ISSN: 2278-4632
Vol-12 Issue-12 No.02, December 2022

DATA VISUALIZATION AND MULTIPLE LINEAR REGRESSIONS FOR BIGMART SALES PREDICTION

Lakshmi Karanam Assistant Professor BBCIT, Kachiguda Hyderabad.
Karanam.lakshmi12@gmail.com

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: Sales Prediction, Big Mart, Data Visualisation, Multiple Linear Regression, Selection of best Regression.

INTRODUCTION

A Linear regression models describes the relationship between a dependent variable and one or more independent variables. The goal is to find the optimal straight line that minimizes the sum of squared residuals of the linear regression model. The principle of least square method is the most commonly used method for approximating the regression line. When a linear regression has only one independent variable it is simple linear regression and multiple linear regression contains more than one independent variable.

The multiple linear regression model is defined as follows:

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

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 $\hat{\beta}=(X^TX)^{-1}X^TY$

2. METHODOLOGY

For multiple independent variables, it is appropriate to use stepwise regression. The goal of stepwise regression is to maximize estimated performance using the minimum number of independent variables. Stepwise regression is a combination of forward and reverse selection that includes an automated independent variable selection process and can be easily described as follows:

- Set up a launch model that contains predefined terms.
- Set limits for the final model – The type of model you need, whether you use linear terms, square terms, or vice versa.
- Set an evaluation threshold (in this case, whether the total root-mean-squared error (SSE) is significantly reduced).
- Retest the model after adding or removing terms.

ISSN: 2278-4632
Vol-12 Issue-12 No.02, December 2022

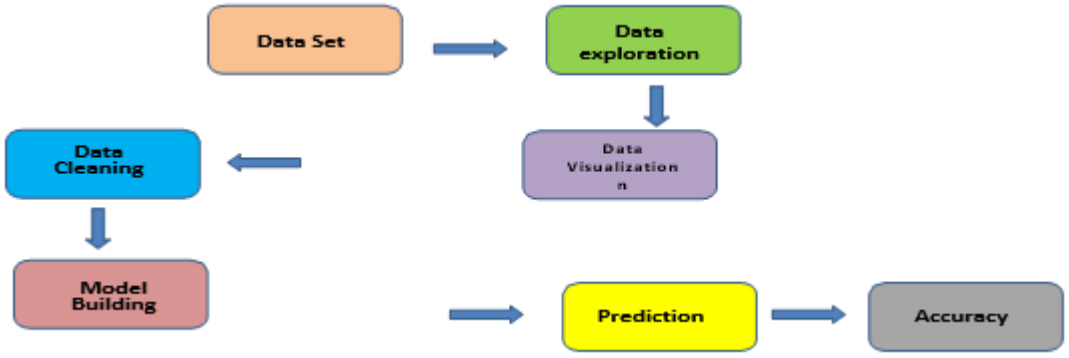
- Stepwise regression will stop if no further improvement in the estimated value occurs.

After each step in which the variable is added, there is a forward selection change in that all candidate variables in the model are checked to see if their significance is below a certain tolerance level. Forward selection starts with no variables in the model and is repeated to add each variable. If a variable that is not important is found, it will be removed from the model. Back select works as well, but removes the variable if it turns out to be insignificant. Therefore, stepwise regression requires two levels of importance. One is to add variables and the other is to remove variables.

3. PROPOSED SYSTEM

Use Big Mart's sales dataset to build a model that predicts accurate results perform a few step sequences, as shown in Figure 1.For this we propose a model that uses stepwise regression. Each step plays an important role in creating the proposed model. The model used the 2018 Big Mart dataset. After pre processing and entering the missing values, I used data visualization to examine the behaviour of different independent variables and compare different items in the dataset. By comparing different selection procedures, we proposed the best model for sales forecasting and applied the square of multiple determination– R2 method to find the model's accuracy.

PROPOSED MODEL



DATA SET DESCRIPTION

The dataset is recorded in Kaggle and consists of two CSV files (Train.csv and Text.csv). There is a dataset "Train" (8523) and a "Test" (5681), and the dataset "Train" has both input and output variables.We need to forecast the sales of test dataset.

- Item_Identifier: Unique Product ID
- Item_Weight: Product Weight
- Item_Fat_Content: Fat content in the product is low or not
- Item_Visibility: Percentage of total display area for all products in the store assigned to a particularproduct
- Item_Type: Product Category to which the store belongs
- Item_MRP:list price of the product
- Outlet_Identifier: Unique store ID
- Outlet_Establishment_Year: Year when the store opened
- Outlet_Size: Size of the store on the covered floor area
- Outlet_Location_Type: The store is located City type
- Outlet_Type: Whether the store is just a grocery store or a supermarket
- Item_Outlet_Sales: Selling goods at a particular store. This outcome variable to be predicted..

6. DATA EXPLORATION

During this phase, useful information about the data was extracted from the dataset. It seeks to identify information from hypotheses and available data. This indicates that the outlet size and item weight attributes are facing the problem of missing values. Also, the minimum visibility of an item is zero, which is virtually impossible. Establishment year of Outlets vary in year from 1985 to 2009.

ISSN: 2278-4632
Vol-12 Issue-12 No.02, December 2022

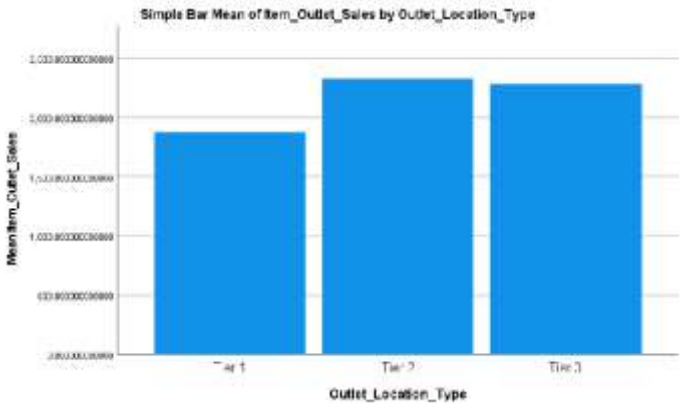
These values may not be appropriate in this format. Therefore, you need to convert them according to age of a particular outlet. The dataset has 1559 unique products and 10 unique outlets. The item type attribute contains 16 unique values. There are two types of fat content in an item, some of which are normal misspellings rather than normal low fats and LF instead of low fats. Response variable, Item outlet sales were positively skewed. Therefore, a log operation was performed on the item's outlet saleto remove the skewness of the response.

DATA CLEANING

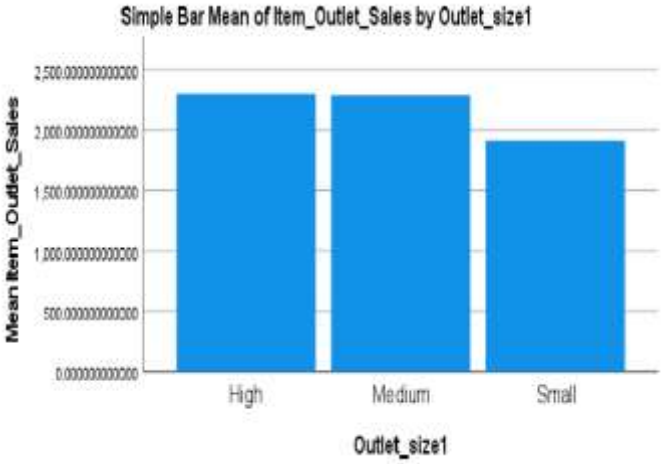
It was observed from the previous section that the attributes Outlet Size and
Item Weight has missing values. In our work in case of Outlet Size missing value we replace it by the mode of that attribute and for the Item Weight missing values we replace by mean of that particular attribute. The missing 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.

DATA VISUALIZATION

Understanding outlet_Location_type depending on item_outlet_sales. It is observed that sales in Tier2 more whencompare to Tier 1 and Tier 3

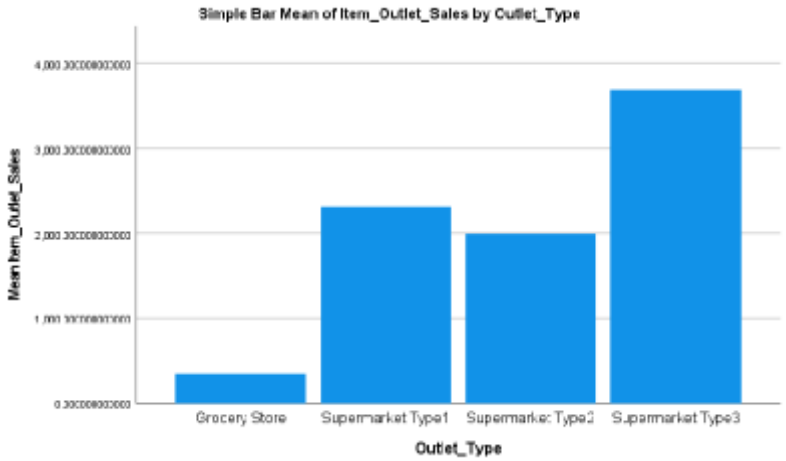


 # Understanding outlet_size depending on item_outlet_sales Small size store have less sales when compare toHigh and Medium size outlets



ISSN: 2278-4632
Vol-12 Issue-12 No.02, December 2022

 # Understanding to outlet_type with respective to mean item outlet sales. Outlet sales are very less in GroceryStores when compare to other Supermarkets



MODEL BUILDING

After completing the previous phases, the dataset is now ready to build proposed model. Once the model is build it is used as predictive model to forecast sales of Big Mart. In our work, we propose a model multiple regression Methods – All possible and backward stepwise Regression. The main purpose of this analysis is to know to what extent is the sales of retail store influenced by the 11 independent variables and what are those measures that should be taken based on the results obtained with using SPSS.

After completing the data visualization and model building the dataset is now ready to build the proposed modeland this model is used to predict sales at Big Mart. In our work, we propose a model that uses multiple regression methods, including backward stepwise regression. The main purpose of this analysis is to find out how much influence the eleven independent variables have on retail store sales, and what measures should be taken based on the results.

BACKWARD STEPWISE

BACKWARD STEPWISE REGRESSION is a stepwise regression method that starts offevolved with a full (saturated) version and at every step steadily eliminates the variables from the regression to get best reduced model which explains the data which also referred to as Backward Elimination regression. The stepwise method is beneficial as it reduces the predictors, decreasing the multi collinearity hassle and it is the one of the method to solve the over fitting.

PROCEDURE

- **Run: STATISTICS->REGRESSION -> BACKWARD STEPWISE REGRESSION...**
- **Select Response variable and Predictor variables .**
  - **REMOVE IF ALPHA > choice defines the Alpha-to-Remove value. At every step pick up the variable forelimination – variables, whose partial F p-value is more or same to the alpha value.**
- **Select the SHOW CORRELATIONS choice to consist of the correlation coefficients matrix to the report.**
  - **Select the SHOW DESCRIPTIVE STATISTICS which consist of the mean, variance and s.d. of every timeperiod to the report.**
  - **Select the SHOW RESULTS FOR EACH STEP to reveal the regression model and summary for everystep.**

R^2 (COEFFICIENT OF DETERMINATION, R-SQUARED) - is the square of the Multiple determination among the PREDICTOR variables and dependent (response variable). In general, R^2 is a percent of variation in explained by the more than one dependent variable .That means R^2 gives the accuracy of the prediction. The largeR^2 is, the total variation in dependent variable reduced to greater extent. The definition of the R^2 is

ISSN: 2278-4632
Vol-12 Issue-12 No.02, December 2022

$$R2 = 1 - \frac{ss\ error}{ss\ total}$$

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.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.87	.757	.758	1316.1536288

a. Predictors: (Constant), Outlet_size2, Item_Type1, Item_Visibility1, Item_Weight_1, Item_MRP, Item_Fat_Content1, Outlet_Location_Type1, Outlet_Establishment_Year, Outlet_Type1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	253.680	19.303		13.142	<.001
	itemmrp	1.026	.016	.531	65.049	.000
	itemvisibility	.530	.021	.292	25.719	<.001
	outletlocationtype	2.382	.073	.324	32.461	<.001
	itemfatcontent	-.026	.029	-.007	-.901	.368
	outlettype1	.683	.015	.556	46.275	.000
	establishmentyear	-36.829	2.574	-.152	-14.308	<.001

a. Dependent Variable: itemoutletsales

R-Square=0.757 (75.7%)

Linear Regression 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
```

ISSN: 2278-4632
Vol-12 Issue-12 No.02, December 2022

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Outlet_size2, Item_Type1, Item_Visibility1, Item_Weight_1, Item_MRP, Item_Fat_Content1, Outlet_Location_Type1, Outlet_Establishment_Year, Outlet_Type1[b]	.	Enter
2	.	Item_Weight_1	Backward (criterion: Probability of F-to-remove >= .100).
3	.	Item_Type1	Backward (criterion: Probability of F-to-remove >= .100).

a. Dependent Variable: Item_Outlet_Sales
b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1		.757	.758	1316.1536288
2		.757	.758	1316.0961823
3	.87	.757	.758	1316.0404836

a. Predictors: (Constant), Outlet_size2, Item_Type1, Item_Visibility1, Item_Weight_1, Item_MRP, Item_Fat_Content1, Outlet_Location_Type1, Outlet_Establishment_Year, Outlet_Type1
b. Predictors: (Constant), Outlet_size2, Item_Type1, Item_Visibility1, Item_MRP, Item_Fat_Content1, Outlet_Location_Type1, Outlet_Establishment_Year, Outlet_Type1
c. Predictors: (Constant), Outlet_size2, Item_Visibility1, Item_MRP, Item_Fat_Content1, Outlet_Location_Type1, Outlet_Establishment_Year, Outlet_Type1

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	253.680	19.303		13.142	<.001
	itemmrp	1.026	.016	.531	65.049	.000
	itemvisibility	-.530	.021	-.292	-25.719	<.001
	outletlocationtype	2.382	.073	.324	32.461	<.001
	itemfatcontent	-.026	.029	-.007	-.901	.368
	outlettype1	.683	.015	.556	46.275	.000
	establishmentyear	-36.829	2.574	-.152	-14.308	<.001
2	(Constant)	253.437	19.301		13.131	<.001
	itemmrp	1.026	.016	.531	65.060	.000
	itemvisibility	-.530	.021	-.292	-25.722	<.001
	outletlocationtype	2.382	.073	.324	32.461	<.001
	outlettype1	.683	.015	.556	46.277	.000
	establishmentyear	-36.826	2.574	-.152	-14.307	<.001

a. Dependent Variable: itemoutletsales

After Backward stepwise
Fitted Linear Regression Model is
Output sales=253.437+1.026+.530+2.382+.683+(-36.826)
Accuracy of the Linear Regression Model=75.7%

CONCLUSION

The profit made business organization is directly proportional to the accurate sales; the Big marts are aims for greater prediction in order that the business enterprise will now no longer suffer any losses. In this article we've designed a predictive model with the aid of using Multiple Linear Regression it at the 2018 Big Mart dataset for predicting sales of the product from all selected outlet. The expected results may be very beneficial for the executives of the business organization to estimate approximately their sales and profits. This can even gives the ideas for new places and new locations of Big Mart.

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)

ISSN: 2278-4632
Vol-12 Issue-12 No.02, December 2022

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. arXivpreprint arXiv:0904.3664.

17. MacKay, D. J., & Mac Kay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge university press.

18. Daumé III, H. (2012). A course in machine learning. Publisher, ciml. info, 5, 69.

18. Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

19. Cerrada, M., & Aguilar, J. (2008). Reinforcement learning in system identification. In Reinforcement Learning. Intech Open.

19. Welling, M. (2011). A first encounter with Machine Learning. Irvine, CA.: Universityof California, 12.

20. Learning, M. (1994). Neural and Statistical Classification. Editors D. Mitchie et. al, 350.

21. Mitchell, T. M. (1999). Machine learning and data mining. Communications of the ACM, 42(11), 30-36.

22. Downey, A. B. (2011). Think stats. " O'Reilly Media, Inc.".

23. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.