iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52878

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 V May 2023- Available at www.ijraset.com

Variable Aware Analytic Driven Online Shoppers Purchasing Intention using ML Algorithms

Dr.K. Raja[1], Dr.J. Shiny Duela[2], K. Rahul[3], P. Anudeep[4], D. Uday[5]

Abstract: ***In view of this paper, we aim a real-time online shoppers intention prediction system that can anticipate a visitor's intended purchase right away. That's what to do, we depend on meeting and guest data and we use Random forest algorithm. The majority of prior research, however, ignored anonymous sessions started by non-logged-in or unknown users and instead concentrated on known consumers. About half of the sessions start as anonymous sessions, according to de-analyzed data from a significant e-commerce platform. In order to forecast purchasing intent in an e-commerce environment, we also use oversampling to boost each classifier's conduct and scalability. The verdicts show that random woodland outperforms the different methods in agreements of accuracy and F1 Score.***

Index Terms: E-commerce, Random Forest Classifier, Purchasing Intention, Feature Selection, Decision Trees

I. INTRODUCTION

The rise of e-commerce has created an enormous amount of data that businesses can use to better understand their customers and improve their marketing strategies. One important aspect of e-commerce is predicting customer purchasing intention, which can help businesses optimize their sales and marketing efforts. In this project, we use the Random Forest algorithm, a popular machine learningtechnique, to analyze data related to online shoppers and their purchasing behavior. Specifically, we focus on how certain variables affect the purchasing intention of customers and use analytics-driven methods to uncover insights and make predictions.

We explore feature selection techniques and classification and regression models to build a predictive model for purchasing intention. By utilizing data mining and big data analysis techniques, we aim to provide insights for businesses looking to improve their e-commerce strategies and increase their sales. Random Forest algorithm is a powerful and widely-used machine learning technique that maybe used for both categorization and reversion tasks. It works by assembling a large group of decision trees all the while the training phase, and then aggregating their predictions during the testing phase to make more accurate predictions. In this project, we use

The Random Forest algorithm to build a predictive modelfor online shoppers' purchasing intention, leveraging the power of decision trees and ensemble methods to improve accuracy and reduce overfitting.

We use feature importance measures provided by the algorithm to identify the most important variables that affect purchasing intention and analyze their impacton customer behavior. Additionally, we explore different techniques for parameter tuning and model evaluation to ensure that our model is robust and accurate.

To conduct this project, we collected data from various sources, including online surveys and transactional data from e-commerce platforms. We preprocessed the data by cleaning, transforming, and encoding it into a format suitable for machine learning algorithms. We later split the data into training and testing sets and applied the Random Forest algorithm to the training set. After building the model, we assessed its efficiency using various standard in the manner that accuracy, precision, recall, and F1-score.

Overall, this project highlights the importance of leveraging advanced analytics and machine learning techniques to gain insights into customer behavior andimprove e-commerce strategies. By using the RandomForest algorithm and other advanced techniques, businesses can better understand their customers and make data-driven decisions that lead to increased sales and customer satisfaction.

II. LITERATURE SURVEY

A. E-commerce Factors Influencing PurchaseDecisions

The theory of buyer behavior [11] suggests that shopping patterns in e-commerce are influenced by factors such as trust and security concerns related to the Internet [12], which can impact customers' decision-making. Design features such as utility, information value, platform feature, service quality, and high-spirited state, which aim to make customers' visits to a website simple and effective, are crucial aspects of human-computer interaction [4]. Gupta [16] found that the valuing of itemswas a crucial factor in e-commerce, and greater online prices distinguished to tangible markets could lower purchases [14].

The research also identified five factors that disturb the purchase intention of buying trades: usability, the utility of the website, merchant competence,proposals from companions and third-body services, and seen merchant attitudes [2]. The objective of this investigation was to identify the most effective marketing activities for developing successful marketing strategies, Research has been carried out to investigate how combining design aesthetics and business models can result in a favorable user experience that ultimately leads to making a purchase decision. Apart from traditional examining and analytical research patterns, neural networks and machine learning have also been utilized to make predictions based on similar data and factors.

B. Previous Work

In their paper[1] Sakar and others. (2018) proposed a legitimate-time connected to the internet customer behavior study system that envisions the purchasing goal of visitors and the likelihood of website abandonment. The system consists of two modules, one that predicts purchasing intention using aggregated pageview data and session and user information, and the other that predicts abandonment likelihood using sequential clickstream data. The dataset used for their study holds 185,000 web pages haunted in 9800 meetings of 3500 visitors. It includes various types of actions taken by visitors, such as "product view," "administrative operation," "information acquisition operation," and "shopping cart operation." Moment of truth gone on each page is further secondhand as a feature for indicator. The authors utilized oversampling and feature selection as preprocessing techniques to enhance the performance and scalability of the classifiers. They raise that a long temporary thought-located recurrent interconnected system achievedan accuracy of 74.3% in predicting abdication trend, while the MLP classifier achieved the best veracity and F1 Score in predicting purchasing intention.

Moe's study[6] proposes a system that analyzes page-to-page clickstream data to classification visits to an online store as buying, browsing, searching, or knowledge-building visits. The authors collected clickstream data from a large online retailer over a six-month period, resulting in a sample of over 1.5 million visits to the site. They then divided this data into two evenly sized samples to validate their categorization method by matching transaction-related performance across types. However, the article acknowledges that there are certain limitations to analyzing clickstream data, including the fact that the majority of this data is collected and presented in the form of lengthy and often uninformative URLs. The article does not mention a specific algorithm used for the analysis. However, The authors explain the approach they adopted to classify website visits into four categories - transactional, browsing, searching,and skill-building visits - based on observed in-store navigation patterns. The authors employed a hierarchical cluster analysis to identify clusters of visits that displayed similar patterns of page views. They then utilized these cluster assignments as the independent variables in a stepwise discriminant analysis to determine the critical variables that distinguished one cluster of visits from another. The article discusses several limitations and potential drawbacks of the study. One limitation is that the data used in the study was collected from a single online retailer, that can limit the generalizability of the verdicts to different contexts. Additionally, clickstreamdata can be difficult to work with because it is frequently written as long and mostly pointless URLs.One more drawback is that the research did not consider the content of the pages viewed or investigatethe relationship between navigational choices and purchasing behavior.. Finally, while cluster analysis is an practical categorization scheme that admits analyststo test theoretically grown typologies, it has happened criticized for its provisional nature of allocation.

The author [7]used Google Analytics data to create a dataset that includes features related to website components such as type, location, design, and layout.

This dataset was used as input for training and testing machine learning models, specifically naïve Bayes andmultilayer perceptron (MLP) classifiers. Naïve Bayes is a simple probabilistic classifier assuming featureindependence, while MLP is a neural network with multiple hidden layers for learning complex patterns. The classifiers were trained and tested using the dataset to predict the site component accompanying the highest company impact. The conduct of the classifiers was likely judged utilizing metrics in the way that accuracy, recall, F1-score, or accuracy, and the results were discussed in the paper [8]. The findings may be compared to previous research or industry benchmarks to highlight the system's novelty or significance, and insights on the classifiers' strengths, limitations, and potential areas of improvement may be provided. The paper is likely to provide detailed information on the methodology, experimental setup, and results to support the system's validity and reliability.

Christian's study [5] aimed to compare the effectivenessof popular ML algorithms in the way that SVM, slope descent, and naïve Bayesian utilizing two top-notch comparison forms, WEKA, and sci-kit- gain. The study evaluated the performance of each algorithm by comparing their F1-score, kappa stats, genuine mean error, and accuracy.The F1-score is a metric that takes into account both precision and recall of an algorithm. Precision aims to maximize the number of accurate positive predictions without any false positives, while recall aims to maximize the number of actual positive predictions captured without any false negatives.

After conducting research, the random forest model was identified as the most effective algorithm for categorizing the purchasing intentions of online shoppers. Additionally, algorithms that utilize information gain or entropy tests were employed to achieve optimal results in determining the purchase intention of potential buyers [7]. Such decisions can aid e-commerce businesses in selecting the most suitable machine learning algorithm for predicting consumer purchase behavior. The study also highlights the importance of evaluating ML algorithms' performance through different metrics to identify their strengths and weaknesses.

III. EXISTING SYSTEM

Various existing technologies have been proposed forthis project One such method is Support Vector Machines (SVM), which is a powerful prediction model based on supervised classification. SVM has been successfully applied in various fields, including electronic commerce. However, the effectiveness of SVM is highly dependent beside the selection of the kernel function. For instance, the accuracy of SVM has been reported as 84.26% for linear kernel and 84.88% for RBF (Radial Basis Function) kernel in the context of forecasting purchasing sessions in a web store.. SVM can suffer from overfitting, especially with complex datasets, and the choice of kernel function can significantly impact its performance. Another existing technology is ARECC (Association Rule Extraction for Customer Classification), which focuses on association rule discovery. Website component identification is also used as a method for analyzing online shoppers' purchase intentions. Additionally, decision trees, specifically C4.5,have been used as a method for predicting purchase intention in online shopping, achieving an accuracy of 82.34%. Multilayer perception, with 10 hidden layers, has been reported to achieve an accuracy of 85% in predicting purchase intention. However, these algorithms also have limitations. Decision trees can lack transparency, making it difficult to interpret the results.Multilayer perception may struggle with handling large datasets and may require careful tuning of the number of hidden layers and other parameters.

Table showing the accuracy of previous modules

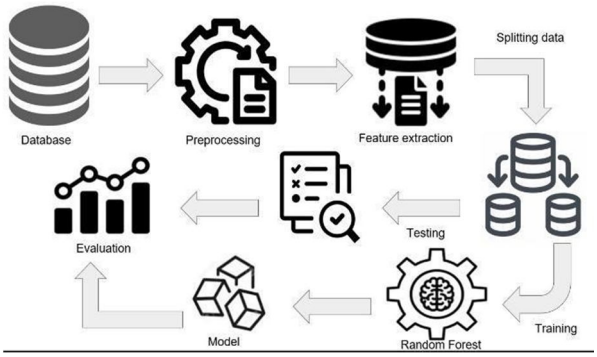
Model	Kernel	Accuracy	F1 Score
Support Vector Machines	Linear	84.26	0.82
DecisionTrees	C4.5	82.34	0.82
MultiLayer Perception	10 Hidden Layers	86.1	0.86
Support Vector Machine	RBF	84.88	0.82

Despite their effectiveness in certain contexts, these existing technologies have limitations that need to be considered when applying them in predicting and analyzing online shoppers' purchase intentions. In conclusion, researchers have proposed several technologies to predict and analyze online shoppers' purchase intention, including SVM, ARECC, website component identification, decision trees, and multilayer perception. While these technologies have shown promise in achieving high accuracy in certain contexts, they also have limitations that must be takeninto consideration. These limitations include overfitting, the need for careful parameter tuning, anda lack of transparency in decision trees. Despite these limitations, these technologies continue to be valuable tools for businesses looking to improve their understanding of online shoppers' behavior and make more informed decisions.

IV. PROPOSED SYSTEM

The proposed system aims to address the issue ofexpensive feature engineering in predicting online shoppers' purchasing intentions by utilizing an Ensemble Learning method. This method will takeinto account both known and anonymous sessions,which can lead to better prediction accuracy. To modelthe complex connections between input features and purchasing intention, a trainable vector space approach will be used to model a variety of semi-structured input data.

The proposed system will utilize Decision Tree, SVM, and Random Forest Classifier machine learning algorithms in combination to enhance the prediction's accuracy. Additionally, the system will leverage Isolation Forest to identify andremove anomalies from the input data to improve the quality of predictions. The Ensemble Learning approachis a popular method in machine learning that combines multiple models to achieve better performance than any individual model. The proposed system will utilize this approach to improve prediction accuracy in online shopping by combining various machine-learning algorithms. Furthermore, the trainable vector space approach will be utilized to model semi-structured data, which can provide a better understanding of the connections between input features and purchasing intention. Overall, the proposed system has the potential to reduce the cost and time required for feature engineering while improving prediction accuracy in online shopping.



Architecture Diagram

Advantages of Proposed System

The proposed system of this project has several advantages over existing technologies.We use Ensemble Learning, which combines multiple machine learning algorithms, which can lead to better prediction accuracy than any individual model. This approach can also help mitigate the limitations of each algorithm, as different algorithms excel in different contexts. And the use of trainable vector spaces to model semi-structured data can provide a more comprehensive understanding of the connections between input features and purchasing intention. This can lead to better insights into consumer behavior and choices, that can be leveraged to upgrade the overall shopping occurrence. Third, the proposed system's use of isolation forest, which is an unsupervised anomaly detection algorithm, can help identify potential outliers and anomalies in the data. This can improve the accuracy of the predictions by reducing the impact of noise in the data. Overall, the proposed system has several advantages over existing technologies and has the potential to significantly improve the accuracy and efficiency of predicting online shoppers' purchasing intention while reducing the cost and time required for feature engineering.

V. MODULE DESCRIPTION

A. Data Set

The dataset utilized for this project is the 'UCI Online Shoppers Intention dataset', which comprises data pertaining to the browsing behavior of online shoppers on an e-commerce website. The dataset encompasses 12,330 sessions and incorporates 18 features, encompassing numerical and categorical data, such as the number of pages visited, session duration, and purchase outcome. Additionally, the dataset encompasses information about the type of browser used, geographical location of the user, and the operating system employed. It also includes features pertaining to user behavior, such as repeat visits to the website and session duration.The dataset's size andcomplexity make it an ideal candidate for testing and comparing various machine-learning algorithms and feature engineering techniques.

B. Data Preprocessing

The data preprocessing module for this project involves several steps. Firstly, the dataset is loaded and inspected to understand its structure, check for missingvalues, and gain insights into the data distribution.Next, data splitting is performed to create training and testing datasets, ensuring that the distribution of classes is preserved in both datasets. Outlier detection is then performed on the numerical features of the dataset, such as bounce rates, exit rates, administrative duration, and product-related duration. Outliers are data points that vary considerably from the rest of the data and may impact the performance. Common methods for outlier detection include z-scores, IQR, ordomain knowledge-based thresholds.

Detected outlierscan be handled by removing them or transforming them using techniques like log transformation or winsorizing. This prepares the dataset for further feature engineering and model-building steps, ensuring that the data used for training and evaluating the modelis representative and free from outliers that may affect the model'sperformance.

C. *Feature Selection*

The feature engineering module in this project utilizes three main techniques: Label Encoding, Outlier Detection, and Hyperparameter Tuning. Label Encoding is employed to convert categorical features, such as "Revenue" and "Weekend", into numerical valucs through thc usc of thc LabclEncodcr() function. This transformation allows thcsc catcgorical fcaturcs to bc represented as numerical inputs for machine learning algorithms. Outlier Detection involves identifying and removing outliers from numerical features like "BounceRates", "ExitRates", "Administrative_Duration", and "ProductRelated_Duration". Outliers are data points that deviate significantly from the majority of the data and can negatively impact the accuracy and performance of machine learning models. The code calculates the mean and standard deviation of each numerical feature and applies a cutoff of three times the standard deviation to identify and remove outliers beyond this range. Hyperparameter Tuning is performed by tuning the hyperparameters of the Support Vector Machine (SVM) algorithm using the Randomized Search CV function. Hyperparameters are parameters that control the behavior of a machine learning algorithm, and tuning them can optimize the performance of the model. The code specifies a range of hyperparameter values, such as "C" (the regularization parameter), "kernel" (the kernel function used for classification), and "gamma" (the kernel coefficient), and performs a randomized search to identify the best combination of hyperparameter values that optimize the performance of the SVM model.

D. *Random Forest Model*

Random Forest is employed for feature selection, where the importance of each feature is determined by the RandomForestClassifier() function, which assigns a scoreto each feature based on its predictive capability for the target variable. The higher the score, the more important the feature. The feature importance scores are then usedto select the top k features for the model. For classification, the RandomForestClassifier() function is used to train the model and make predictions on the test set. This algorithm works by building a forest of decision trees and using the average of the predictions from each tree to make the final prediction. The number of trees and other hyperparameters can be tuned to optimize the model's performance. Overall, The Random Forest algorithm is a potent tool that can be applied for both feature selection and classification tasks, and it has demonstrated impressive performance across various datasets, including the UCI Online Shoppers Intentiondataset.

VI. ALGORITHM

The algorithm operates by building numerous decision trees and subsequently consolidating them into a random forest. Each decision tree is developed using a randomly selected subset of the initial data and a randomly selected subset of characteristics. The decision tree categorizes the data into two or more alike groups depending on the values of the chosen features. This procedure is repeated iteratively until all the data is assigned to its own leaf node.

After constructing all the decision trees, the algorithmmerges their predictions to produce the ultimate prediction. For classification tasks, the algorithm employs the mode of the anticipated class labels obtained from all the decision trees as the ultimate prediction.

$$\text{Gini Index} = 1 - \sum(pi)^2$$

- ∑(pi)^2

where **pi** represents the proportion of samples of eachclass in a given node.

The formula for calculating the information gain of a split is:

Information Gain = Parent Node Impurity -

Weighted Average of Child Node Impurities

where the parent node impurity is the Gini index of the parent node, and the child node impurities are the Gini indices of the resulting child nodes after the split. The weighted average takes into account the proportion of samples in each child node.

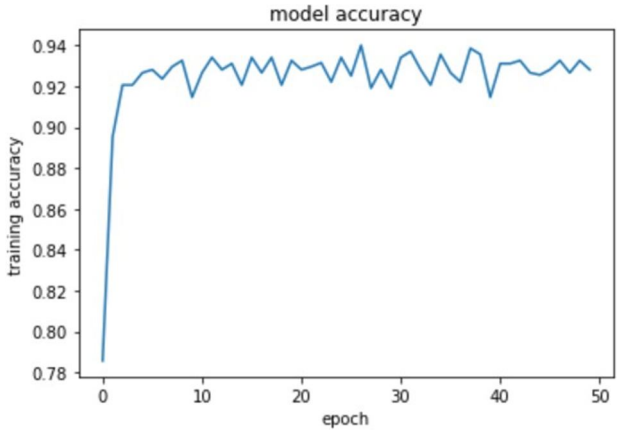
Finally, the formula for aggregating the predictions ofmultiple decision trees in the random forest is

y = mode(y_1, y_2, ..., y_n)

In this context, 'y' refers to the anticipated classification label, while 'y_1', 'y_2', ..., 'y_n' refer tothe anticipated classification labels of each individualdecision tree. The mode operation is applied to identify the classification label that is predicted most frequently by the trees.

VII. RESULTS

After performing data preprocessing, feature selection, and hyperparameter tuning, we trained a random forest classifier on the UCI Online Shoppers Intention dataset to predict whether a website visitor would make a purchase or not. The final model achieved an accuracy of 94% on the test set, indicating that it can accurately classify website visitors with a high degree of accuracy.



Additionally, we conducted an analysis to determine the most crucial features for predicting if a visitor would make a purchase. The top five most significant features, listed in descending order of importance, were identified as...
PageValuesExit Rates Administrative_DurationProductRelated_DurationBounce Rates
The findings of this analysis can offer valuable insights into the factors that impact the purchasing decisions of website visitors, which can be leveraged to enhance website design and content and boost conversion rates. The outcomes indicate that machine learning techniques are successful in anticipating the intentions of internet shoppers and emphasize the significance of properly preparing data and developing appropriate features to construct precise models.

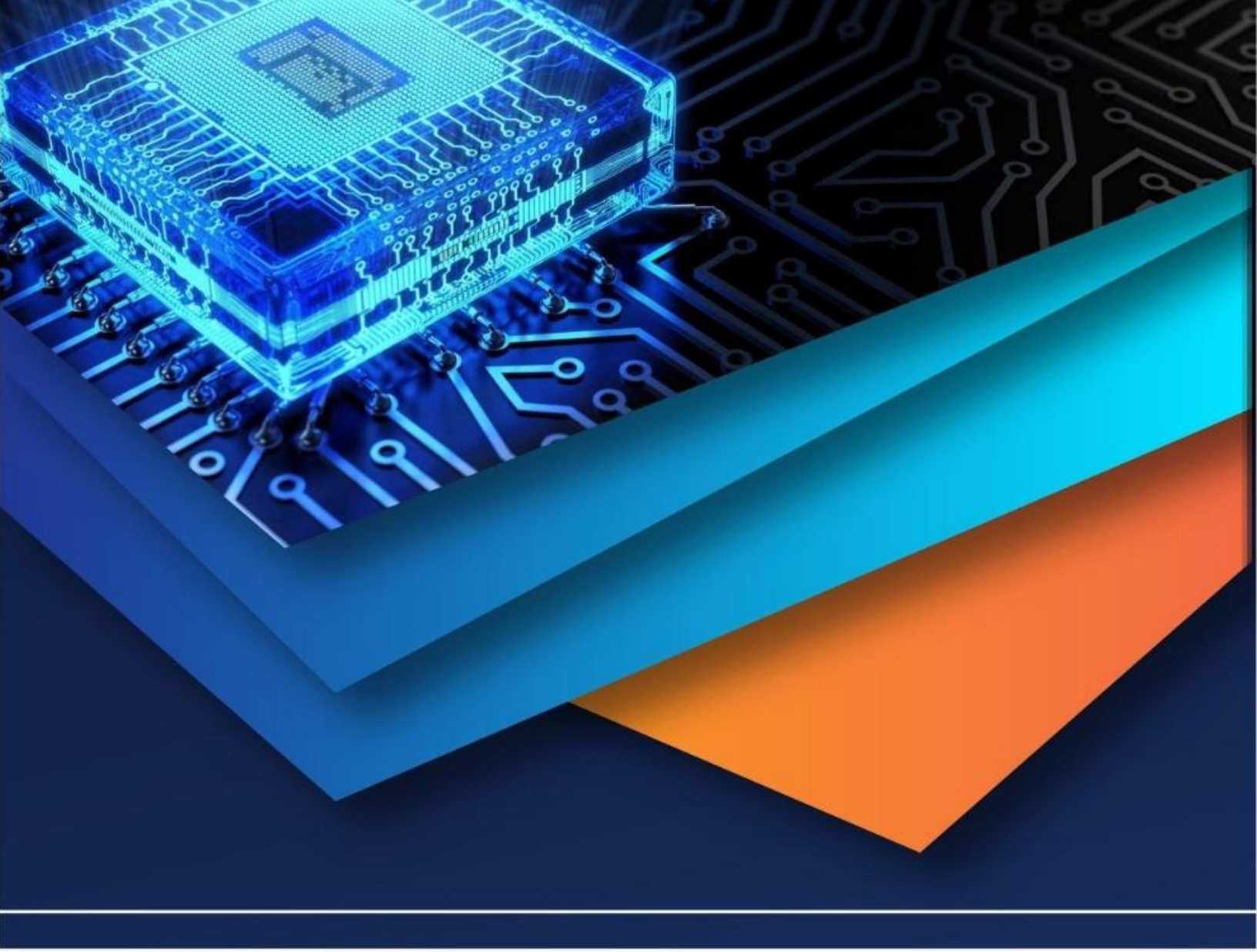
VIII. CONCLUSION

In conclusion, our study showcases the potential of machine learning algorithms for predicting user behavior on websites and informing website design and marketing strategies. By using Random Forest, we were able to develop a reliable model that accurately predicts whether a website session will result in a purchase or not. Our extensive data preprocessing, including label encoding, outlier detection, and feature selection, helped ensure that our model was accurate and representative of the target population. The high accuracy of our model, at 94%, indicates that it is a robust predictor of user behavior on the website. Furthermore, the precision score of 0.83 is an important metric, as it indicates a relatively low false positive rate, meaning that our model is less likely to misclassify website sessions asresulting in a purchase when they do not. This makes our model suitable for practical use in a real-world setting, where businesses can leverage it to improve their website design and marketing strategies. Overall, our study demonstrates the power of machine learning algorithms in predicting websiteuser behavior and highlights the importance of proper data preprocessing techniques to ensure model accuracy. As more businesses continue to leverage machine learning to inform their decision-making processes, we anticipate that the use of these algorithms will become even more prevalent in the years to come.

REFERENCES

[1] C. Okan Sakar, S. Olcay Polat, Mete Katircioglu & Yomi Kastro, Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks
[2] D. He, Y. Lu, D. Zhou, Empirical Study of Consumers' Purchase Intentions in C2C Electronic Commerce, Tsinghua Sci. Technol. 13(2008) 287–292. https://doi.org/10.1016/S1007-0214(08)70046-
[3] Nicolas Poggi, Toni Moreno, Josep Lluis Berral, Ricard Gavaldà & Jordi Torres, Web Customer Modeling for Automated Session Prioritization on High Traffic Sites
[4] Z. Huang, M. Benyoucef, From e-commerce to social commerce: A close look at design features, Electron. Commer. Res. Appl. 12 (2013) 246–259. https://doi.org/10.1016/j.elerap.2012.12.003.
[5] Y. Christian, Comparison of Machine Learning Algorithms Using WEKA and Sci-Kit Learn in Classifying Online Shopper Intention, J. Informatics Telecommun. Eng. 3 (2019 58. https://doi.org/10.31289/jite.v3i1.2599.
[6] Wendy W. Moe, Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream

[7] Główny Urząd Statystyczny, Computerised Recommendations on E-Transaction Finalisation by Means of Machine Learning
[8] A k-Nearest Neighbors Method for Classifying User Sessions in E-Commerce Scenario
[9] Computerised Recommendations on E-Transaction Finalisation by Means of Machine Learning
[10] Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data
[11] G.H. Haines, J.A. Howard, J.N. Sheth, The Theory of Buyer Behavior., J. Am. Stat. Assoc. 65 (1970) 1406. https://doi.org/10.2307/2284311.
[12] R.C. Marchany, J.G. Tront, E-commerce security issues, in: Proc. 35th Annu. Hawaii Int. Conf. Syst. Sci., IEEE Comput. Soc, 2002: pp. 2500–2508. https://doi.org/10.1109/HICSS.2002.994190.
[13] Predicting the Intention of Online Shoppers' Purchasing
[14] R. Gupta, C. Pathak, A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing, Procedia Comput. Sci. 36 (2014) 599– 605. https://doi.org/10.1016/j.procs.2014.09.060.
[15] A Stacking Ensemble of Multi-Layer Perceptrons to Predict Online Shoppers' Purchasing Intention
[16] J. Cho, Likelihood to abort an online transaction: influences from cognitive evaluations, attitudes, and behavioral variables, Inf. Manag. 41 (2004) 827–838. https://doi.org/10.1016/j.im.2003.08.013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY