OPTIMIZING SALES STRATEGIES WITH MULTIPLE LINEAR REGRESSION AND INTERACTIVE DATA VISUALIZATIONS

K. Lakshmi[1], B. Rama[2]

[1,2]Assistant Professor, Department Statistics BBCIT, Kachiguda Hyderabad, India.

ABSTRACT

Sales forecasts are important for forecasting future demand. It depends on two important factors: owning the right data and drawing the right conclusions from the right data. Mow a days most of the Business Organizations are focused on sales forecasts. Forecasts help to plan and reduce unnecessary costs. This means that we can offer the goods at a reasonable price. This allows companies to decide whether to add new products or remove failed products that are not in demand in the market. This article proposes a predictive model using multiple regression techniques from companies like Big Mart, a one-stop shopping centre, discussed to predict the sales of different types of products and the impact of different factors on the sale of items. MLR is an extension of linear regression. MLR improves model generalization and provides accurate results.

Keywords: Data Visualisation, Linear Regression, step wise regression, backward elimination

1. INTRODUCTION

An The relationship between a dependent variable and one or more independent variables is described by linear regression models. The objective is to identify the ideal straight line that reduces the linear regression model's sum of squared residuals. The most popular technique for estimating the regression line is the least squares method. Simple linear regression is the type of linear regression in which there is only one independent variable; multiple linear regression, on the other hand, involves numerous independent variables. The following is the definition of the multiple linear regression model

$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_pX_{ip}+\varepsilon_i$, i=1, 2,…n

where yi is the dependent variable, $X_{i1}\ldots X_{ip}$ is independent variable (predictor), β_0 the intercept, and $\beta_1\ldots\beta_n$ are regression coefficients. The value of ε_{ii} represents the error residual. The model is considered as a matrix, with each row represent a data point. In matrix form it can be written as

$Y=X\beta+\varepsilon$

Using ordinary least squares estimation, the vector of estimated regression coefficients is $\beta\hat{}= (X^TX)^{-1}X^TY$

2. METHODOLOGY

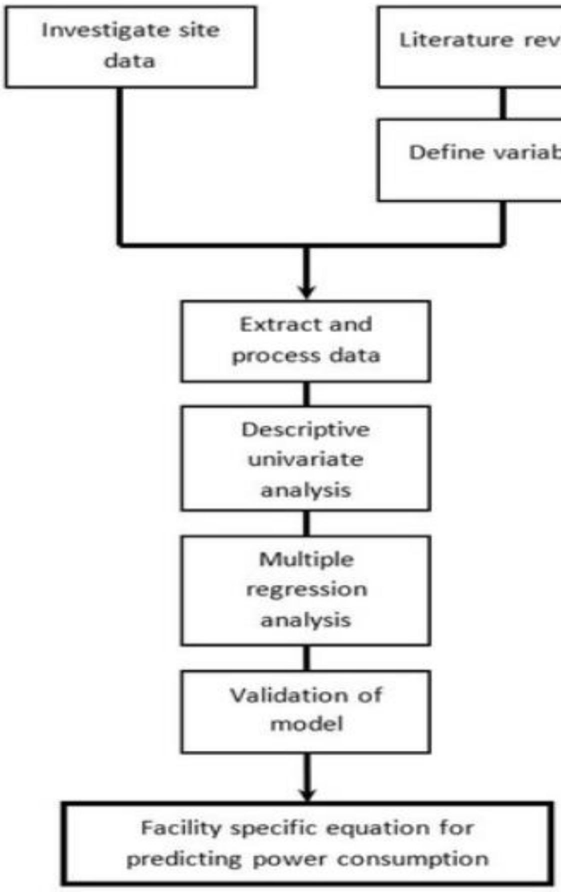
Stepwise regression is ideal for several independent variables. Using the fewest possible independent variables, stepwise regression seeks to optimize estimated performance. Stepwise regression, which combines forward and reverse selection with an automated method for selecting independent variables, is simply explained as follows:

- Establish a launch model with terms that are predetermined.
- Define the parameters for the final model. This includes the kind of model you require and whether to employ square terms or linear terms.
- Establish a cutoff point for the evaluation (in this example, if the overall root-mean-squared error, or SSE, is considerably decreased).
- After terms are added or removed, retest the model.
- If there is no more improvement in the estimated value, stepwise regression will terminate.
- Regression in steps If there is no more progress in the predicted value,

Every time a variable is introduced, there is a forward selection change where all potential variables in the model are examined to see whether or not their importance is less than a certain threshold. Forward selection is applied to each variable in the model one at a time, beginning with none. A variable will be eliminated from the model if it turns out to be unimportant. Although it eliminates the variable if it proves to be unimportant, back select also functions effectively. Stepwise regression hence necessitates two degrees of significance. The first involves adding variables, and the second involves removing them.

3. PROPOSED SYSTEM

Create a model that accurately predicts outcomes by following a few step sequences, as illustrated in Figure 1, using the sales dataset from Big Mart.We provide a stepwise regression model for this. Every stage is crucial to the creation of the suggested model. The 2018 Big Mart dataset was utilized by the model. I utilized data visualization to compare various items in the dataset and analyze the behavior of various independent variables after preprocessing and inputting the missing values. We determined the optimum model for sales forecasting by comparing several selection processes, and we used the square of multiple determination, or R^2, approach to determine the model's accuracy.


Investigate site data
Literature review
Define variables
Extract and process data
Descriptive univariate analysis
Multiple regression analysis
Validation of model
Facility specific equation for predicting power consumption

DATA SET

The dataset is stored in two CSV files (Train.csv and Text.csv) and is logged in Kaggle. A "Test" (5681) and a "Train" (8523) dataset are available. The "Train" dataset includes both input and output variables. The test dataset's sales must be predicted.

- Item Weight: Product Weight;
- Item Identifier: Unique Product ID;
- Item Fat level: Is the product's fat level low or not;
- Item Visibility: Percentage of the store's total display area allotted to a certain product Product Category to which the store belongs:
- Item Type; Product List Price:
- Item MRP;
- Outlet Identifier:
- Special store ID
- Outlet_Size: The shop's size on the covered floor space;
- Outlet_Establishment_Year: The year the store began;
- Outlet_Location_Type: The location of the store Type of city
- Outlet_Type: Indicates if the establishment is a supermarket or merely a grocery shop.
- Item_Outlet_Sales: Merchandising products at a certain retailer.

4. EXPOLATORY DATA ANALYSIS

In this stage, the dataset was examined to extract pertinent data information. It looks for patterns in the existing data and theories. This suggests that missing values are an issue for the outlet size and item weight characteristics. Additionally, an item's minimal visibility is zero, which is essentially unattainable. Outlets were established in a variety of years between 1985 and 2009. It's possible that these values don't fit this format. As a result, you must convert them in accordance with the age of the specific outlet. Ten distinct stores and 1559 unique goods make up the dataset. There are sixteen distinct values in the item type attribute.

As was noted in the preceding section, there are missing values for the characteristics Outlet Size and Item Weight. In our work, we substitute the mean of that specific attribute for the missing values of Item Weight and the mode of that attribute for the missing values of Outlet Size. The correlation between the imputed qualities is reduced when the mean and mode are substituted for the missing numerical data. We are presuming that the measured characteristic and the imputed attribute have no connection for our model.

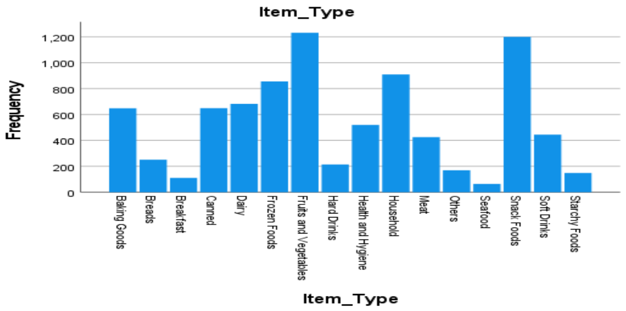
Observations from data set

- The data set shows that the attributes Outlet size and Item weight face the problem of missing values, also the minimum value of Item Visibility is zero which is not actually practically possible.
- There are 1559 unique products, as well as 10 unique outlets, present in the dataset.
- The attribute Item type contains 16 unique values.
- Whereas two types of Item Fat Content are there but some of them are misspelled as regular instead of 'Regular' and low fat, LF instead of Low Fat.

DATA VISUALIZATION

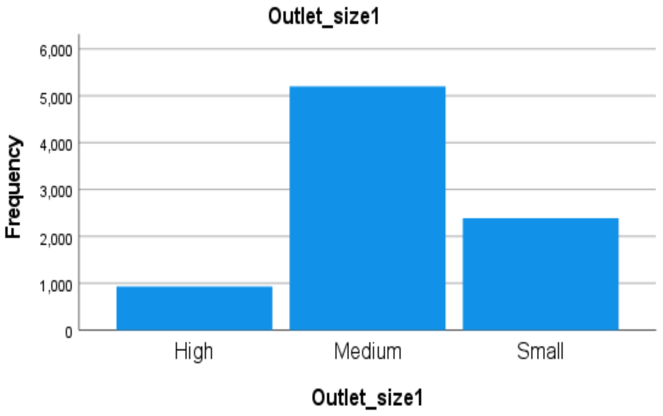
Distribution of the variable Item_Type

#Understanding to item_type per year with respective to mean of each respective year item outlet sales



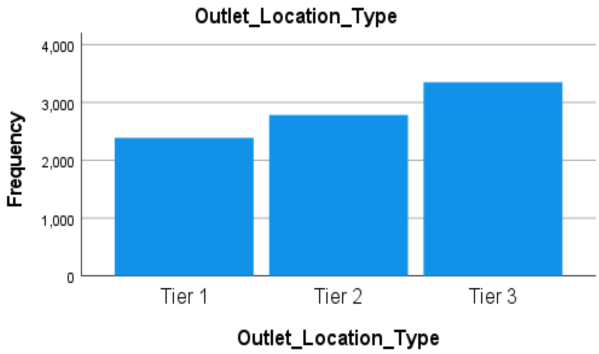
For Item_Type we have 16 different types of unique values and it is high number for categorical variable. Therefore we must try to reduce it.

Distribution of the variable Outlet_Size #Understanding to Oulet_Size



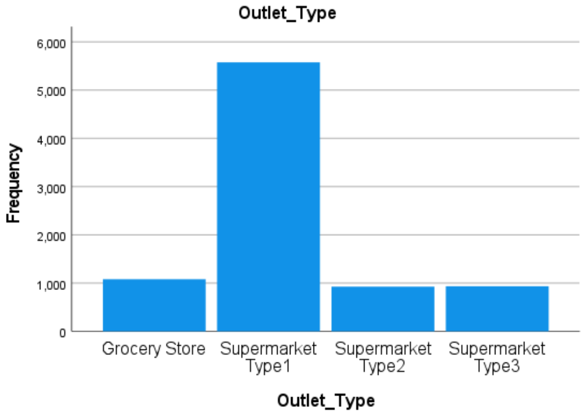
There seems to be less number of stores with size equals to "High". It will be very interesting to see how this variable relates to our target.

Distribution of the variable Outlet_Location_Type



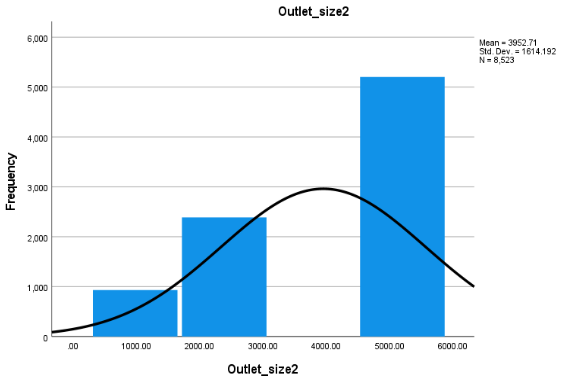
From the above graph we can see that Bigmart is a brand of medium and small size city compare to densely populated area.

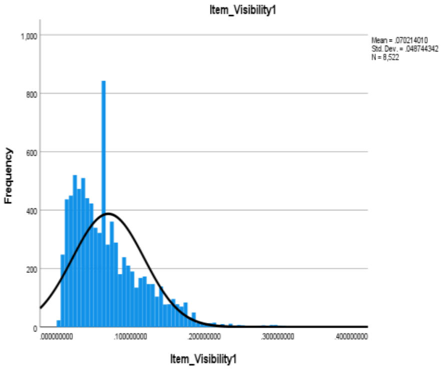
Distribution of the variable Outlet_Type

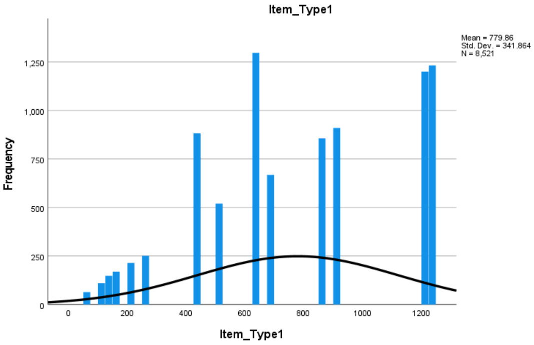


There seems like Supermarket Type2 , Grocery Store and Supermarket Type3 all have low numbers of stores, we can create a single category with all of three, but before doing this we must see their impact on target variable.

VISUALIZING THE SKEWNESSS OF THE DATASET







5. MODEL BUILDING

Fitting Linear Regression Model

Linear Regression Model fitted**: Enter**

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT Item_Outlet_Sales
/METHOD=ENTER Item_MRP Outlet_Establishment_Year Item_Weight_1 Item_Visibility1
Outlet_Location_Type1 Outlet_Type1 Item_Type1 Item_Fat_Content1 Outlet_size2.
```

R-Square=0.625(63%)

Linear Regression5 Model fitted: **Backward**

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT Item_Outlet_Sales
/METHOD=BACKWARD Item_MRP Outlet_Establishment_Year Item_Weight_1 Item_Visibility1
Outlet_Location_Type1 Outlet_Type1 Item_Type1 Item_Fat_Content1 Outlet_size2.
```

R_Square=0.622(62%)

Fitted Linear Regression Model is

Output sales=343.437+2.058+0.684+1.952+0.562+29.731)

Accuracy of the Linear Regression Model=62%

6. CONCLUSION

A firm's ability to make money is closely correlated with how accurately sales are predicted; Big Marts strive for higher sales forecasting to ensure that their business will never experience a loss. In order to estimate product sales from every chosen outlet, we have developed a predictive model in this post utilizing Multiple Linear Regression and the 2018 Big Mart dataset. The anticipated outcomes might be quite advantageous for the company's leaders in terms of approximating their sales and earnings. This may even inspire ideas for brand-new Big Mart sites.

7. REFERENCES

[1] Shrivas, T.: Big mart dataset@ONLINE (Jun 2013),analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/

[2] Smola, A.J., Sch¨olkopf, B.: A tutorial on support vector regression. Statistics and computing 14(3), 199–222 (2004)

[3] Chu, C.W., Zhang, G.P.: A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of production economics 86(3), 217–231 (2003)

[4] Makridakis, S., Wheelwright, S.C., Hyndman, R.J.: Forecasting methods and ap- plications. John wiley & sons (2008)

[5] Punam, K., Pamula, R., Jain, P.K.: A two-level statistical model for big mart sales prediction. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON). pp. 617–620. IEEE (2018)

[6] Schneider A, Hommel G, Blettner M. Linear regression analysis: Part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107:776-82

[7] Freedman DA. Statistical Models: Theory and Practice. Cambridge, USA: Cambridge University Press; 2009.

[8] Elazar JP. Multiple Regression in Behavioral Research: Explanation and Prediction. 2nd ed. New York: Holt, Rinehart and Winston; 1982

[9] Makridakis, S., Wheelwright, S.C., Hyndman, R.J.: Forecasting methods and applications. John wiley & sons (2008).

[10] Kadam, H., Shevade, R., Ketkar, P. and Rajguru.: "A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression." (2018)

[11] C. M. Wu, P. Patil and S. Gunaseelan: Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2018).

[12] Das, P., Chaudhury: Prediction of retail sales of footwear using feed forward and recurrent neural networks (2018)

[13] Das, P., Chaudhury, S.: Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2007)

[14] Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. Cambridge University, UK, 32, 34.

[15] Saltz, J. S., & Stanton, J. M. (2017). An introduction to data science. Sage Publications.

[16] Shashua, A. (2009). Introduction to machine learning: Class notes 67577. arXiv preprint arXiv:0904.3664.