iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54192

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 VI Jun 2023- Available at www.ijraset.com

Predicting Online Customer Purchase using Gradient Boost Classifier

Arushi Sreekumar[1], Renuka Devi S M[2]
[1]*Scholar,* [2]*Professor, ECE Dept., Narayanamma Institute of Technology & Science, Affiliated to JNTUH, Hyderabad TS , India*

Abstract: ***Understanding client buying behavior is crucial for firms in order to enhance targeting consumers, maximize marketing efforts, and boost overall sales success. In order to identify whether a client is inclined to make a purchase or not, the Gradient Boosting algorithm is used in this paper's predictive modeling technique. The Gradient Boosting technique is used to create a model for prediction because it can handle complicated connections and detect non-linear patterns. The technique includes a number of steps, such as feature engineering, model training, and performance evaluation. The data preparation methods used include feature encoding, feature selection, and missing value imputation. On the labeled dataset, the Gradient Boosting classifier is trained, with prediction accuracy being optimized.***
Keywords: Gradient Boosting algorithm, prediction of customer

I. INTRODUCTION

Predicting the future of digital marketing is challenging, as the industry is constantly evolving with new technologies, consumer behaviors, and regulatory changes. However, based on current trends and the direction of the industry, some of the predictions for digital marketing includes

1) Increased Use of Artificial Intelligence (AI):
2) Growing Significance of Voice Search and Smart Assistants:
3) Emphasis on Privacy and Data Protection:
4) Rise of Influencer Marketing and User-Generated Content:
5) Integration of Augmented Reality (AR) and Virtual Reality (VR):
6) Continued Growth of Mobile Marketing:
7) Expansion of Video Marketing:

It's important to note that these predictions are based on the current trajectory of the digital marketing landscape. However, the industry is dynamic, and new innovations and trends can emerge, influencing the future direction of digital marketing.

The digital marketing has opened up the doors of big data analysis. The moment the customer browses for an item, the traces regarding the requirements, type of the customer, the interests, the priorities are all learnt and analysed by the system, resulting in recommendations provided to the customer. This is possible due to predictive analysis on big data using statistical and machine learning algorithms. In [1] prediction of customer behaviour using behavioural informatics is tried for better business decisions.

II. RELATED WORK

In the modern era of advanced technology, it is crucial to anticipate market trends in order to understand consumer behavior as trends tend to be unpredictable. Drawing upon advancements in machine learning and previous research on behavior prediction, our study aims to develop a model that can forecast consumer behavior. In [2] the research is to explore the correlation between various consumer behaviour parameters and their inclination to make purchases. Initially, we investigate the relationship between consumer behavior and changing factors such as the environment, organization, individuals, and interpersonal dynamics. The paper proposes a time-evolving random forest classifier that utilizes innovative feature engineering to accurately predict consumer behavior, which significantly influences their purchasing decisions. The results obtained from the random forest classifier demonstrate higher accuracy compared to other machine learning algorithms.

Customer buying behaviour is influenced by various factors, including personality traits such as quality, motivation, occupation, income level, perception, psychology, references from others, and demographics[3]. In today's world, data mining is commonly employed to study customer shopping activities using diverse algorithms and methods. This technology has gained popularity across numerous industries. Each customer's actions and preferences are recorded as data in a database, capturing information about their purchasing habits, frequented items, and quantities bought, often without their awareness.

The research[3] focuses on analyzing and categorizing customers based on their purchase behavior using a dataset. The classification is conducted using the SVM algorithm, utilizing publicly available inventory and sales data. The performance of the methodology is evaluated through experimentation, demonstrating its effectiveness in understanding customer behavior.

The Internet has witnessed a surge in online shoppers, with the statistics plot showing the buyers in India. To maintain a thriving, diverse, and organized marketplace while meeting the comprehensive shopping needs of consumers, it is crucial to accurately analyze and predict their purchasing behaviors. This study [4] utilizes user-product interaction data from JD.com within a specific timeframe to forecast whether users will make purchases in designated categories and shops in the upcoming week.

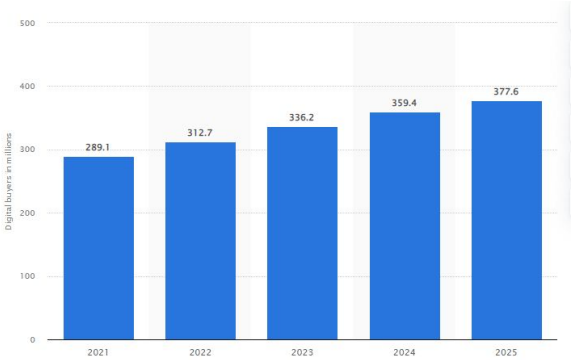


Fig1: Digital customers[4]

To address the challenge of weakly related original features, robust features are constructed using feature fusion techniques. To tackle computational and overfitting challenges in high-dimensional systems, random forests are employed to filter features, reducing model complexity. Finally, a fusion algorithm based on XGBoost and LightGBM is proposed, outperforming random forest and GBDT algorithms, enabling merchants to conduct targeted marketing activities based on accurate predictions.

X. Zhai et al., [5] accurately predicted online customer purchase behaviour which crucial for improving e-commerce performance. Existing studies often rely on limited proxies like purchase intention and sales rank, lacking practical applicability. This research utilizes deep learning techniques on a large dataset of over 50,000 unique web sessions to enhance understanding. Platform engagement and customer characteristics are employed as predictors for online purchases. Comparing to traditional machine learning methods, the deep learning approach outperforms in predictive capability. These findings provide valuable insights for platform designers and contribute to the academic understanding of purchase prediction in e-commerce. In[7] recent decades, predicting customer purchasing behavior has gained significant attention in consumer behavior research. Most business models rely on linear equations to estimate the influence of factors such as age, gender, income, product price, and promotions. However, compared to other research fields, the prediction methods for purchase behavior have been overly focused on linear models. With the increasing volume of data collected through information and communication technologies in retail and marketing, linear models alone are insufficient. As a result, machine learning techniques like Bayes classifier and support vector machine (SVM) are being explored as alternative approaches for knowledge discovery and data mining. In the context of Retail 4.0, there is a growing need for accurate prediction of consumer purchase intentions. To address this, a decision support prediction model has been developed at the attribute level, offering an influential e-commerce platform for customers. Social perception scores of brands and review polarity are computed through social network mining and sentiment analysis. Regression analysis is applied, and suitable instances are identified for each attribute to predict the relevant product attributes. Notably, camera attributes such as sensor, display, and image stabilization attract customer attention during the search. These findings benefit e-commerce retailers and enhance the search platform's efficiency in providing desired durable goods. Sensitivity analysis ensures the model's robustness.

E-commerce platforms are witnessing a rise in online transactions as suggested by Fig 1 and as consumers are increasingly opting for the convenience of digital shopping. Analyzing complex behavioral patterns from these interactions through predictive analytics helps businesses understand consumer needs. However, a comprehensive systematic review of recent research in this area is lacking. This paper fills that gap by presenting a systematic literature review which has its main focus on customer purchase prediction in the E-commerce area. The contributions include an innovative analytical framework and a research agenda. The framework identifies three main tasks: predicting customer intents, buying sessions, and purchase decisions. The employed predictive methodologies are analyzed from different perspectives, and the research agenda highlights key areas for further exploration in online purchase behavior prediction.

III. PROPOSED WORK

The data consists of 22 features, which include Home Page, Home Page Duration, LandingPage, Landing Page Duration, Product Description Page, Product Description Page Duration, Google Metric: Bounce Rates, Google Metric: Exit Rates, Google Metric: Page Values, Seasonal Purchase, Month Seasonal Purchase, OS, Search Engine, Zone, Type of Traffic, Customer Type, Gender, Cookies Setting, Education, Marital Status, WeekendPurchase.

First, we deal with the missing values using a simple imputer. The' SimpleImputer' is a class from the sci-kit-learn library that provides a simple strategy to handle missing values in a dataset. The 'SimpleImputer' class follows a straightforward approach to imputation. It takes a user-defined strategy as a parameter, determining how missing values should be replaced. The available strategies are Mean: Replaces missing values with the mean of the non-missing values in the same column, Median: Replaces missing values with the median of the non-missing values in the same column, Most frequent: Replaces missing values with the most frequent value (mode) of the non-missing values in the same column, constant: Replaces missing values with a specified constant value.

We are using the 'Most frequent' as our strategy, we will perform ordinal encoding on our entire data as there are categorical data present. Simple imputer replaces all the missing values with the most frequent occurrence.

The columns Customer Type, Gender, Cookies Setting, Education, Marital Status, and Month Seasonal Purchase contain categorical values. One hot encoding is used to deal with these columns. One-hot encoding is a popular technique used to convert categorical variables into a binary representation that machine learning algorithms can process more effectively. It creates binary columns for each unique category in the original variable, where each column indicates the presence or absence of that category in a given instance. One-hot encoding is useful because it avoids the issues that can arise from assigning arbitrary numerical values to categorical variables, which could inadvertently introduce misleading relationships. By transforming categorical variables into a binary representation, one-hot encoding enables machine learning algorithms to effectively interpret and utilize categorical information during model training. After performing one hot encoding there is a total of 42 columns in the data frame.

Made purchase is the column that contains true or false based on whether the customer purchased an item or not. We use a label encoding technique to convert it into numerical values, and store it in a separate data frame named labels. The training and testing data is split where it's observed that both the classes have an imbalance, the number of rows corresponding to true is much lesser than the ones in the false class. To address this SMOTE is used. SMOTE (Synthetic Minority Over-sampling Technique) is a popular technique used to address the issue of imbalanced datasets in machine learning. It works by synthesizing new minority class samples based on the existing minority class samples. The process involves finding out a minority class sample then figuring the k nearest neighbors in the feature space. Then, synthetic samples are created by randomly selecting one of the nearest neighbors and creating a new sample along the line connecting the original sample and the selected neighbor. This way, SMOTE increases the number of this minority class samples and balances the class distribution, which helps improve the performance of machine learning models in handling imbalanced datasets. By generating synthetic samples, SMOTE effectively introduces diversity into the minority class, aiding in better generalization and reducing the risk of overfitting. SMOTE is only performed on training data, and auto strategy is employed. The last step in preprocessing is standard scaling, 'StandardScaler' is a popular technique used for feature scaling in machine learning. It works by transforming the features of a dataset to have zero mean and unit variance. The process involves subtracting the mean value of each feature from the dataset and then dividing it by the standard deviation. This normalization ensures that each feature is centered around zero and has a similar scale, which is important for many machine learning algorithms. By standardizing the features, 'StandardScaler' helps in mitigating the impact of varying scales and brings the features to a comparable range, making them more suitable for accurate and reliable model training. Additionally, 'StandardScaler' assumes that the distribution of each feature is approximately Gaussian or follows a bell curve. Standard scaler is trained and applied to the entire data.

IV. RESULTS

The model chosen are the gradient boost classifier, Gradient Boosting Classifier (GBC) is a powerful and widely used machine learning algorithm and Support vector machine (SVM). GBC is versatile and can be applied to both classification and regression problems. GBC is capable of capturing complex nonlinear relationships between features and the target variable. By building an ensemble of weak learners (usually decision trees), GBC combines their predictions to create a powerful model that can capture intricate patterns and interactions in the data. A grid search CV is used to decide the best parameter for the model, and trained on the training dataset, finally, we test the model. The comparison of both of the models is given in Table 1

'learning_rate': [0.1, 0.01],
'n_estimators': [100, 200, 500],
'max_depth': [3, 5, 7]

Best parameters found for gradient Boost classifier are: n_estimators=100, learning_rate=0.1, max_depth=3

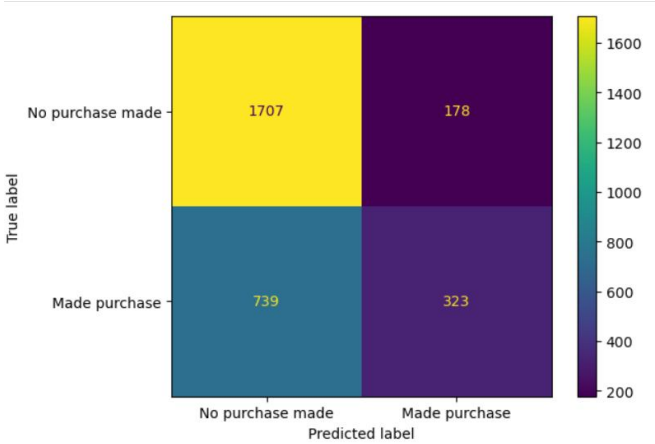


Fig 2: Confusion matrix for Gradient boost classifier

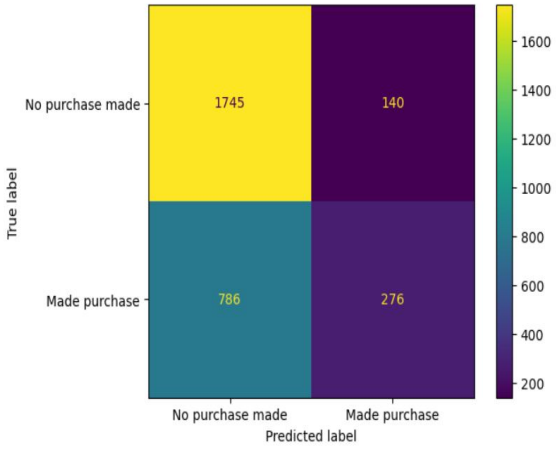


Fig 3: Confusion matrix for SVM

S.NO	METRIC	GBC	SVM
1	Accuracy	0.68883610451	0.685782151340
2	Precision	0.67129232948	0.676456174209
3	Recall	0.60485670897	0.592808224310
4	F1 score	0.41330774152	0.373477672530
5	R2 score	-0.3499373086	-0.36318642069
6	MSE	0.31116389548	0.314217848659
7	RMSE	0.55782066606	0.560551379143

Table 1: Metrics calculated form the models

V. CONCLUSIONS

In this work, we used a comprehensive set of data made up of customer traits to forecast consumer purchasing behavior using the gradient boosting technique and support vector machine. Although both our model's accuracy and F1 score were not as high as we were anticipating, it is crucial to highlight that there is still a lot of space for future enhancements and further development. Despite this, our study gave useful insight into the dataset and probable factors affecting purchasing decisions. A better grasp of the data properties was obtained as a result of the intensive feature engineering and exploration carried out during the study, which also showed the significance of more features such as the past transactional history to dive deeper and get better results. Fig 2 shows the confusion matrix for GBC and fig 3 shows SVM.

There are various directions for future research that can be taken in order to increase the model's capacity for prediction. The addition of supplemental characteristics, including data on consumer browsing habits, demographics, or past purchases of the customer, may significantly enhance model performance. Even though our results might not have been as accurate as we had hoped, the research done for the present investigation will serve as a starting point for future developments in consumer purchase behavior prediction. The knowledge collected from this study can help firms create more successful marketing plans, individualized deals, and client retention programs. In conclusion, this study has paved the way for future research and development. Future efforts in this area show excellent promises for improving the estimation of customer purchase behavior and aiding informed business choices with more data, the improvement of features, and the development of different modeling methodologies.

REFERENCES

[1] Asniar and K. Surendro, "Predictive Analytics for Predicting Customer Behavior," 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, 2019, pp. 230-233, doi: 10.1109/ICAIIT.2019.8834571.

[2] H. Valecha, A. Varma, I. Khare, A. Sachdeva and M. Goyal, "Prediction of Consumer Behaviour using Random Forest Algorithm," 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Gorakhpur, India, 2018, pp. 1-6, doi: 10.1109/UPCON.2018.8597070

[3] H. Valecha et al., research explored the relation between purchaser behaviour parameters and readiness to buy the item. This paper presents a survey to classify the purchaser and then k-nn, logistic egression and the Random Forest classifier in prediction of the willingness to buy the item. It is observed that RF fives beest accuracy of prediction i.e., 94% accuracy

[4] K. Maheswari and P. P. A. Priya, "Predicting customer behavior in online shopping using SVM classifier," 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Srivilliputtur, India, 2017, pp. 1-5, doi: 10.1109/ITCOSP.2017.8303085.

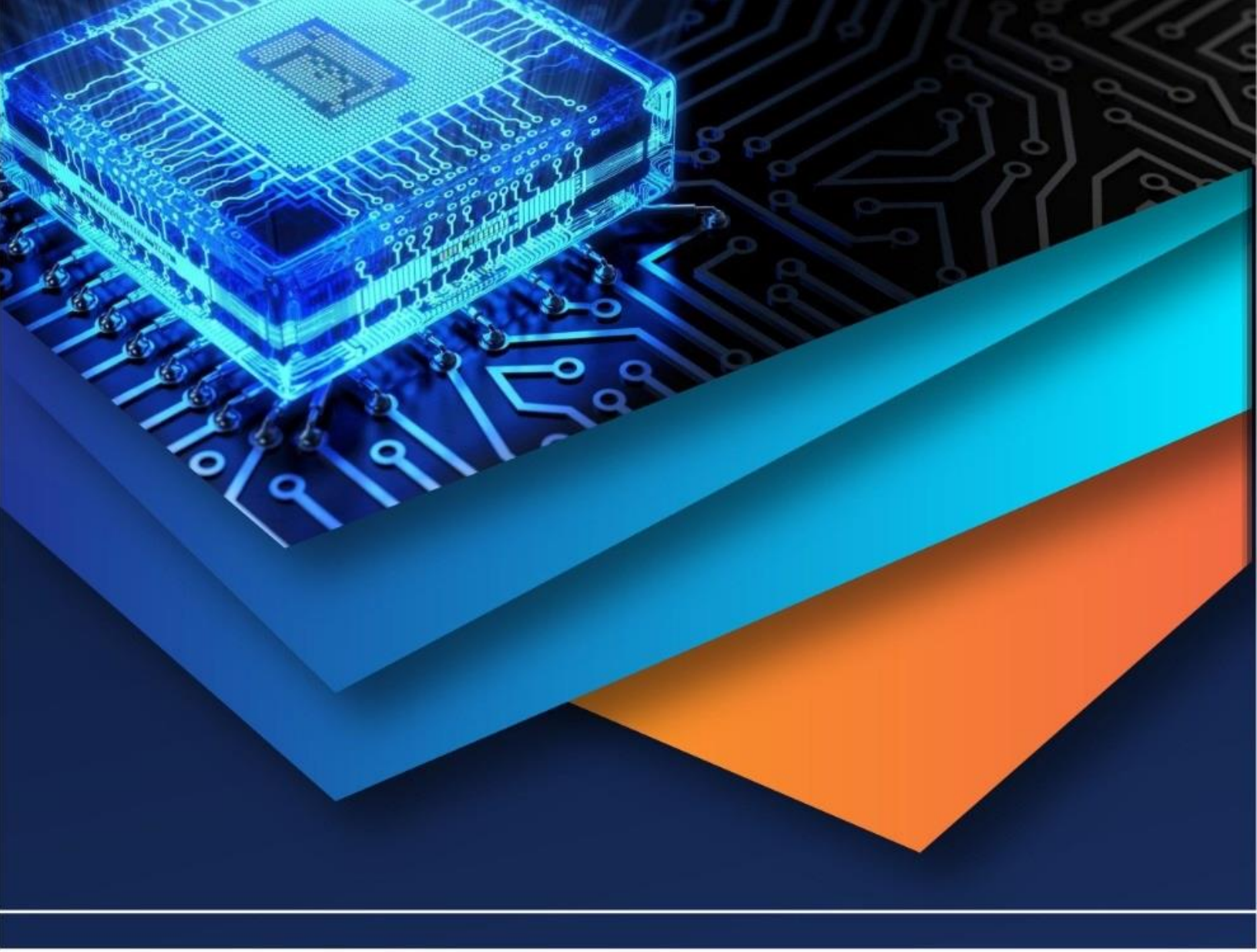
[5] X. Zhai, P. Shi, L. Xu, Y. Wang and X. Chen, "Prediction Model of User Purchase Behavior Based on Machine Learning," 2020 IEEE International Conference on Mechatronics and Automation (ICMA), Beijing, China, 2020, pp. 1483-1487, doi: 10.1109/ICMA49215.2020.9233677.

[6] Neha Chaudhuri, Gaurav Gupta, Vallurupalli Vamsi, Indranil Bose, On the platform but will they buy? Predicting customers' purchase behavior using deep learning, Decision Support Systems, Volume 149, 2021, 113622, ISSN 0167-9236

[7] Y. Zuo, K. Yada and A. B. M. S. Ali, "Prediction of Consumer Purchasing in a Grocery Store Using Machine Learning Techniques," 2016 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), Nadi, Fiji, 2016, pp. 18-25, doi: 10.1109/APWC-on-CSE.2016.015.

[8] Cirqueira, D., Hofer, M., Nedbal, D., Helfert, M., & Bezbradica, M. (2019, September). Customer purchase behavior prediction in e-commerce: A conceptual framework and research agenda. In International Workshop on New Frontiers in Mining Complex Patterns (pp. 119-136). Springer, Cham.

[9] S. Bag, M.K. Tiwari, F.T.S. Chan, Predicting the consumer's purchase intention of durable goods: An attribute-level analysis, J. Bus. Res. 94 (2019) 408-419. doi:https://doi.org/10.1016/j.jbusres.2017.11.031.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY