “Regular”. It was also found that since some of the products are Non-Consumable, thus, accordingly the fat content was modified to “Non-Edible”.

d) Convert categorical variable to numerical variable

All the categorical variables were converted to into numerical variable since scikit learn accepts only numerical input. Consequently, dummy variables were obtained for each of these numerical categorical variables.

5. Model Building

In order to build a predictive model, various Machine Learning algorithms such as Linear Regression, Decision Tree and Random Forest were used. Firstly, the model was built using Linear Regression, the algorithm was trained on the training dataset and a cross-validation method was used to check the performance of the model. To check the performance, a CV score (Cross-validation score) was computed and a mean of 1129 was obtained. Moreover, to check the accuracy of prediction, a RMSE score of 1127 was obtained. However, to improve the performance and accuracy of the model, Decision Tree algorithm was implemented and to verify the model, a RMSE score of 1058 was obtained which was much better than the RMSE score obtained on Linear regression model. This algorithm was also used to determine the Feature Importance of the predictors. The predictors with highest Feature importance rate are Item_MRP, Outlet_0, Outlet_5, and Outlet_years. The feature importance shows the impact of the variable on the target variable (Sales). Thus, to further enhance the accuracy of the model, it was fitted to Random Forest algorithm but a RMSE score of 1069 was obtained which is higher than the one obtained with Decision Tree model.

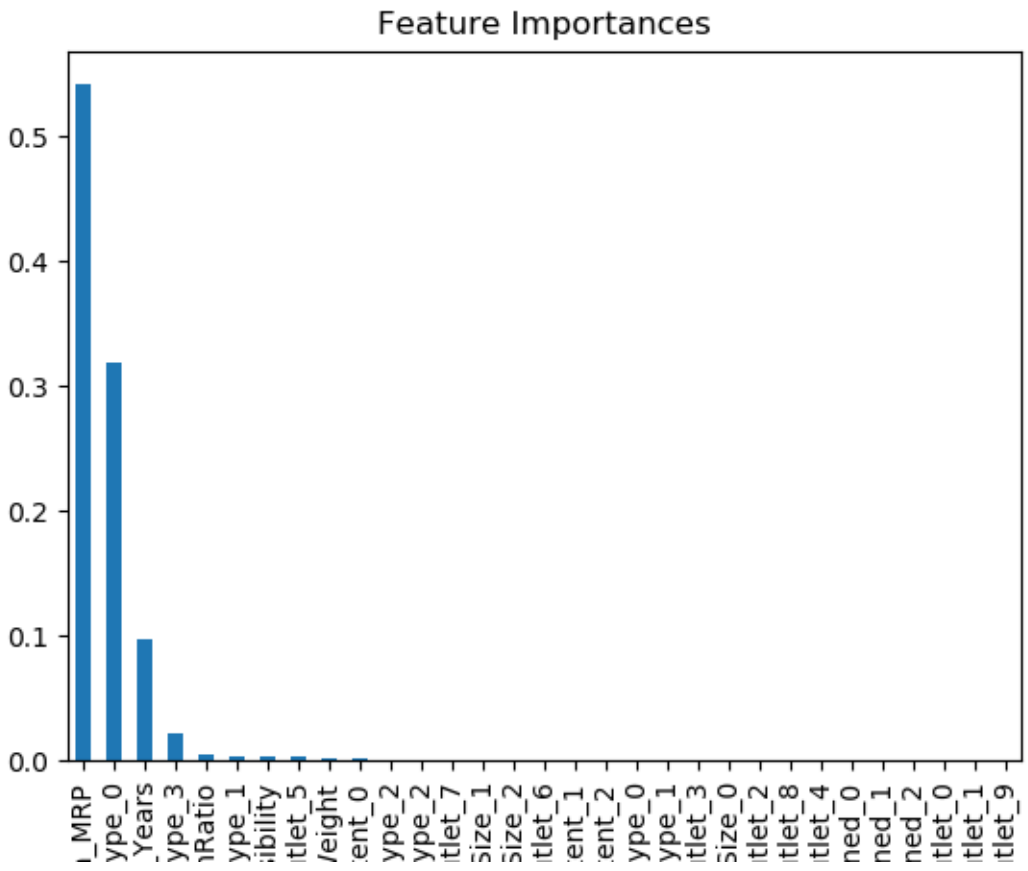


Fig: Bar graph representing the feature importance of the predictors upon implementing Decision Tre model.

Conclusion

This model would be beneficial in estimating the sales of each product at different stores which would definitely help the company improve their business by taking measures analyzing key variables from the dataset that could greatly impact their business. Analysis of different Outlet properties can also help company improve the services at different stores, this may enhance customer's experience. The accuracy of this model can further be enhanced by introducing more variables that may have direct impact on the sales, this may further reduce the RMSE score.

References

Analytics Vidhya. (2016, May 25). Big Mart Sales Practice Problem. Retrieved from https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/

Donges, N. (2018, Feb 22). The Random Forest Algorithm. Retrieved from https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Jha, V. (2017, June 18). Decision Tree Algorithm for a predictive model. Retrieved from https://www.techleer.com/articles/120-decision-tree-algorithm-for-a-predictive-model/

Microsoft Azure. (2017, Nov 03). The Team Data Science Process Lifecycle. Retrieved from https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

Statistics Solution. (2013). What is Linear Regression. Retrieved from https://www.statisticssolutions.com/what-is-linear-regression/