ISSN: 0974-5823 Vol. 7 No. 1 January, 2022

International Journal of Mechanical Engineering

Machine Learning Multiple Linear Regression Algorithm for Fast Moving Consumer Goods: An Exploratory Research

Dr.SK.Dhastagiri Bhasha

Associate Professor, Department of Management Studies

PBRVITS, Kavali

Dr.Shaik Karim

Associate Professor, Department of Management Studies

Sree Vidyanikethan Institute of Management, Tirupati

Dr.K.Vidyasagar

Assistant Professor, Department of Management Studies

Sree Vidyanikethan Institute of Management, Tirupati

Dr.K.Balaji

Assistant Professor, Department of Management Studies

Sree Vidyanikethan Institute of Management, Tirupati

Dr.P.Chenchu Reddy

Assistant Professor, Department of Management Studies

Vikrama Simhapuri University, SPSR Nellore (Dt), Andhra Pradesh

Abstract:-

Purpose/Aim: - The aim of the exploratory research study was accurate prediction of sales using factors of sales forecasting which facilitate the manufacturers and retails, especially in the fast moving consumer goods. *Outcome:* - The outcome of the research facilitates to forecast the sales of the company in short run and long run to maximize the sales followed by customer satisfaction. *Research Methodology/Approach:*-It is an exploratory research design, where researcher used structure questionnaire with a sample of 150 and applied multiple linear regression analysis using R-Programming. The R2 value is above 90% which explains 90% of the variance in the sales being explained by the independent variables of product differentiation, market focus, offers and benefits and customer relationship. *Practical Implications: -The* outcome of the research will facilitate to predict the sales of the FMCG product under any market segment in India. *Social Impact:* - The outcome of the research is having social impact, where maximum customer satisfaction can be understood with the help of developed concept. *Originality/Novelty:* - The research is new in the contemporary research, the same factors might not have used in any research to predict the sales of the company. *Type of the Research:*-It is an exploratory research design, where factors used in the research are extremely unique to estimate or to predict the sales of the company.

Index Terms:-Multiple Linear Regression, machine learning algorithm, fast moving consumer goods, exploratory research

INTRODUCTION

The sales forecasting of FMCG is essential concept of production planning in manufacturing organizations. Unless no proper sales forecasting it will seriously impact on production schedule. Procurement of factors of production is mainly depends up on the sales forecasting. The fast moving consumer goods market is a market leader in the competitive economy. The machine learning algorithms plays a crucial role predicting the future aspects of the sales especially the predictive analytics. The present research study mainly concentrates on what are the various factors which will impact on sales forecasting in the competitive edge.

LITERATURE REVIEW

In fact, the holistic nature of approach it is purely depends on the sales forecasting.[1]. Lacking the proper demand forecasting the inventories is ever increasing which becoming serious issue in later stages. The major categories of products which falls under this category are: electronic goods, fashions, apparels, shoes and footwear's are the basic components goods of this [2]. The time series analysis is the sophisticated concept of forecasting the sales, based on the historical trends of the analysis of the sales [3]. There

are many concepts of time series analysis like: seasonality and non-seasonality. Mere simple statistics are not sufficient to forecast the sales of the company need to apply the concept of machine learning applications. There are some techniques like: artificial neural network, decision tree analysis and regression models [4]. It is evident from the machine learning algorithms are giving better results in the forecasting of short-term sales [5]. The machine learning algorithms are giving best results like: ANN and other sorts [6]. The Deep learning algorithms is a more complex form of machine learning computer algorithm which gives better results in performing the statistical analysis [7]. In fact, the machine learning algorithms are giving the best results while extracting the reliable and valid results in the contemporary research in any segment. The sales/demand forecasting with the machine learning algorithms are giving the good results in the especially by applying the time series analysis tool integrated moving average (ARIMA) model in the contemporary research in sales forecasting [8]. The greater accuracy of the sales forecasting will be done through machine learning models whereas the traditional models with the manual statistical tools are not up to the mark with respect to sales forecasting [9]. Introduced machine learning networks sales data with the forecasting techniques in the retail sales forecasting for better production decisions which facilitates to increase the revenue [10]. The Bayesian networks analysis based on conditional probabilities are enhancing or increasing the accuracy of the sales forecasting [11]. Therefore, it is evident that the machine learning algorithms are giving the accuracy while forecasting the demand or sales which mainly influences the production.

NEED & IMPORTANCE

Estimation of factors which impact majorly on sales is being influenced by various factors like: product differentiation offers and benefits, market focus and customer relationship. Improving the sales of the company plays a significant role in the contemporary competition. Therefore, it is evident that there is a need to conduct the research in the area to identify the various factors which impact on to maximize the sales of the company.

SCOPE OF THE RESEARCH

The scope of the research with respect to objectives restricted to estimate the factors which impact on to accelerate the growth of the sales of FMCG. The scope with respect to the location restricted to the Andhra Pradesh.

OBJECTIVES OF THE RESEARCH

1. To review the concepts of machine learning concepts related to sales forecasting.
2. To identify the factors which influence more to develop the sales of the company
3. To suggest the best factors for based on sophisticated machine learning algorithms to develop the sales of the company.

RESEARCH METHODOLOGY

4. Data Sources: Collected both primary and secondary data sources for conducting the research in a systematic manner. In fact, it is a descriptive research design.
5. Statistical Tools: Applied both inferential and descriptive statistical tools to analyze the data in all dimensions. The predictive analytics, multiple linear regression analysis applied to predict the sales of the FMCG goods.
6. Statistical Software: R-Programming software used to analyze the data in all aspects. The machine learning structural equation modeling algorithm used to analyze the data in all aspects.

HYPOTHESIS

7. Ha(1): There is a significant positive relationship between product differentiation and sales of the company.
8. Ha(2): There is a significant positive relationship between offers and benefits and sales of the company.
9. Ha(3): There is a significant positive relationship between market focus and sales of the company.
10. Ha(3): There is a significant positive relationship between customer relationship and sales of the company.

Descriptive Analytics:- There are four factors which influence the sales like: product differentiation, offers and benefits given by the seller, customer relationship and market focus. The following Table 4.1 exhibits the details of the descriptive statistics.

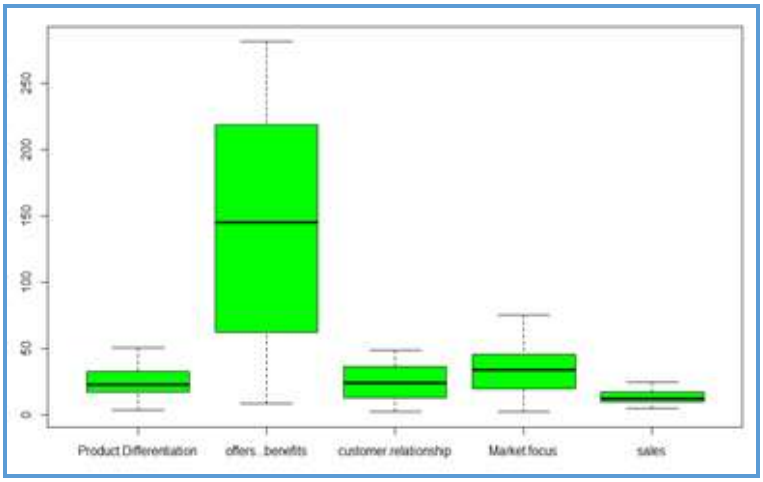
TABLE1: 1Descriptive Analysis

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Sales	1	44	25.06	4.74	25.35	25.30	6.00	14.6	31.4	16.8	-0.32	-0.95	0.71
Product.Differentiation	2	44	50.74	5.29	51.45	50.85	6.75	40.2	58.6	18.4	-0.21	-1.23	0.80
Offers...Benefits	3	44	26.89	3.25	27.10	26.80	4.00	21.3	37.0	15.7	0.48	0.24	0.49
Market.Focus	4	44	20.07	5.16	21.10	20.14	6.38	11.7	28.3	16.6	-0.17	-1.36	0.78
Customer.Relationship	5	44	20.07	6.05	21.55	20.10	5.71	10.0	31.2	21.2	-0.22	-1.15	0.91

Sources: Author Complied

The above Table 4.1 shows the data is following the normal distribution as the skewness and kurtosis values falls in between +/- 2, which explains that the data is following the normal distribution. The total samples of each observation will be 30 followed by the highest mean value of offers and benefits given by the seller in the competitive market followed by the standard deviation will be 89.11 the maximum value will be 281 and minimum value will be 272.8. Therefore, it can be conclude that the data is suitable for conducting parametric tests followed by the correlation among the independent variables can be drawn and the dependency and independency can be tested through multiple linear regression analysis.**Outliers:**The outliers are the extremely different values in the data, which deviates the outcome of the research in different direction. Need to identify the outliers in the data that has to be eliminated which brings the accuracy in the data. The boxplot is a good measurement to identify the outliers existed in the data. The following figure: 1 shows the outliers in the data.

FIGURE: 1 Whisker Box-Plot for Outliers Detection



The above figure.1 depicts the outlier's concept in the data. In fact, the above figure shows that there are no outliers in the data. The outlier points can be seen in the extremely upper side or extremely down side of the lower limit. In fact, to enhance the accuracy of the analysis need to identify the outliers in the analysis that has to be eliminated. The central line in the above diagram shows the Mean value of four different factors. The offers and benefits of the products will have more weightage compared to the remaining product diversification, customer relationship and market focus.

Correlation with Scatterplot: The scatter plot matrix explains about the correlation between the dependent variable and independent variables followed by the linear/non-linear relationship between the variables. Except the market focus remaining all other variables have shown the significant (P<.000) relationship and positive correlations among the variables. The correlations value between sales and product differentiation is(r=0.41*) followed by offers and benefits to sales is (r=0.79*) and customer relationship to sales is (r=0.42) and all the variables have shown liner relationship and some of the variables have not shown the linear relationship with the dependent variable sales.

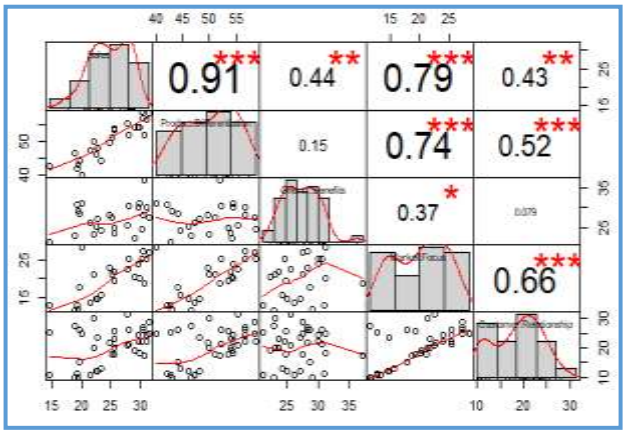
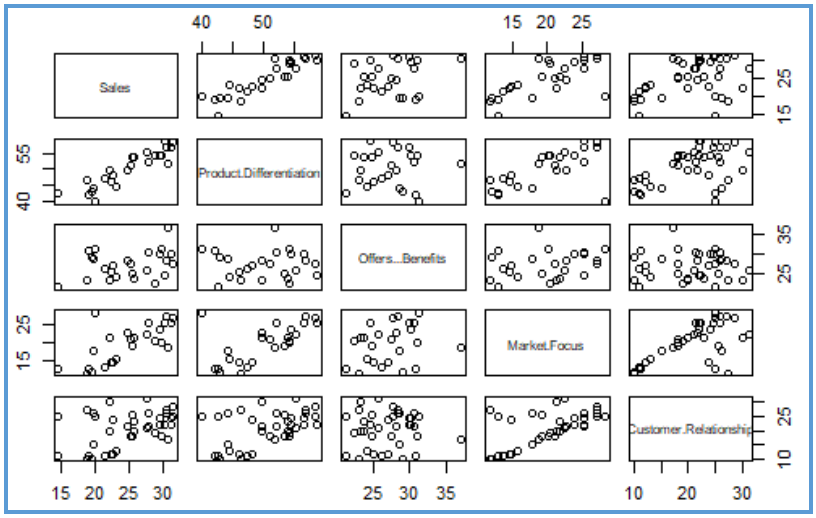


FIGURE: 2: The Correlation Matrix Chart

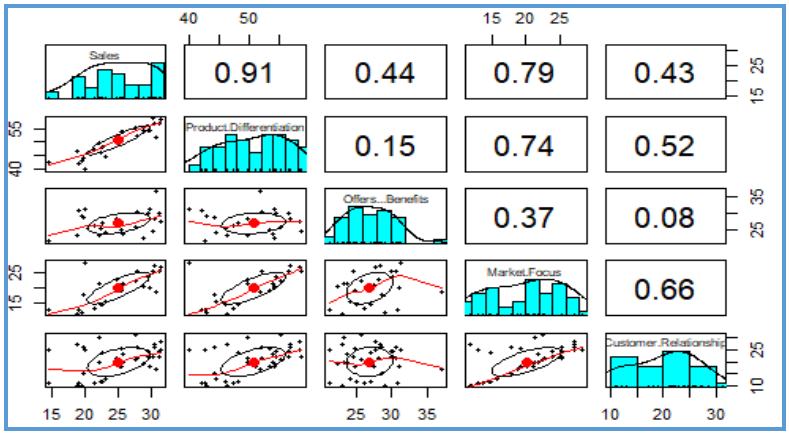
Correlation Matrix Plot: -The following figure: 1 shows the scatter plots among the variables. When the both the variables are increasing trend which explains about there is a positive correlation among the variables. The market focus have not shown any sort of relationship with the sales followed by product differentiation has shown the positive relationship with the sales of the company and offers and benefits also shown there is a significant relationship with the sales. In fact scatter plot is the best measurement for the identification of linear relationship among the variables in the analysis.

Figure: 3 Correlation Matrix Plot



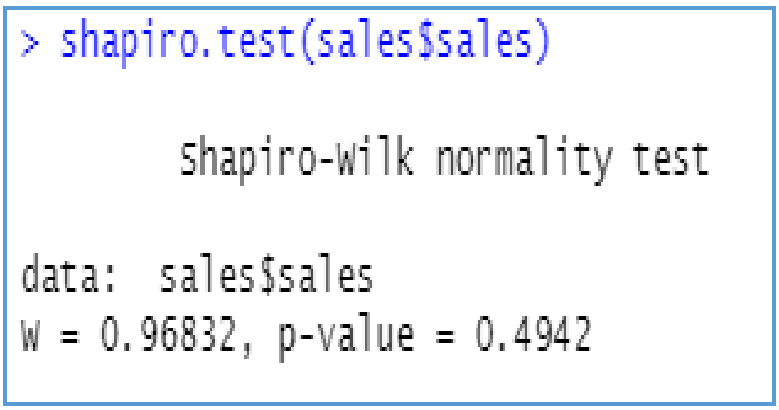
Homoscedasticity:-The residual error plot explains about the homoscedasticity and hetroscedasticity existed in the data. The data has to show homoscedastic relationship among the variables. Which explains about the variance has been equally distributed among the variables. The straight line and variables which are concentrated in the middle which explains about the variance is equally distributed among the variables.

Figure 4: Scatter plot Matrix Analysis



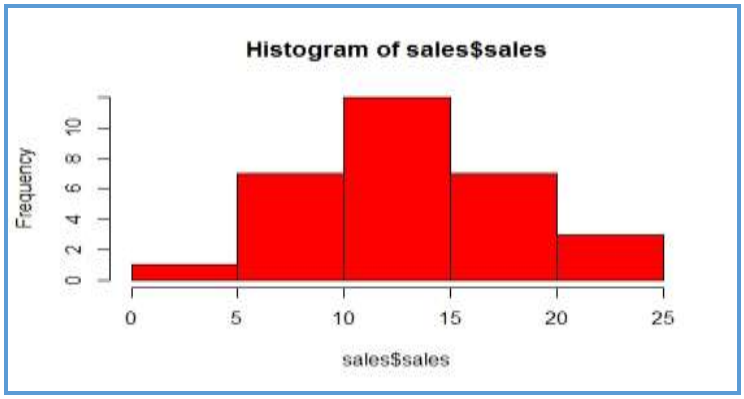
Wilki-Shapiro test: The wilki-Shapiro and histogram are the best to test the normality of the data. If data do not follow the normal distribution it is not possible to conduct the parametric test. The figure 4.3 shows the p-value=0.4942 which is greater than the standard significant value at 5% level 0.05. Therefore, it can be conclude thatthe data is being followed the normal distribution and the w=0.96832 which is wilki-Shapiro value.

Figure 5: Normality test



The histogram, Q-Q plot and P-P plots are the best measurements for to estimate the normality of the data. From the below graph the histogram shows that the data is following the normal distribution. From the wilki-Shapiro value came to know that the p-value 0.4942 which is greater than the standard significant value at 5% level of significance. Hence, Therefore it can be conclude that the data is being followed the normal distribution.

Figure 6: Histogram Analysis for Normality test



Regression Analysis: The regression analysis explains about the relationship between independent and dependent variables in the analysis. In this case the dependent variable is sales and the independent variables are product differentiation, offers and benefits, customer relationship and market focus. The R-Squared value is 0.8981 which is 89% of variance in the dependent variable sales is being explained by two independent variables and its standardized co-efficient are ($p<0.000212$) and customer relationship are ($p<0.0000320$). Therefore, it can be conclude that the offers and benefits and customer relationship plays a significant impact with respect to sales.

Figure .5: Regression Relationship with respect to Dependent & Independent

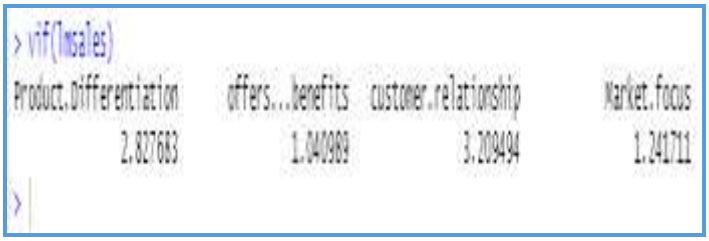
```
Residuals:
    Min      1Q  Median      3Q     Max
-2.2383 -0.6627 -0.1742  0.2349  3.6792

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             -23.08488    2.72529  -8.471 2.26e-10 ***
Product.Differentiation   0.72179    0.05365  13.455 3.10e-16 ***
Offers...Benefits         0.37871    0.06469   5.854 8.27e-07 ***
Market.Focus              0.16882    0.06866   2.459   0.0185 *
Customer.Relationship    -0.10223    0.04255  -2.402   0.0211 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.225 on 39 degrees of freedom
Multiple R-squared:  0.9393,    Adjusted R-squared:  0.933
F-statistic: 150.8 on 4 and 39 DF,  p-value: < 2.2e-16
```

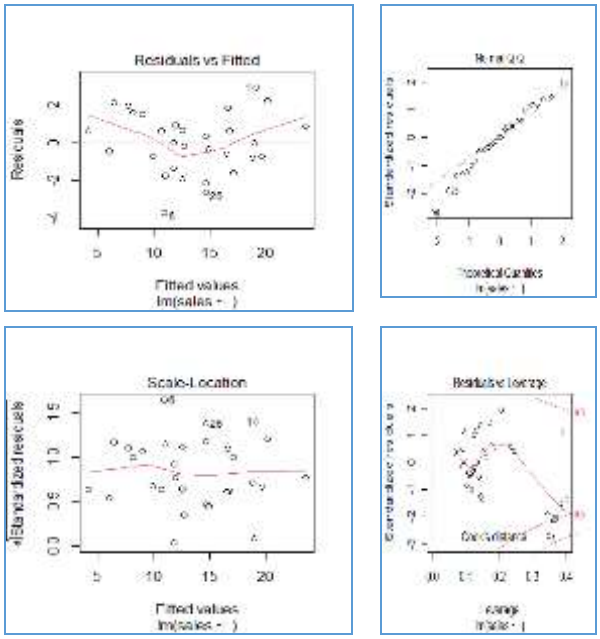
Variance Inflation Factor: - The variance inflation factor explains about the multicollinearity problem in the analysis, if the correlations among the independent variables are ($r=>.70$) then it indicates that there is a multicollinearity problem among the variables. The all the independent variables have proven that the ($VIF<5$), when VIF is less than the 5, it indicates that there is no multicollinearity problem in the analysis. The product differentiation is 2.8276683 followed by offers and benefits 1.040989, customer relationship is 3.209494 and market focus is 1.241711.

Figure.6: Variance Inflation Factor for Multicollinearity



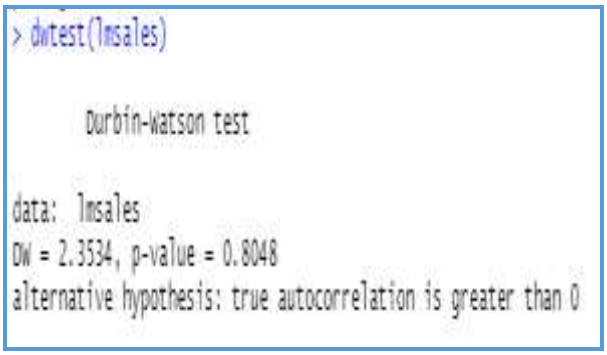
If the variance Inflation factor $VIF<5$, then it tells that there is no multicollinearity issue in the data if the $VIF<10$, there is a little or no multicollinearity in the data, if $VIF>100$, there is a high multicollinearity in the data which explains that the independent variables are highly correlated with each other.

Figure.7: Error residuals & Q-Q plot



The Q-Q plot explains about data normality. If all the points plotted along with the line, which explains about data is being normal distribution. As per the above diagram all the values are plotted on the line with slight deviations which explains that the data is being following the normal distribution. The residual plot also has shown the straight line which explains that data is following the homoscedasticity rather than hetroscedasticity.The cooks distance value which is <1. Therefore, it can be conclude that the data is being followed the homoscedasticity and even following the normal distribution. Auto Correlation:-The Auto Regression explains about the correlation among the successive error terms. The Durbin-Watson test is the best measurement for the estimation of Auto correlation. If the DW<4 is a good measurement which tells that there is no Auto correlation issue in the analysis. If the (DW>4) then there is an Auto correlation issue in the analysis. The following figure.5 explains about there is no auto correlation among the variables.

Figure.8: Durbin- Watson test



Training / Test Data: -The concept of train and test data has been introduced to predict the sales of the company. The total data has been divided in to 80:20 ration. The 80% of the data moved to train data set and followed by remaining 20% of the data is being moved to test dataset. Creating the model with 80% of the and validating the data with test data set. From the train dataset it is also evident that 91.03% of the variance in the sales being explained by offers and benefits, customer relationship and market focus also. The overall significant value is (p<.001) which is less than the standard significant value at 1%. Therefore, the model has shown 90% accuracy in the analysis.

Figure9: Multiple Linear Regression Analysis

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.3397 -0.5845 -0.0050  0.2193  3.5114

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -22.55085    2.74003  -8.230 2.69e-09 ***
Product.Differentiation  0.68674    0.05486  12.517 1.17e-13 ***
Offers...Benefits        0.43486    0.06778   6.416 3.78e-07 ***
Market.Focus             0.15231    0.06933   2.197   0.0356 *
Customer.Relationship   -0.09504    0.04573  -2.079   0.0460 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.175 on 31 degrees of freedom
Multiple R-squared:  0.9436,    Adjusted R-squared:  0.9363
F-statistic: 129.7 on 4 and 31 DF,  p-value: < 2.2e-16
```

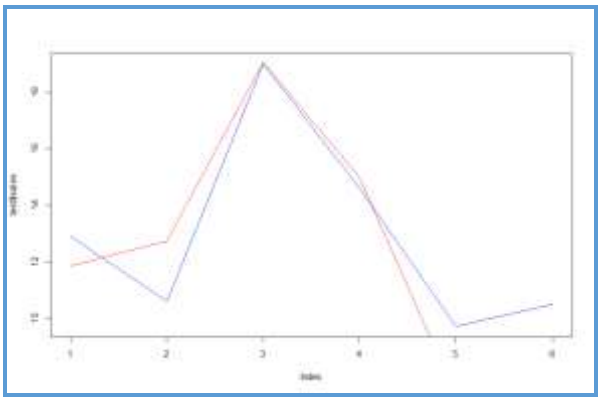
Prediction Analysis: - A model has been created with the help of train data and prediction has been done with the test data weather prediction is happening in right direction or not. The model predicted in such a manner where the actual value is 12.9 and the predicted value is 11.85 for the fifth observation followed by 10.6 for the actual value and the predicted value is 12.73 for the tenth observation and 19.07 for predicted value and the actual value is 19 for the 15th observation and 14.97 of predicted value and 19.0 is the actual observation fir the 20th observation. Therefore, it can be conclude that the accuracy of the analysis around 80%. Therefore, the prediction is happening in right direction.

Table10: Actual (Vs) predicted Sales

```
> #prediction Analysis
> pred <- predict(model, test)
> pred
       5       10       15       20       25       30
11.858062 12.733633 19.073019 14.971430  7.269382  7.546600
```

Predicted value	11.858062	12.73	19.07	14.97	7.26	7.54
Actual value	12.9	10.6	19.0	14.6	9.7	10.5

Figure11: Line chart for mapping Prediction (Vs) Actual Sales



The above line graph shows that the predicted line and actual sales almost similar. Therefore, it can be conclude that our prediction is going on right direction. Even, with the help of current model we can forecast the sales of FMCG with the new data. From the above graph the blue line is actual sales of the FMCG goods from test dataset whereas the red color line is predicted sales from actual dataset. Therefore, both train dataset model and test dataset model both are matching, in equal common parlance.

FINDINGS

1. The 93% of the variance in the dependent variable, the sales of FMCG is being explained by the product differentiation, offers and benefits, market focus and customer relationship.
2. There is a significant positive correlation among the dependent and independent variables.

3. All the independent variables like: product differentiation, offers and benefits, market focus and customer relationship have shown significant relation (.000) with each other.
4. There is no multicollinearity problem in the analysis. That is correlation among the independent variables.
5. The model validated with 90% accuracy. Still the model can be regenerated by adding some other independent variables.
6. No Auto correlation problem in the analysis followed by the data has followed the normal distribution.

CONCLUSION

Therefore, it can be conclude that the factors (product differentiation, market focus, customer relationship, offers and benefits) have shown the significant relationship with the sales of the fast moving consumer goods in the competitive market. In fact, the model can be redeveloped for accuracy by adding the other variables. Hence, the predictive analytics model which is multiple linear regression analysis will give better results to predict the future of the market trend in the competitive market.

REFERENCES

[1] Arvan, M., Fahimnia, B., Reisi, M., Siemen, E. (2018). *Integrating Human Judgment into Quantitative Forecasting Methods:* A Review. Omega.

[2] Nagashima, M., Wehrle, F. T., Kerbache, L., Lassagne, M. (2015). *Impacts of adaptive collaboration on demand forecasting accuracy of different product categories throughout the product life cycle. Supply Chain Management*, Vol.20 (4), 415-433.

[3] Da Silva, I.D., Moura, M.D.C., Didier Lins, I., López Droguett, E., Braga, E. (2017) *Non-Stationary Demand Forecasting Based on Empirical Mode Decomposition and Support Vector Machines*. IEEE Latin America Transactions, 15 (9), art. no. 8015086, 1785-1792.

[4] kabane, G. (2008). Gestão strategic das technologies cognitive: conceitos, methodologies e applications. *Saraiva, São Paulo,*

[5] Lu C.-J., Shao Y.E. (2012). *Forecasting computer products sales by integrating ensemble empirical mode decomposition and extreme learning machine. Mathematical Problems in Engineering.*

[6] Lu, CJ., Chang, CC. (2014). *A Hybrid Sales Forecasting Scheme by Combining Independent Component Analysis with K-Means Clustering and Support Vector Regression*. Scientific World Journal.

[7] Good fellow, I., Bengio, Y and Courville, A. (2015). *Deep Learning. Cambridge*: The MIT Press.

[8] Kandananond, K. (2012). *A comparison of various forecasting methods for auto correlated time series*, International Journal of Engineering Business Management.

[9] Wu J., Zheng S. (2015). *Forecasting for fast fashion products based on web search data by using OS-ELM algorithm*. Journal of Computational Information Systems, 11(14), 5171-5180.

[10] Chen F.L., Ou T.Y. (2011). *Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry*. Expert Systems with Applications.