An empirical study of BigMart's sales

Hanna Carucci Viterbi, Giulia Di Martino, Martina Bozzi,
Olimpia Sannucci, Cosimo Poccianti, Carlo Ardito

Contents

1 Introduction

Data for 8523 products at BigMart has been collected across 10 stores in different cities. Also, various characteristics for each product and store have been described. The object of the analysis is to build a significant regression model which determines the main attributes that have an influence on the sales of a product. Using this model, we will try to understand the properties of products and stores which play a key role in increasing sales.

2 Data Description

The objective of the empirical analysis is the study of Big Marts' sales. The data set provided is composed of 8523 observations and 12 variables (7 character and 5 numeric, the target variable is included). The variables are: **Item Outlet Sales** (represents the dependent variable of the linear regression model and it identifies the sale of a specific product in a specific branch); **Item Identifier** (describes the id of the product and is a categorical variable); **Item Weight** (indicates the weight of the product); **Item Fat Content** (it is a categorical variable of the type "Low fat" or "Regular"); **Item Visibility** (the percentage of the area that a branch dedicates to a product); **Item Type** (delineates the category of the product. There are 16 categories of the product, one of them is generic and defined as 'others'); **Item MRP** (outlines the maximum retail price of the product); **Outlet Identifier** (the unique ID of the branch. It is a categorical variable and there is data from 10 different branches regarding the sales of the products); **Outlet Establishment Year** (the year in which the branch was opened. It is a categorical variable and the data set is based on 9 different establishment years); **Outlet size** (the size of the shop in terms of the surface. It is a categorical variable that can assume 3 values: high, medium, small); **Outlet Location Type** (the type of area in which the branch is located. It is a categorical variable and it divides the cities according to a Tier system. The values accepted are: Tier-1, Tier-2 and Tier-3); **Outlet Type** (typology of the branch. It is a categorical variable that has four variations: supermarket of type 1, type 2, type 3 and grocery stores).

2.1 Outlet Identifier

The variable Outlet Identifier is being analyzed in detail to obtain a broader understanding of the data set. The latter comprises 10 different branches. As a result of a multilateral analysis based on three main variables (Outlet Identifier, Outlet Location Type, Outlet Type) it has been discerned each branch to its particular type and location. Thus, OUT019 and OUT010 fall into the category of Grocery. The peculiarity is that these types of Outlet are located respectively in Tier 1 and Tier 3. Differently, a larger number of branches (OUT046,OUT049,OUT035,OUT045,OUT017,OUT013) are Supermarkets of Type 1. The distribution of the branches is spread across the three types of areas. In particular, OUT013 is the only placed in Tier 3. Ultimately, the only supermarket of type 2 is OUT018 located in Tier 3, as well as OUT027 which is the unique Supermarket of Type 3. From this it can be deducted that Tier 3 must be a highly populated area since all the four types of Outlet are present there. Contrarily, in areas labeled as Tier 1, Groceries and Supermarkets of type 1 are the only types of Outlet available. As Groceries are only in Tier 1, and these types of activities function mostly in small areas, it can be assumed that such areas have a lower population density. Finally, in Tier 2 there are only Supermarkets of type 1, leading to the conclusion that these cities are medium populated (Appendix 5.1)

3 Data Analysis

In order to carry out the empirical analysis of the model the data frame has been modified since it presented missing values. 526 data points have visibility equal to zero, while the Item Weight and Outlet size have, respectively, 1463 and 2410 missing values. To prevent omitting all the missing values, Item Weight's have been substituted with their mean and Outlet size's missing values replaced with their mode which is the size small. To compute the regression it is useful to remove the variables Outlet Establishment Year and Item Identifier which contain superfluous information for the analysis.

From the graph (Appendix 5.1) results that exists a positive linear relation (0.567) between Item Outlet Sales and Item MRP: as the number of sales of a product increases, so does its Maximum Retail Price, and vice versa. For what concerns the other two numerical variables (Item Visibility and Item Weight) the relations are weak.

In order to prevent collinearity a new data frame has been created: dummy variables were generated in order to deal with this issue. For each categorical variable a category has been dropped: the value left out can be thought of as the reference value and the fit values of the remaining categories represent the change from this reference. After having dropped a category for each of the six categorical variables, the correlation has been assessed. In this way all the variables that are perfectly correlated have been eliminated. Three variables in particular resulted to be highly correlated: Outlet Location Type, Outlet Type and Outlet Size. The latter were therefore removed. The plot now shows only mild correlations left in the predictors which are ready to be used in the regression analysis (Appendix 5.3).

The target variable has been specified as Y. The latter, being right-skewed, has been transformed to normality by taking the cubic root. The initial behaviour of Y is represented as follows (Appendix 5.2). After the transformation the dependent variable follows a normal distribution as represented by the histogram: (Appendix 5.2).

Before fitting the model data has been split through a proportion of 75% for training data and the rest for the testing set.

3.1 Complete Model Analysis

The equation of the complete model is the following:

$$\begin{gathered}\mathrm{Y} = \beta_0 + \beta_1 ItemWeight + \\ \beta_2 ItemFatContent + \beta_3 ItemVisibility + \beta_4 ItemType + \\ \beta_5 ItemMRP + \beta_6 OutletIdentifier + +\beta_7 OutletSize + \beta_8 OutletLocationType + \beta_9 OutletType + \epsilon\end{gathered}$$

An interpretation of the equation is assessed. The intercept is equal to 8.575, representing the expected value of the sales of a specific product when the independent variables are equal to zero. Given that the mean of the sales of a product is 11.992 and that the Residual Standard Error is 2.008, the percentage error is 16.744%. It's also worth noting that the Residual Standard Error is calculated with 6360 degrees of freedom. Theoretically, degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account the latter.

The R-squared (R^2) statistic provides a measure of how well the model is fitting the actual data and it is equal to 0.6899. Indeed, 69% of the variance found in the response variable can be explained by the predictor variables. The adjusted R^2 is 0.6884. The latter value is more precise and thus is the one taken in consideration when evaluating the fit of the model.

3.2 Verification of the assumptions of the residuals

After having estimated the complete linear regression model, it is necessary to verify whether the assumptions on the residuals are satisfied. In first instance it is checked whether the mean of the errors is not significantly different from zero, computing a t test. The first assumption is verified and found valid. Secondly, as the graph indicates, the variance of the error σ^2 is constant, and thus the assumption on homoscedasticity holds. Finally, it is assumed that the error terms are normally distributed as the following graph displays (Appendix 5.4).

3.3 Final model Analysis

The best-fit model according to AIC is the one explained by two variables: Item MRP and Outlet Identifier. Out of the 10 branches that compose Outlet Identifier just five were found to be significant (OUT010, OUT018, OUT019, OUT027,OUT045). The average sales of the product is equal to 8.196 when the Item MRP, Outlet Identifier (the 5 branches) are zero. The residual standard error of the regression model counting the two regressors is 2.011, outlining that 16.770% is the percentage error. The R^2 is now 0.6878, which means that 68.8% of the variance found in the response variable can be explained by the Item MRP, OUT010, OUT018, OUT019, OUT027, OUT045. The R^2 appears to be slightly lower than the one calculated on the complete model (0.6899), suggesting that these two variables are highly explicative. The Adjusted R^2 is now 0.6875.
Let's see in detail how the two predictors behave.

3.4 Item MRP

The coefficient estimate (β_1) for the Item MRP shows that for every 1 unit (in dollars) increase in the Maximum Retail Price, the sales go up by 3.111 (the cubic of 0.031451). Indeed, the standard error of the latter can be used to compute an estimate of the expected model in case the model is run iteratively. The number of sales of a specific product can vary by 0.000404. This value is a good indicator of how the observations are close to the fitted ones since is very small. The p-value is significant.

3.5 Outlet Identifier

Finally, an interpretation of Outlet Identifier variable is assessed. This variable is divided in 5 branches: Outlet Identifier OUT010, Outlet Identifier OUT018, Outlet Identifier OUT019, Outlet Identifier OUT027, Outlet Identifier OUT045. Indeed, just 5 out of the 10 different branches occur to explain the response variable. What stands out from the analysis of this specific variable is the negative estimate of 4 of the branches 010, 018, 019, 045 suggesting a decrease of the number of sales of the product when the branch identified is one of the latter. In detail when the outlet is OUT010, the sales of the product decreases of 228.211 (the cubic of 6.111). As for OUT018, the sales of the product decreases of 0.430 (the cubic of 0.755). When the branch is OUT019, the sales of the product decreases of 217.841 (the cubic of 6.017). While for OUT045, the sales of the product

decreases of 0.026 (the cubic of 0.297). The average standard errors for the above betas is 0.095 suggesting that the observed values are close to the fitted ones, and therefore good estimates.
On the contrary, the estimate of branch 027 is positive. Thus, the number of sales of the product increases of 11.852 (the cubic of 2.280). The standard error is 0.082145, representing the variation of the sales of a product. Lastly, the p-values are all significant, indicating that we can reject the null hypothesis which allows us to conclude that there is a relationship between the number of sales of a specific product and its maximum retail price together with the specific ID of the branch.

3.6 Predictive Performance

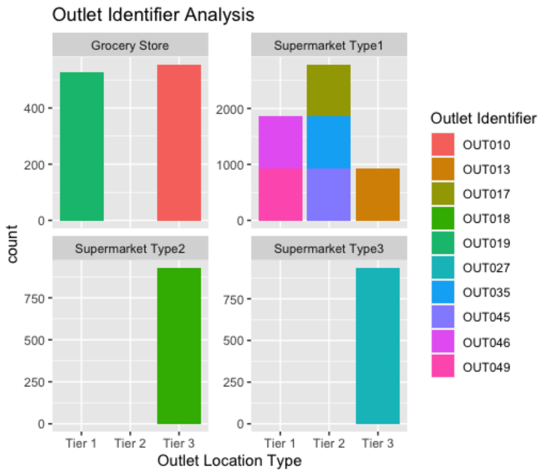
Finally, a comparison of the error on the training set and on the test set has been assessed.
The first one, on the training set is 2.01013 and it is slightly smaller than the second one, calculated on the test set, which is 2.033121. The final model doesn't seem to suffer from overfitting. To compare the values, a benchmark model composed only of the intercept, has been used. The latter has an RMSE respectively for the training set and test set of 3.597343 and 3.534489, which points that the final model is more accurate. The RMSE on the test set decreases of 42% with respect to the benchmark model, while the one on the train set of 44%.

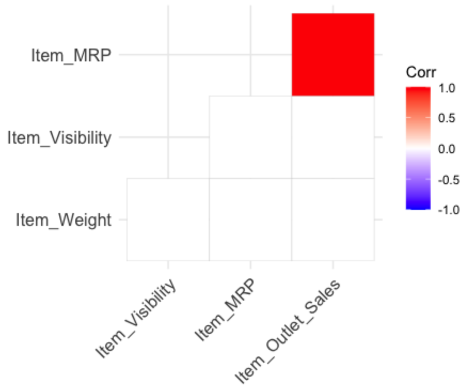
4 Conclusion

In conclusion, the empirical analysis shows that the Item Maximum Retail Price and the following types of Outlet Identifiers OUT010, OUT018, OUT019, OUT027, OUT045 have an influence on the sales of a product supplied by Big Marts, and play a key role in increasing or decreasing the sales of a specific product. The results of this empirical study can be now compared with the ones obtained by Gopal Behera and Neeta Nain who carried out with the Xgboost technique a comparative predictive study of Big Mart Sales. The authors came up with similar conclusions in regards of MRP, pointing out that OutletSales are highly correlated to MRP as found in the final model of the empirical study. Similar are the conclusions drawn for the Outlet Identifier 027 variable: the regression carried out in this paper detected a positive effect on Sales caused by Outlet Identifier 027, whereas the authors concluded in their study that the highest sales are made by OUT027 (as a medium size outlet in the super market type-3).

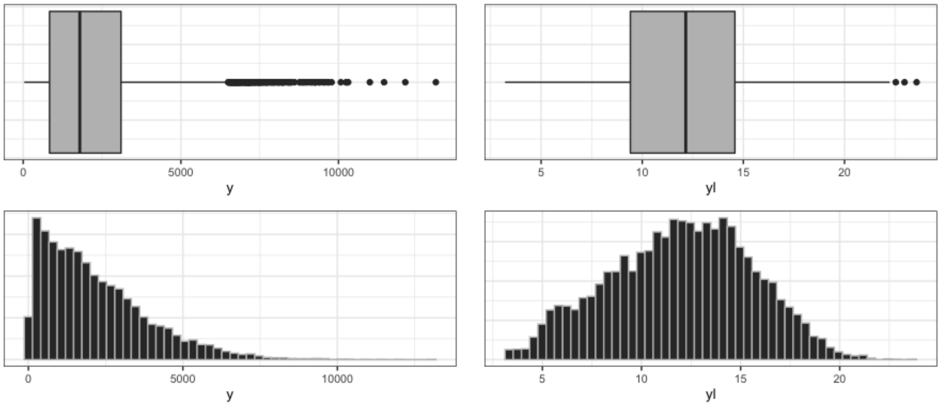
5 Appendix

5.1 Analysis of the variables

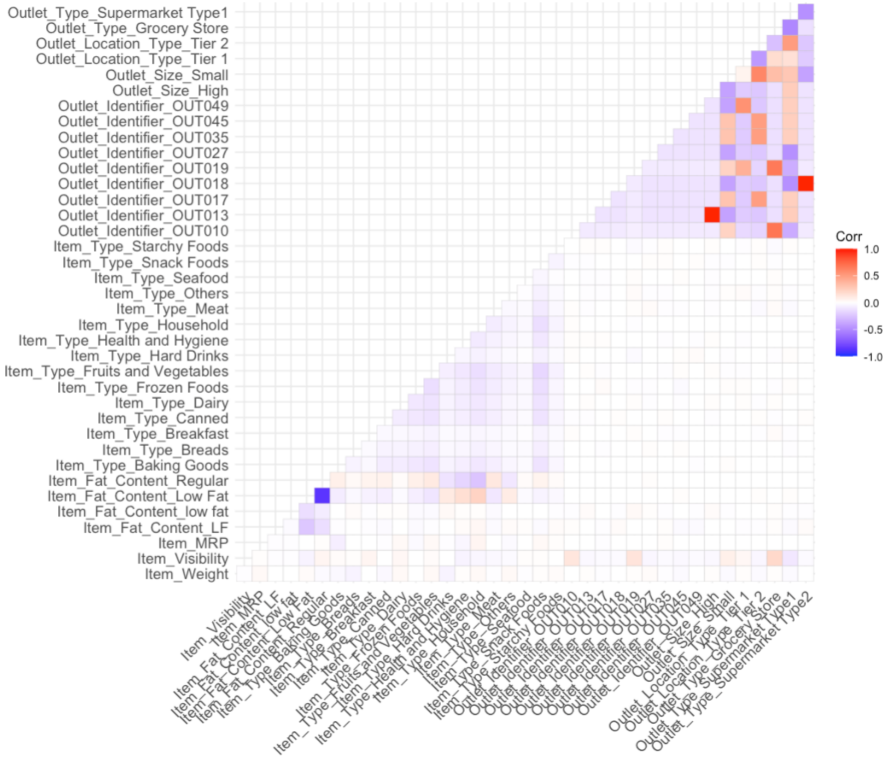


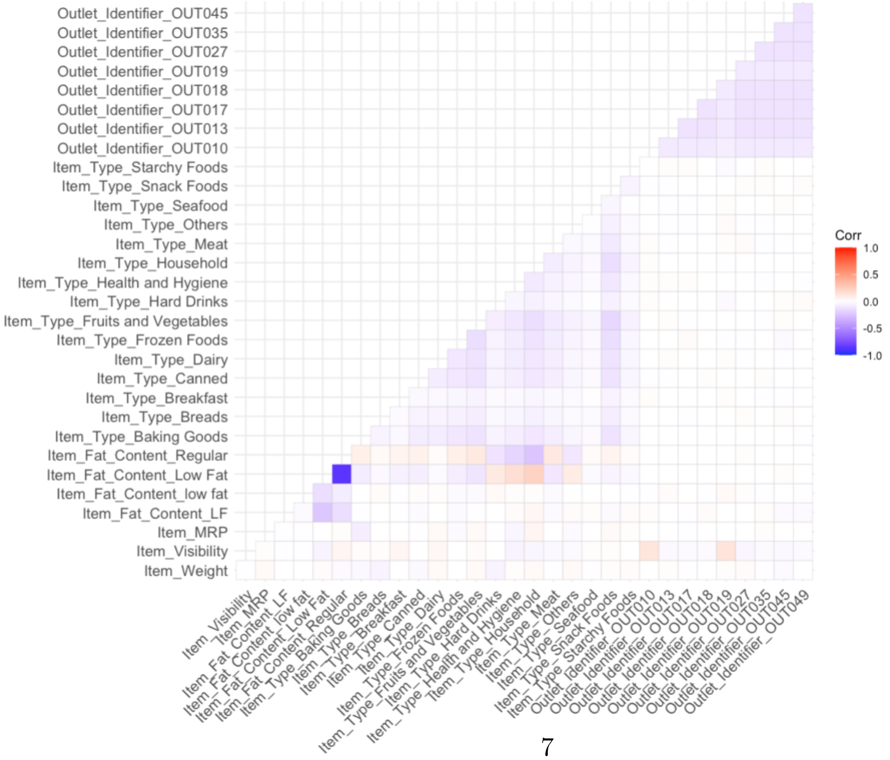


5.2 Y transformation: from right-skewness to normality

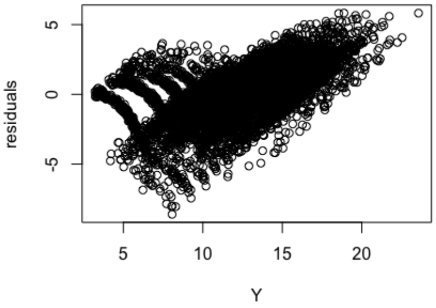


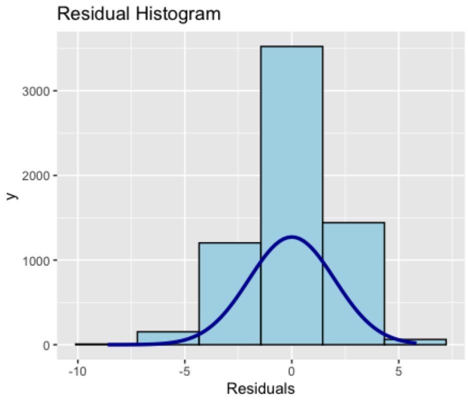
5.3 Associations

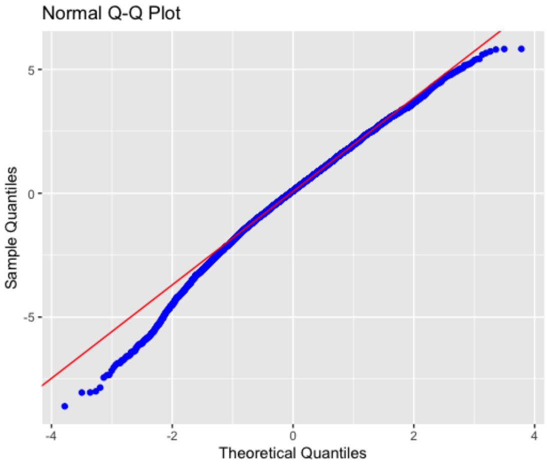


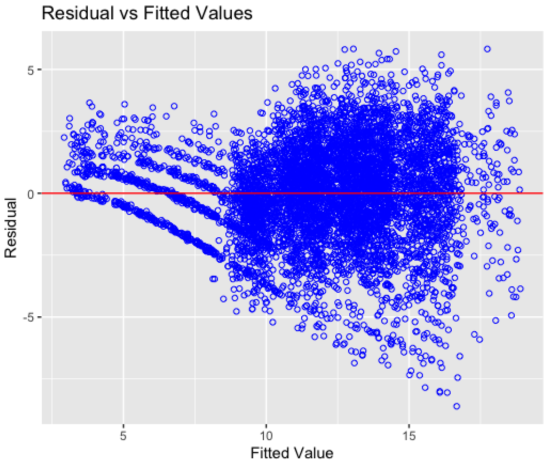


5.4 Residuals graphs


residuals
Y


Residual Histogram
y
Residuals


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


Residual vs Fitted Values
Residual
Fitted Value