iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50189

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Data Analyzation using Regression (Sales Prediction)

Mr. Huzefa Babinwale[1], Ms. Apeksha Shriwas[2], Mr. Omkar Porlikar[3], Ms. Pratiksha Kharode[4], Prof. Rohan Kokate[5]
[1, 2, 3, 4]*Author, UG Scholar, IT Dept. JDCOEM, Nagpur*
[5]*Guide, Assistant Professor, IT Dept. JDCOEM, Nagpur*

***Abstract:** This model is made for overcoming the excessive production of any product that occurs by inappropriate analysis of data. Sometimes the requirement is more but product shortage always leads companies to suffer from various types of branding losses because demand for and shortage of products are at the same page of a company's reputation.Our motto is to minimize the advertisement pricing and also preding the demand of time.*

I. INTRODUCTION

We are using the data analysis of previously purchased items of any customer to predict the requirement of the company's product. Predicting sales of a company needs time series data of that company and based on that the model can predict the future sales of that company or product. So, in this research project we will analyze the time series sales data of a company.

We are using factors like review of the customer, Gross income of the customer, and regional requirements. For achieving that goal, we are using database manipulation by SQL and MYSQL.

For achieving the goal of prediction, we are using multiple regression and classifier algorithms. We are using techniques like Google Collab, Online compiler, Pythons Library, Machine Learning Algorithms. By using all these techniques companies can predict the customer by their previous data. Loss is happening due to the excessive production of the products. Because of the shortage of products, customers get disappointed by the service of the company. Company spends a lot of money for the advertisement as well as for the manpower. Machine Learning Algorithm is used in this model to predict the future sales on the basis of past records of the customer and reducing the manpower.

Each person should get the proper products for them. Companies can make better decisions for their customers. It helps in overall business planning, budgeting etc. **Sales** forecasting allows companies to efficiently allocate resources for future growth and manage its cash flow.

II. RESEARCH GAP

From the above literature review, it is recognized that many of the earlier researchers have focused on YouTube advertisement, consumer behavior, brand awareness, and purchase intention. Marketers who use YouTube Advertisements boost their revenue 49%faster than those who don't. YouTube advertisements drive sales. Moreover, other factors like the relationship between YouTube advertising budget and sales are not clearly defined in previous research. Hence there exists a need to understand the relationship between YouTube advertising budget and sales in social media advertisement and we are including a new factor for more reliable results that is budget of customer.

III. RESEARCH DESIGN

Objective

1) To analyze the relationship between previously purchased items and sales.
2) To build a logistic regression model using the training data set.
3) To make predictions of the sales of a test data set using a **logistic regression** model.

IV. HYPOTHESIS

1) *Null Hypothesis:* There is no statistically significant relationship between **sales** and the **advertising budget**.
2) *Alternate Hypothesis:* There is a statistically significant relationship between **sales** and previously purchased items budget

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

V. PREPARING THE DATA

A Marketing dataset, a preloaded dataset in model, is used for the study of this research. This marketing dataset comprises the effect of three advertising media (YouTube, Facebook, and Newspaper) on **sales**. The outcome variable for this research is sales, and the predictor variable is the YouTube advertising/online marketing budget. The last column in the dataset denotes sales while the remaining column denotes different media advertising in 1000 dollars. The dataset consists of 200 rows. Each row denotes a different advertising experiment. Once the data is loaded, a comment called head is used to look at a few samples of the data set. Similarly, to look at the end of the data set, a comment called tail is used. This is done since, in some data sets, the last few columns may comprise totals or summaries of the data, which can be unrelated. Next, the data is divided into the testing data set (30%) and training data set (70%) such that the training dataset will contain 145 rows, and the test dataset will have 55 rows.

VI. DATA ANALYZATION AND INTERPRETATION

Data visualization :- **logistic regression** means the relationship between the outcome and predictor variables are logistic.A logistic relationship can be easily verified using generating a dot plot of the resulting variable(sales) vs. predictor variable (YouTubeadvertising budget). The following R code will appear the relationship between the outcome variable and predictor variable.plot (marketing, marketing sales) and (model, col=2, lwd=3)

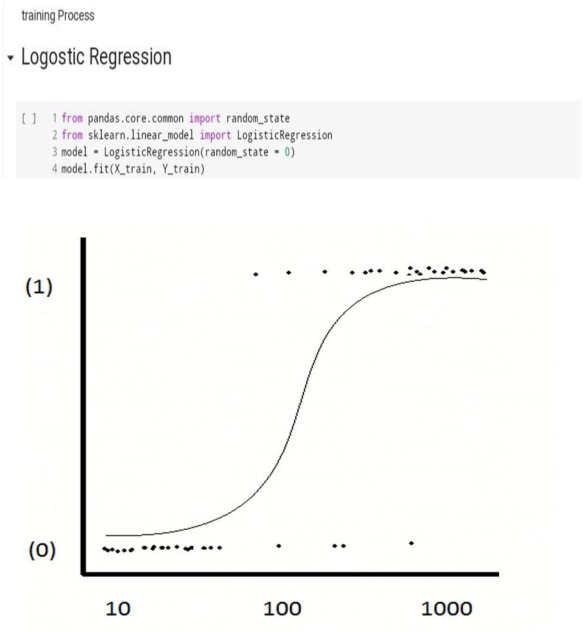
It is clear from the above graph that there is an increasing logistic relationship between the outcome variable (sales) and the predictor variable (budget, previously Buyed & Age), which is a positive aspect. The correlation coefficient between the outcome variable and the predictor variable is calculated using the following R code.

VII. CORRELATION COEFFICIENT

The correlation coefficient cor (marketing sales, marketing $YouTube)0.7822244The level of association between the outcome and predictor variables are measured by correlation coefficient. If the value of the correlation coefficient -1 means a perfect negative relationship between variables and +1 means perfect positive relationship. There is a weak relationship between the two variables if the value is nearer to zero. In this research, the correlation coefficient is 0.78, which indicates a strong positive relationship between the outcome variable and the predictor variable.

VIII. LOGISTIC REGRESSION MODEL

In this paper, a simple logistic model is built to predict sales units based on the advertising budget spent on YouTube. The following code is used to find out the beta coefficients of the model. model. < lm (sales~YouTube, data = market train)



A. Coefficients

(Intercept) youtube
7.821390 0.051034

The above value indicates the beta coefficient and the intercept for the predictor variable. Hence the calculated equation is sales = 7.82 + 0.051*PPI. The intercept (b0) is 7.82, and the coefficient of THE variable is 0.051. Using this equation, for each new YouTube advertising budget, the number of **Sales** units can be predicted. For a YouTube advertising budget equal zero, the expected sale is 7.82 units. For a YouTube advertising budget equal to 1000 dollars, the expected sale is 7.82 + 0.051*1000 = 58.8 units

IX. MODEL SUMMARY

Before using the model for **predictions**, the statistical significance of the model is assessed by exhibiting the statistical summary of the model.

X. INTERPRETATION

The p-value of the t-static can be used to determine whether the null hypothesis can be rejected or not. The higher the t-statistic and the lower the p-value, the more significant the predictor. In this research both p-values for intercept and predictor variables are very significant. Hence the predictor variable advertising on YouTube is significantly related to the **sales** outcome variable. Therefore zero the hypothesis is rejected. It is normal to measure goodness of fit of the **logistic regression** model, once the null hypothesis is rejected. Logistic quality regression fit is typically measured using two related measurements, residual standard error (RSE) a R-squared statistic. In this research for the training data set, the residual standard error is 3.87, which means that the **sales** value deviates from the true regression line by approximately 3.87 units on average. In this average training **sales** value data set is 16.83 so the percent error is 3.87/16.83=22%, which is a low residual error. Thus, by reducing the RSE, the model fits the data best. The value of R square in this research is 0. 8684 The R-squared value typically ranges from 0 to 1 a high R-squared model will be better. logistically the regression model of this research works well accuracy because the R-squared value is high.

XI. SALES PREDICTION

The performance of the regression model is evaluated by making predictions with test data. new data <- data. frame (YouTube = c(0,1000)) predict (model, new data)

1 **2**
7.82139 **58.85568**

The result shows a surge of 58 units in sales for the YouTube advertising budget equal to 1000 dollars. The accuracy of the model is tested by Root mean squared error and R-square values. The model is better with lower RMSE and higher R-square values.

RMSE (res, market_test$sales)
4.069838
R2 (res, market_test$sales)
0.8495986
mean (market_test$sales)
16.65818

The computed R-square value for the test dataset is 0.84. The above R-square value indicates the predicted outcome value by the model is highly correlated with the observed outcome value in the test datasets. The calculated Root mean square value for the test dataset is 4.06. The error rate for the test dataset is 24%(4.06/mean), which is low. Hence the model accuracy is good.

XII. FINDINGS

1) The relationship between YouTube advertising budget and sales is logistic and additive.
2) There is a strong relationship between the outcome variable and the predictor variable.
3) There is a statistically significant relationship between the Previously Purchased Items budget and **sales**.
4) The **logistic regression** model built in this research fits very well with the data.
5) According to the **prediction** of the **logistic regression** model for the YouTube advertising budget of 1000 dollars, an increase of 58 units of sales is expected.
6) If we can shortlist the customer based on previously purchased items then the spending budget will also fall dramatically.

©IJRASET: All Rights are Reserved

XIII. CONFUSION MATRIX

```
[ ]  1 from sklearn.metrics import confusion_matrix , accuracy_score
     2 cn= confusion_matrix(Y_test, y_pred)
     3 print("Confusion Matrix: ")
     4 print(cn)
     5
     6 print (f"Accuracy of model: {(accuracy_score(Y_test,y_pred)*100)}% ")
     7
```

```
Confusion Matrix:
[[61  0]
 [20 19]]
Accuracy of model: 80.0%
```

XIV. CONCLUSION

YouTube, Social Media advertising has realized unbelievable growth in the last decade and is the main marketing channel for many marketers. Online helps to build brand awareness, increase market share, and drive sales. Marketers have conventionally turned to Online for branding. Later they have invested time to understand the impact of YouTube **advertisements** on **sales**. From this study, it is clear that there is a significant relationship between YouTube advertising budget, sales. This research concludes YouTube advertising is a better predictor of company **sales**. If we are shortlisting the peoples/customers then our overall YouTube/online advertisement budget will fall quickly.

REFERENCES

[1] Aburto, L., & Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. Applied Soft Computing, 7(1), 136–144. Retrieved July 27, 2014, from http://linkinghub.elsevier.com/retrieve/pii/S1568494605000311. doi: 10.1016/j.asoc.2005.06.001 [Crossref], [Web of Science ®]Saquib HashmiSaquib Hashmi, KaushtubhSaquib Hashmi, Kaushtubh Saquib Hashmi, Kaushtubh Kumar, Siddhant Khandelwal, Real-Time License Plate Recognition from Video Streams using Deep Learning, International Journal of Information Retrieval Research, (January 2019).

[2] Ait-alla, A., Teucke, M., Lütjen, M., Beheshti-Kashi, S., & Karimi, H. R. (2014). Robust production planning in fashion apparel industry under demand uncertainty via conditional value at risk. Mathematical Problems in Engineering, 2014, 10 pp.

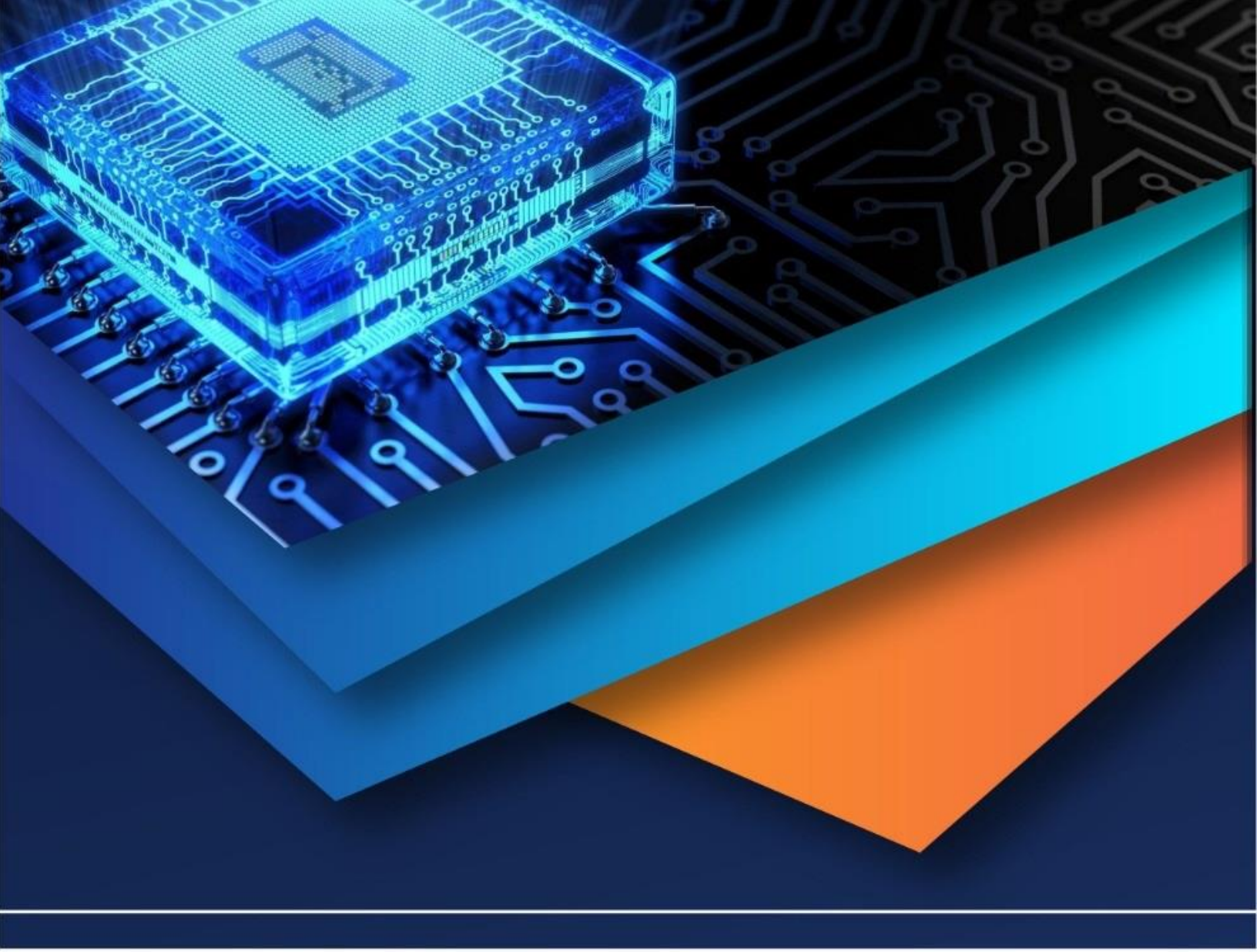
[3] Beheshti-Kashi, S., Karimi, H.R., Thoben, K.D., L¨utjen, M., Teucke, M.: A surveyon retail sales forecasting and prediction in fashion markets. Systems Science &Control Engineering 3(1), 154–161 (2015)

[4] Bose, I., Mahapatra, R.K.: Business data mininga machine learning perspective.Information & management 39(3), 211–225 (2001)3. Chu, C.W., Zhang, G.P.: A comparative study of linear and nonlinear models foraggregate retail sales forecasting. International Journal of production economics86(3), 217–231 (2003)

[5] AshwiniRekha. Banjanagari, Vijaykumar. B, “Retail Giant Sales Forecasting using Machine Learning”, International Journal of Recent Technology and Engineering (IJRTE), Vol. 8, Pgno. 2408–2411.

[6] Khushbu Kumari, Suniti Yadav, “Curriculum In Cardiology – Statistics”, Journal of the practice of Cardiovascular Science. Vol-4, Issue-1, Pgno. 33–36. https://doi.org/10.4103/jpcs.jpcs_8_18

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)