European Journal of Technology (EJT)

Predictive Customer Engagement Using Machine Learning Models for Retail Business

Rami Reddy Kothamaram, Dinesh Rajendran, Venkata Deepak Namburi, Vetrivelan Tamilmani, Aniruddha Arjun Singh Singh, Vaibhav Maniar

European Journal of Technology
ISSN 2520-0712 (online)
Vol.7, Issue 4, pp 56 - 73, 2023

www.ajpojournals.org

Predictive Customer Engagement Using Machine Learning Models for Retail Business

Rami Reddy Kothamaram[1], Dinesh Rajendran[2], Venkata Deepak Namburi[3], Vetrivelan Tamilmani[4], Aniruddha Arjun Singh Singh[5], Vaibhav Maniar[6]

[1]California University of management and science, MS in Computer Information systems, [2]Coimbatore Institute of Technology, MSC. Software Engineering, [3]University of Central Missouri, Department of Computer Science, [4]Principal Service Architect, SAP America, [5]ADP, Sr. Implementation Project Manager, [6]Oklahoma City University, MBA / Product Management

Article history

Submitted 19.10.2023 Revised Version Received 16.11.2023 Accepted 12.12.2023

Abstract

Purpose: Customer interaction in retail business is one of the key sources of loyalty, profitability, and long-term development since it shows the ability of a business to establish relations with its customers. Using the Online Retail data set on Kaggle which contains over half of the transactions, totaling 540,000 this article investigates how machine learning approaches might be utilized to predict consumer participation.

Materials and Methods: The research methodology includes data preprocessing in the form of data cleaning, data normalization, and outlier identification. K-Nearest Neighbors (KNN) and Random Forest (RF) are two algorithms that were tested and evaluated using key performance indicators such as recall, accuracy, precision, F1-score, and ROC-AUC.

Findings: Based on the results of the trial, KNN was the most accurate with a 97.90% score, while RF was the most efficient with a 96.46% score, thanks to its higher recall and F1-score. Comparative analysis with other models, including CNN and XGBoost, confirmed the robustness of the proposed models in outperforming traditional approaches. Retail analytics that incorporate machine learning have the ability to boost decision-making, inventory management, and tailored marketing, according to the results. This study demonstrates that predictive modeling can provide powerful tools for fostering customer engagement and achieving competitive advantage in the retail sector.

Unique Contribution to Theory, Practice and Policy: Future improvements, incorporating advanced models such as Gradient Boosting or deep learning, as well as integrating real-time data streams and customer sentiment from social media or reviews, could provide deeper insights into engagement patterns.

Keywords: *Customer Engagement, Retail Business, Machine Learning, Online Retail Dataset, Retail Analytics*

INTRODUCTION

A retail economy is one that allows people to buy and sell products and services for a small price; it connects the needs of farmers and producers with those of consumers; it meets people's fundamental needs; and it promotes the growth of creative solutions to those needs, which ultimately leads to happiness, success, and community cohesion. A French-Italian term is where the word "Retail" was first used, according to one meaning. One who removes a tiny fragment from an object is known as a retailer. Commercial sales of products and services to consumers for in-home or other non-institutionalized use constitute retail sales. The retail market in India is the fifth biggest in the world [1]. Retailers may be either individuals or companies. Individuals or organizations whose primary occupation is retailing are considered brokers who engage in retail activities within the marketing channel. Various entities engage in retailing, including producers, wholesalers, and retailers. Nonetheless, retailers companies whose primary source of revenue is retail sales perform the vast majority of retailing.

One strategy for advertising websites like Yahoo!, Rediff, etc. is to track user engagement, which can be defined as "the degree to which a user actively participates in the experience of using a given website" [2]. An indicator of this would be the amount of time a customer spends actively engaging with the website. The term "engagement" can mean many things depending on the context. It has been defined by some academics as customer input into the creation of organizational policies and services. For policies and services to be developed, it is necessary to involve consumers in the planning and development process. This is known as consumer engagement. [3]. A rising number of successful applications of AI technologies may be seen in the retail industry, among others. Building a conceptual framework for incorporating AI methods and algorithms into retail organizations' information systems was a good use of the acquired cognitive capacity. Acquiring a new customer can be somewhat costly for the organization. The current explosion in data collection has piqued a lot of people's interest in machine learning (ML) [4]. After studying the data, the computer demonstrated its value in making predictions and discovering patterns. Extensive research into machine learning (ML) skills led to notable achievements in the retail sector [5]. The retail business has just begun to utilize some of the most famous machine learning techniques, like DL [6][7]. DL approaches have yielded more accurate forecasts and outcomes. One more way to improve system performance is via DL.

Motivation and Contribution

The rapid growth of the retail industry, fueled by increasing consumer demands and vast amounts of transactional data, has created a pressing need for intelligent systems that can accurately predict and enhance customer engagement. The traditional retailing practices are usually not able to capture the behavior of the customers properly and thus there is difficulty in customization, optimality of the inventory and customer retention. The high cost of obtaining new customers makes it more and more attractive to businesses to concentrate on retaining and engaging existing customers. The research is inspired by the possibility of machine learning (ML) methods to change retail analytics, revealing the concealed patterns in consumer data and making it possible to make a decision based on data. Using models like KNN and RF on the Online Retail dataset, the proposed research helps resolve the issue of customer engagement prediction accuracy, which guides retailers to devise effective marketing strategies, enhance efficiency in their operations, and build long-term customer loyalty. In terms of research on consumer involvement in retail settings, this study makes a number of essential contributions:

- Development of a systematic framework for applying machine learning to the Online Retail dataset, including data cleaning, normalization, and outlier detection.

- Implementation and comparison of two classification models KNN and RF to analyze customer purchasing behavior.
- Recall, accuracy, precision, F1-score, and ROC-AUC are just a few of the metrics that will be used to evaluate the model's robustness.
- Determination of the most efficient model of retail analytics, which provides practical insights on how to enhance decisions in the field of customer targeting and inventory management.

Novelty and Justification of the Study

This research is novel due to the fact that the Online Retail dataset is put under the methods of machine learning, in this case, KNN and Random Forest, to gain better understanding of customer behaviour and buying trends. Unlike traditional statistical approaches, the use of advanced pre-processing steps such as outlier detection and normalization ensures higher data reliability, while rigorous evaluation using multiple performance metrics provides a holistic view of model effectiveness. This study is justified by the growing need for businesses to leverage data-driven decision-making in competitive retail environments, where accurate predictions of customer activity can significantly enhance personalization, inventory management, and overall profitability. By systematically comparing two distinct algorithms, the work not only identifies the most effective model but also contributes to the broader understanding of applying machine learning in real-world retail analytics.

Structure of the Paper

This paper is structured as follows: Section II provides a survey of the literature on retail consumer interaction, focusing on studies that have used machine learning, datasets, and important results. The research approach, including data collecting, pre-processing, and dataset splitting, is described in Section III. Section IV details the findings and analysis, including a comparison to other models. Section V closes the study and delineates future work directions.

LITERATURE REVIEW

An extensive literature search on consumer participation in retail settings formed the basis of this study, which in turn determined its focus and methodology.

Naz and Popowich (2019) looked at the potential applications of supervised learning models in retail telecoms in this study. In particular, they looked at the potential applications of SVM, Bayesian classifiers, feed-forward ANN, and closest neighbour approaches to retail telecommunications data. Their preliminary results show that they can successfully classify retail telecommunication data according to profitability, with a precision of 95.5%, recall of 94.7%, and f-measure of 95.1% [8].

Arif et al. (2019) developed an innovative approach that contributes to precise forecasting by means of machine learning. Gathering and analyzing a store's historical data is what this strategy does best. Get ready to use this strategy by collecting the necessary data, processing it, and organizing it. Using associated algorithms on process data. Knowledge of the algorithms used for prediction includes KNN, SVM, RF, DT Classifier, and Gaussian Naive Bayes. Market data is gathered in a real-life setting. They may compare the results and accuracy of several algorithms after creating a data set and applying them. When they put them side by side, they discover that Gaussian Naive Bayes provides the most accurate results [9].

Bhatnagar and Srivastava (2019) Python is used to create two supervised machine learning classification models. Since the data utilized is labeled, this can be classified as supervised machine learning. The confusion matrix shows that when it comes to predicting client attrition,

KNN is better than Logistic Regression. When it comes to predicting client attrition, KNN outperforms Logistic Regression by 2.0%. The "accuracy" parameter here shows that the Logistic algorithm's prediction of customer turnover is worse than KNN [10].

Liu et al. (2018) article proposes a smart, unstaffed retail shop model based on artificial intelligence and the internet of things (IoT) in an effort to study the feasibility of this shopping style. To count and recognize SKUs, an end-to-end classification model is trained using the MASK-RCNN approach. On the test dataset, the suggested solution obtains a counting accuracy of 97.7 percent and a recognition accuracy of 96.7 percent, proving that the system can compensate for the shortcomings of conventional unmanned container systems. The dataset includes ten distinct kinds of stock-keeping units (SKUs) and is based on eleven thousand photographs taken in various settings [11].

Kim and Lee (2018) Bayesian Optimization to fine-tune the model that predicts customer attrition. Customer Relationship Management relies heavily on the customer churn forecasting model. On the other hand, achieving great precision requires proper hyperparameter settings. Here, they use a Recurrent Neural Network to fine-tune seven hyperparameters of a model that predicts consumer attrition. Results from the experiment demonstrate a substantial improvement in the predictive model's accuracy. They also show how changing each hyperparameter affects the predictive model's precision [12].

Amnur (2017) Customer Relationship Management (CRM) helps companies learn more about their customers and communicate with them in two ways. Without assuming a relationship between project risk, project management, and organizational performance, CRM systems analyze decision-making based on multiple criteria using the stimulus-organism response paradigm. This study's classification tasks are best handled by machine learning using the Support Vector Machine approach, which is able to simulate nonlinearities in CRM systems. Machine learning and customer relationship management help Bank X maximize earnings by identifying and nurturing new customers, retaining existing ones, and managing their most valued clients [13].

Kaneko and Yada (2016) built A model for sales prediction using point-of-sale (POS) data from a retail store that spanned three years in the current research. Using the sales data from one day, this model can predict how the numbers will change the next day. Consequently, a deep learning model that considers the L1 regularization achieved an 86% success rate in sales forecasting. All of the merchandise is organized in its own section at the store. As the number of attributes for product categories rose from tens to thousands, the expected accuracy remained relatively unchanged, falling only slightly below 7%. On the other side, accuracy was down around 13% when using the logistic regression model [14].

The following Table I presents a consolidated overview of recent studies on customer engagement in retail business, summarizing the proposed models, datasets utilized, key findings, as well as the challenges and future research directions identified in each study.

Table 1: Recent Studies on Customer Engagement in Retail Business

Author	Proposed Work	Dataset	Key Findings	Limitations & Future Work
Naz & Popowich (2019)	Explored supervised learning models (KNN, Feedforward ANN, Bayesian Classifier, SVM) for retail telecom data classification	Retail telecommunications data	Achieved precision 95.5%, recall 94.7%, F1-score 95.1%, demonstrating effective categorization based on profitability	Limited to initial dataset; future work could include larger datasets and additional models for better generalization
Arif et al. (2019)	Proposed ML-based method for store sales prediction using KNN, SVM, Gaussian Naive Bayes, RF, DT, and regression	Real-life market/store data	Gaussian Naive Bayes achieved highest accuracy among tested algorithms	Limited dataset; future work could explore ensemble models and real-time prediction systems
Bhatnagar & Srivastava (2019)	Implemented KNN and Logistic Regression for customer churn prediction	Labeled customer churn data	KNN outperformed Logistic Regression by 2% in accuracy	Focused only on two algorithms; future work can explore hybrid models and feature engineering to improve predictions.
Liu et al. (2018)	Intelligent, unattended storefront powered by AI and the Internet of Things using MASK-RCNN for stock-keeping unit identification and counting.	11,000 images of 10 SKU types	Counting accuracy 97.7%, recognition accuracy 98.7%	Dataset limited to 10 SKUs; future work could expand SKU types and explore other deep learning architectures.
Kim & Lee (2018)	Optimized customer churn prediction using Recurrent Neural Network with Bayesian Optimization of hyperparameters	Customer churn dataset	Accuracy significantly improved by optimizing 7 hyperparameters.	Limited to RNN and hyperparameter optimization; future work could explore other neural architectures and larger datasets.
Amnur (2017)	Implemented CRM solutions that utilize Machine Learning (SVM) for classification	-	SVM effectively models nonlinearities in CRM tasks, enabling Bank X	Dataset details and scalability across different domains not discussed. Future

	and multi-criteria decision-making analysis tools.		to optimize profit by managing high-value customers, attracting new customers, and recovering lost potential customers.	work could involve testing with larger, diverse datasets, comparison with other ML algorithms, and integration with real-time CRM platforms.
Kaneko & Yada (2016)	Developed DL model for daily sales prediction using POS data with L1 regularization	3 years POS data from retail store	Achieved 86% accuracy; predictive accuracy decreased by only ~7% with more product attributes	Limited to single store; future work could test generalization across multiple stores and explore alternative regularization techniques

MATERIALS AND METHODS

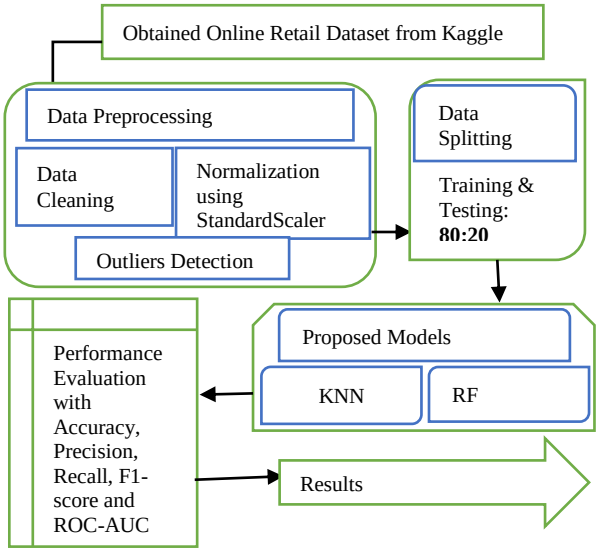

Figure 1: Proposed Flowchart for Customer Engagement in Retail Business

The proposed method starts with obtaining the Online Retail dataset from Kaggle. After that, it goes through a comprehensive data pre-processing step that includes data cleaning, standardization with the StandardScaler, and outlier identification to make sure the data is consistent and of high quality. After pre-processing, the dataset is divided into training and testing sets with a ratio of 80:20. This allows for efficient construction and assessment of prediction models. Two proposed models, KNN and RF, were trained on the dataset that was prepared. Next, the models are assessed using standard classification metrics as F1-score, recall, accuracy, precision, and ROC-AUC to guarantee a comprehensive assessment of their prediction capabilities. Finding the optimal model for the dataset is a matter of comparing the results. The whole procedure of the suggested approach is shown in Figure 1.

The following section elaborates on each step outlined in the proposed Customer Engagement in Retail Business flowchart

Data Gathering and Analysis

This study utilized the Online Retail dataset from Kaggle, containing 541,909 transactions with 8 features: Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, CustomerID, and Country. The dataset provides detailed information on customer purchases and sales patterns, including numerical and categorical attributes, with some missing CustomerID values requiring pre-processing. The EDA performed a high-level review of the information in search of overarching patterns or correlations that might serve as a framework for the rest of the research. Some data visualizations of the data are given below:

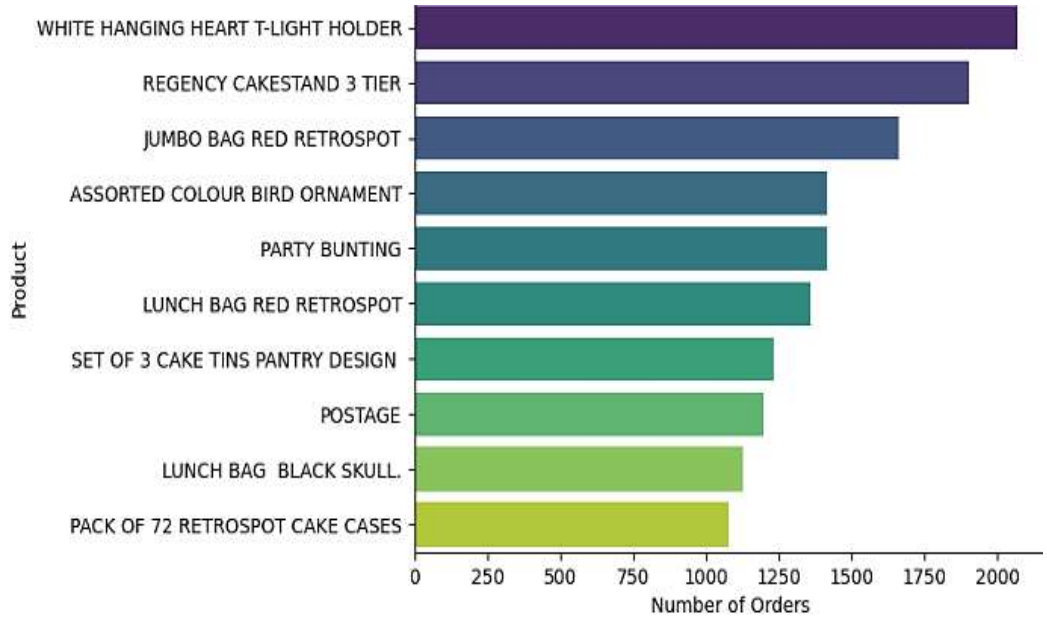

Figure 2: Top 10 Selling Products

Figure 2 is a horizontal bar chart that shows the 10 most popular products. The y-axis displays the product names, while the x-axis shows the total number of orders in descending order. The "WHITE HANGING HEART T-LIGHT HOLDER" has just under 2,000 orders, putting it first, while the "PACK OF 72 RETROSPOT CAKE CASES" ranks tenth with just over 1,000. A gradient color palette highlights popularity, with darker purple-blue for higher orders and lighter green-yellow for lower. The chart effectively visualizes and compares product demand.

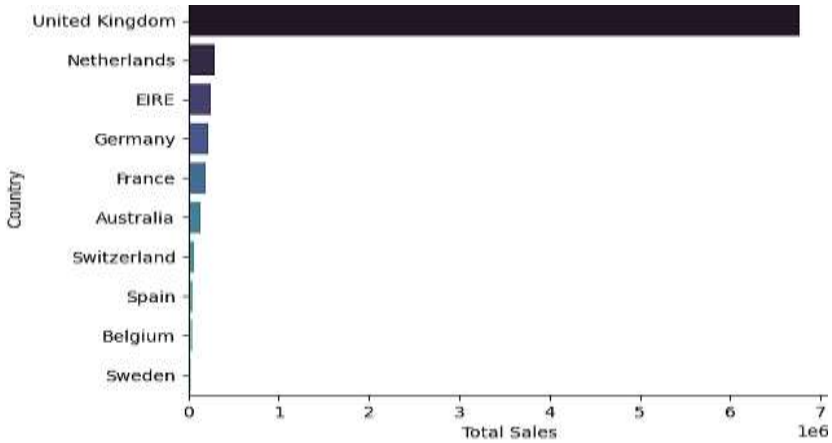

Figure 3: Top 10 Countries by Sales

In Figure 3, they can see the top-selling countries in terms of overall sales with a horizontal bar chart. Sales in millions, arranged descendingly, are shown on the x-axis, while the y-axis lists countries. The United Kingdom leads by a wide margin, exceeding 6 million in sales, followed by the Netherlands, EIRE, Germany, and others. A color gradient from dark purple to light teal highlights the disparity in sales among countries, effectively visualizing their relative performance.

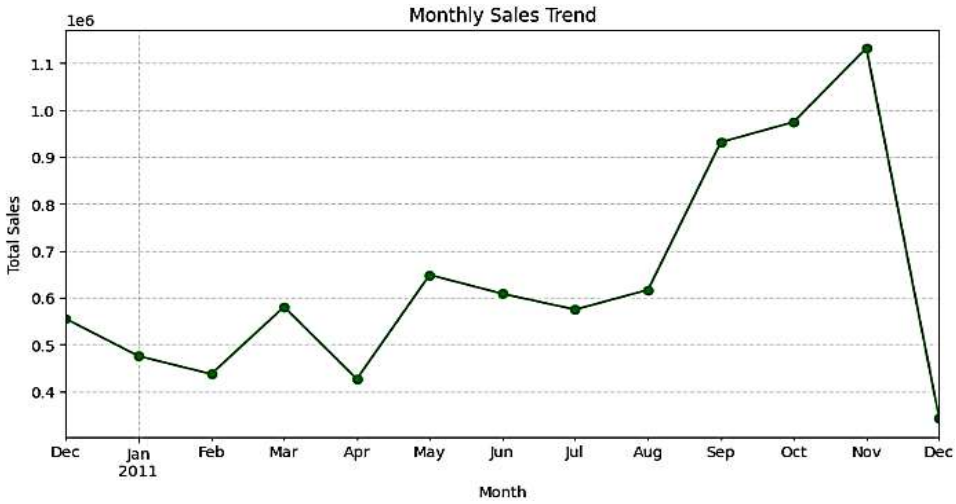

Figure 4: Monthly Sales Trend

Figure 4 is a line graph showing the monthly sales pattern from 2010 to 2011. The y-axis shows total sales, while the x-axis shows the number of months. Sales start low, dip in February, gradually rise from April, and peak in November at over 1.1 million. A sharp drop follows in December, reaching the lowest point of the year, highlighting a cyclical sales pattern.

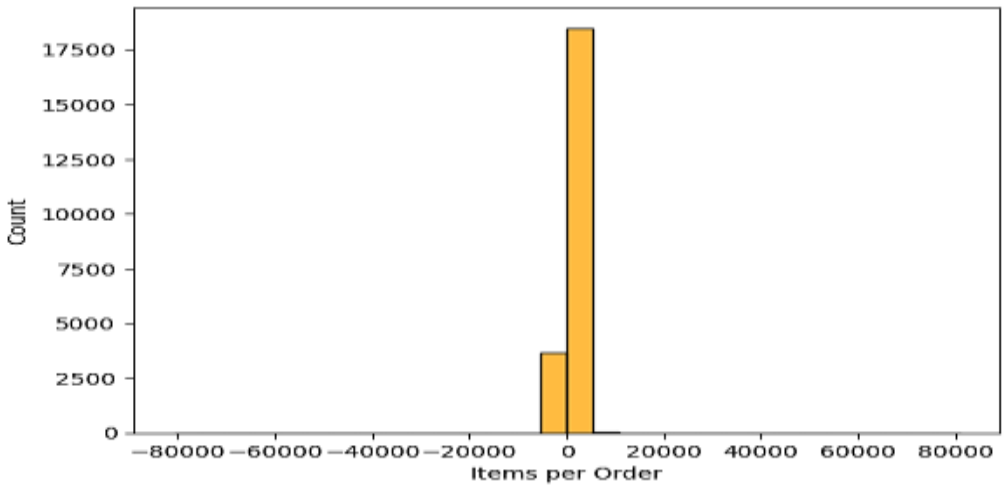

Figure 5: Distribution of Items Per Order

The distribution of objects per order is shown in Figure 5, a histogram. On one side, they have the quantity of things, and on the other, they have the frequency of orders. Most orders contain a small, positive number of items, while a few orders have negative values, likely representing returns or cancellations. Extreme values are rare, indicating that most orders are small, with some outliers requiring further investigation.

Data Pre-processing

A number of preparation processes were performed on the dataset to improve its quality before analysis. Data preprocessing incorporates stages of cleaning, normalization, and outlier detection.

- **Data Cleaning:** During data cleaning, duplicate and irrelevant items were removed. Then, mode imputation was used for categorical variables and mean imputation for continuous variables to manage missing values.
- **Data Normalization using StandardScaler:** Continuous variables were normalized using the following Equation (1) to ensure a uniform scale:

$$z = \frac{x-\mu}{\sigma} \quad \text{...} \quad (1)$$

In the given data set, z and x are scaled variables, whereas μ and σTand for the average and standard deviation of the training sets, respectively.

- **Outliers Detection:** A small number of users who clicked or purchased an excessive number of things on a single day can cause overall internet activity to be abnormally high or low. They classify these consumers as anomalies since their behaviours do not line up with what the model suggests; it's possible that these interactions are the product of bots or large purchasers.

Data Splitting

An 80:20 train-test split of the dataset was used to train and evaluate the model, guaranteeing a valid and robust performance assessment.

Proposed Models

The suggested plan to use k-NN and RF effectively and efficiently for customer engagement in retail business is shown in this section. These models are detailed below:

K-Nearest Neighbors

Using clusters is a breeze for achieving an approximation KNN that is efficient. The goal of a c-clustering procedure is to split data into $c$distinct groups while minimizing a set of criteria that vary with each clustering approach. Here, cIs the number of clusters used. After that, for any given sample x, The k-nearest neighbor search begins with getting the closest cluster, denoted by its centroid. After that, run the standard kNN search, but this time they only look for elements in that cluster [15]. Equation (2) shows the size of the training set, m. With this information, one can find the expected or theoretical number of distances for input classification:

$$f(m, c) = c + \frac{m}{c} \quad \text{..} \quad (2)$$

The first step is to find the cluster that is closest, which requires c lengths. In the next phase, they search within the retrieved cluster using an average of $\frac{m}{c}$ distances, since all the samples must belong to a cluster. To be sure, the distribution of samples within each cluster and the cluster that was obtained in the first phase determine the actual number of distances for a specific test sample.

Random Forest (RF)

An RF regression model is an approach that uses averaging to improve prediction accuracy and reduce over-fitting. It fits many grouping decision trees on different subsets of the data. By default, bootstrap=True limits the sub-set size to what the max_samples border allows. If this is

not the case, then each tree is built using the entire dataset The RF ensemble approach may perform classification and regression using a number of decision trees, the Bootstrap technique, and aggregation, often known as bagging. The key idea is to combine multiple decision trees into one, rather than relying on just one to get the task done. As its foundational learning paradigm, RF makes use of several DT.

Evaluation Metrics

The evaluation takes five variables into account, which capture both the quantitative performance and the practical usability of the product. Accuracy, precision, recall, F1-score, and area under the curve (AUC) are some of the evaluation metrics used to estimate its performance, as specified in equations (3) through (6):

- **True Positives (TP):** Customers accurately anticipated becoming involved.
- **True Negatives (TN):** Accurately predicting that customers would not be engaged.
- **False Positives (FP):** Accurately estimating customers as engaged.
- **False Negatives (FN):** Clients were erroneously assumed to be disengaged.

1) Accuracy

The percentage of consumers whose orders were accurately filled out. A high value indicates that the model's predictions are generally accurate. It is presented as Equation (3):

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+Fp+TN+FN}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

2) Precision

The percentage of expecting consumers who are actually participating. In retail terms, it measures how many of the customers the model identifies as engaged truly show purchasing or interaction behavior. The capability of the classifier to accurately detect positive classes is quantified in Equation (4):

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

3) Recall

The percentage of actively involved consumers that the model accurately detects. In retail, it shows how effectively the model captures all genuinely engaged customers. In mathematical form it is given as Equation (5):

$$Recall = \frac{\text{TP}}{TP+FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

4) F1 score

Accuracy and memory as a harmonic mean. Correctly recognizing engaged clients (recall) while avoiding false alarms (accuracy) is balanced by it. It can be expressed mathematically as Equation (6):

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

5) ROC-AUC

Criteria for the model's capacity to distinguish between engaged and unengaged customers across different levels. The discriminative power of a model to predict consumer involvement is inversely proportional to its AUC.

FINDINGS

An NVIDIA GeForce RTX 3070 Ti Laptop GPU and 8 GB of RAM were used to test this model. In Table II, they can see the experimental results of the proposed models for customer interaction in the retail business, which compare K-Nearest Neighbors (KNN) with Random Forest (RF). The table demonstrates that KNN attained a 97.90% accuracy, a 93.3% precision, a 70.0% recall, and an 80.0% F1-score. While KNN offers better precision, the RF model showed improved recall (98.8%) and F1-score (98.7%), suggesting that RF is better at accurately identifying engaged clients. The RF model's slightly lower accuracy (96.46%) was offset by these improvements. In general, the two models have high performance, but RF demonstrates better balance, recall and general predictive capability.

Table 2: Experimental Results of the Proposed Models for Customer Engagement in Retail Business

Performance matrix	KNN	RF
Accuracy	97.90	96.46
Precision	93.3	93.3
Recall	70.0	98.8
F1-score	80.0	98.7

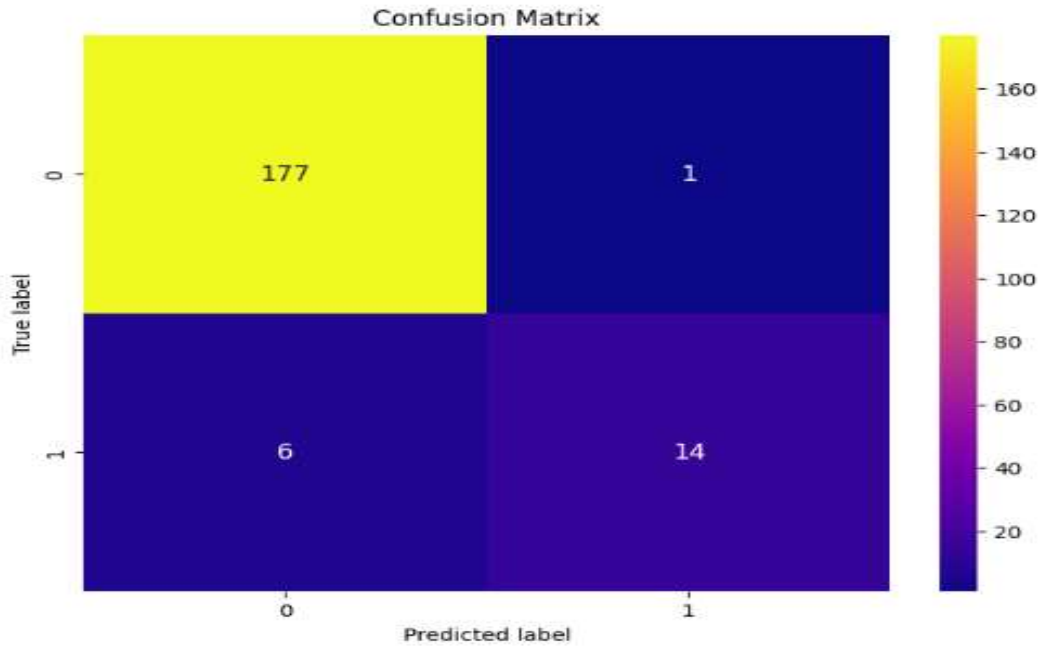

Figure 6: Confusion Matrix for the RF Model

The Random forest model's misclassifications are shown via a confusion matrix in Figure 6. There are 177 accurately recognized negative cases, called True Negatives (TN), in the top left cell and 1 false positive, or FP, in the top right cell. There are 6 FNs, or positive cases that were mistakenly identified as negative, in the bottom-left cell, and 14 TPs, or positive cases that were accurately identified, in the bottom-right cell.

European Journal of Technology
ISSN 2520-0712 (online)
Vol.7, Issue 4, pp 56 - 73, 2023

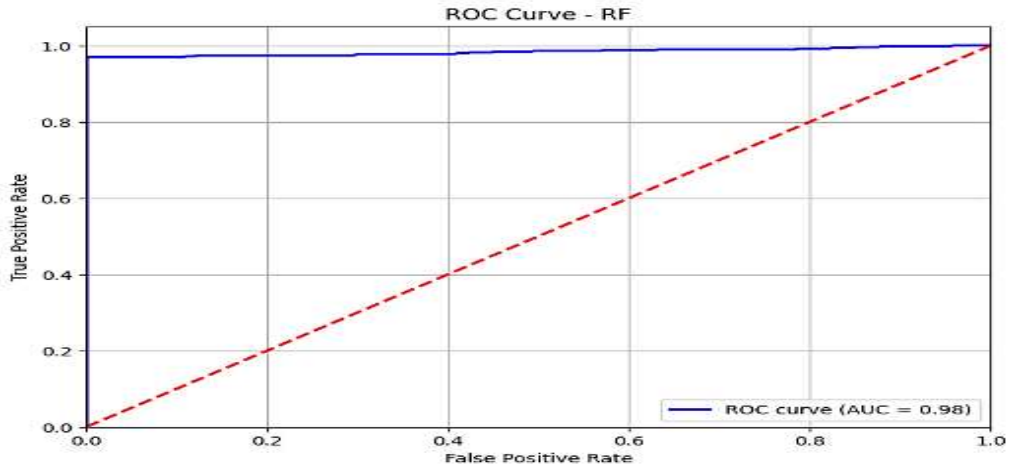

Figure 7: ROC for the RF Classifier

The ROC curve, representing the Random Forest (RF) model's performance, is shown in Figure 7. The TPR (True Positive rate) is on the bottom axis while the FPR (False Positive rate) is on the top axis. Dashed red lines represent a random classifier baseline, whereas solid blue lines represent the RF model. An improved model will have a blue curve that is more closely aligned with the top left corner. Differentiating between the two groups is clearly excelled by the RF model, with an area under the curve (AUC) of 0.98.

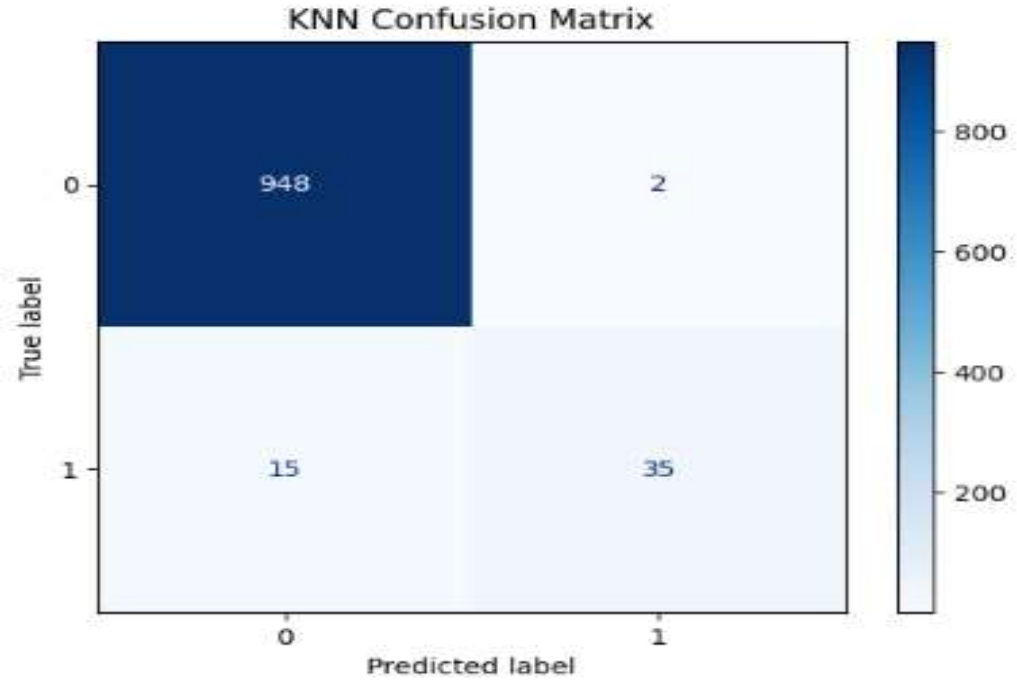

Figure 8: Confusion Matrix of the KNN Model

An overview of the model's predictions compared to the actual results is shown in Figure 8, which is a confusion matrix. The True Negatives (948) and True Positives (35), which signify the cases accurately identified for each class, are displayed in the matrix. The varying shades of blue provide a visual indication of the classification performance.

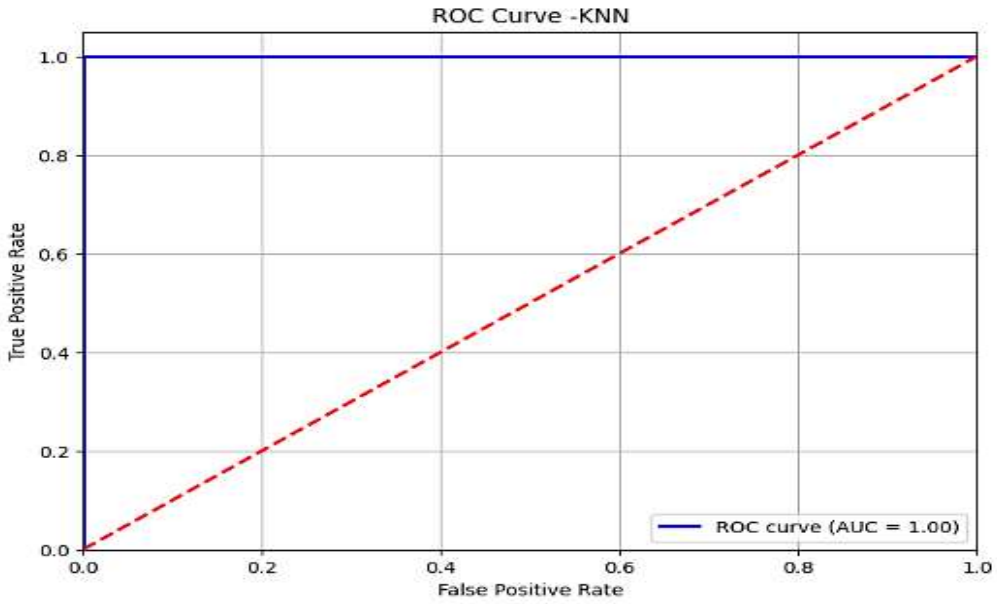

Figure 9: ROC for the KNN Model

Figure 9 shows the ROC curve for the K-Nearest Neighbors (KNN) model. Two axes are at your disposal: one showing the True Positive Rate (TPR) and the other the False Positive Rate (FPR). The solid blue line represents the KNN model, whereas the dashed red line shows a random classifier baseline. Good performance is shown by the curve's proximity to the top-left corner, and a 1.00 AUC indicates that the categorization between the two classes is flawless.

Comparative Analysis

The performance of the proposed KNN and RF models was determined by the comparative analysis with other reliable predictive models. A comparison of performances of various models in engaging the customers in the retail business has been provided in Table III. The highest model accuracy, 97.90, was obtained using K-Nearest Neighbors (KNN), and very closely there was Random Forest (RF) at 96.46%. Convolutional Neural Network (CNN) also achieved high accuracy of 94% and XGBoost accuracy was low with 71.75%, respectively. The findings suggest that KNN and RF are more effective than the other models, which proves their applicability to customer engagement prediction in the retail sphere.

Table 3: Performance Comparison of Different Models for Customer Engagement in Retail Business

Models	Accuracy
CNN [16]	94
XGBoost [17]	71.75
KNN	97.90
RF	96.46

The suggested KNN and RF systems are highly effective in predicting customer engagement in the retailing business. These two models remain highly effective in comparison to the traditional machine learning and deep learning models, in that they demonstrate their ability to produce a reliable and accurate result. KNN has high accuracy in determining the customers who are really engaged and is useful in target marketing and the use of individualized engagement strategies. Conversely, RF has a better balance between the recall and total predictive performance where majority of the engaged customers are accurately recognized. Collectively, these models have

strong solutions to improve customer engagement, assist in making better decisions, and achieve better business results in the retail industry.

CONCLUSION AND RECOMMENDATIONS

Conclusion

Customer interaction in retail business is crucial for fostering loyalty, increasing sales, and ensuring long-term profitability, as it reflects the depth of engagement and trust between businesses and consumers. This paper explored the use of ML models to determine customer engagement using the Online Retail dataset. Two models, KNN and RF, were implemented and tested to evaluate their effectiveness in identifying engaged customers. The results showed that KNN achieved an accuracy of 97.90%, while RF reached 96.46%. These findings demonstrated the strong predictive ability of both models in retail analytics. High-quality predictions such as these can enable businesses to tailor marketing campaigns, optimize inventory management, and improve customer retention. Although KNN slightly outperformed RF in terms of accuracy, the latter displayed better balance across performance metrics, making it more versatile for different engagement scenarios. Overall, the study highlights that machine learning offers valuable tools for analyzing customer engagement and supporting data-driven decision-making.

Recommendations

Future improvements, incorporating advanced models such as Gradient Boosting or deep learning, as well as integrating real-time data streams and customer sentiment from social media or reviews, could provide deeper insights into engagement patterns.

REFERENCES

[1] A. Kumar, "Retail Sector : Growth and challenges perspective in India," *Int. J. Emerg. Technol.*, vol. 5, no. 1, pp. 69–73, 2014.

[2] V. Shankar, "Big Data and Analytics in Retailing," *NIM Mark. Intell. Rev.*, vol. 11, no. 1, pp. 36–40, May 2019, doi: 10.2478/nimmir-2019-0006.

[3] H. Mahobia and R. K. D. Dubey, "A Study on Factors Influencing Consumer Engagement in Retail," *Int. J. Res. IT Manag.*, vol. 6, no. 2, 2016.

[4] T. K. Rao, "Reimagining Retail : AI-Driven Personalization and the Future of Customer Experience," *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 8, no. 2, 2019, doi: 10.15680/IJIRSET.2019.0802106.

[5] A. M. Choudhury and K. Nur, "A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior," in *2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)*, IEEE, Jan. 2019, pp. 242–247. doi: 10.1109/ICREST.2019.8644458.

[6] Z. H. Kilimci *et al.*, "An Improved Demand Forecasting Model Using Deep Learning Approach and Proposed Decision Integration Strategy for Supply Chain," *Complexity*, vol. 2019, no. 1, Jan. 2019, doi: 10.1155/2019/9067367.

[7] S. S. S. Neeli, "The Significance of NoSQL Databases : Strategic Business Approaches and Management Techniques," *J. Adv. Dev. Res.*, vol. 10, no. 1, p. 11, 2019.

[8] F. Naz and F. Popowich, "Mining Retail Telecommunication Data to Predict Profitability," in *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, IEEE, Aug. 2019, pp. 1–5. doi: 10.1109/PACRIM47961.2019.8985083.

[9] M. A. I. Arif, S. I. Sany, F. I. Nahin, and A. S. A. Rabby, "Comparison Study: Product Demand Forecasting with Machine Learning for Shop," in *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, IEEE, Nov. 2019, pp. 171–176. doi: 10.1109/SMART46866.2019.9117395.

[10] A. Bhatnagar and S. Srivastava, "A Robust Model for Churn Prediction using Supervised Machine Learning," in *Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing, IACC 2019*, 2019. doi: 10.1109/IACC48062.2019.8971494.

[11] L. Liu, B. Zhou, Z. Zou, S. C. Yeh, and L. Zheng, "A Smart Unstaffed Retail Shop Based on Artificial Intelligence and IoT," in *IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks, CAMAD*, 2018. doi: 10.1109/CAMAD.2018.8514988.

[12] K. Kim and J.-H. Lee, "Bayesian Optimization of Customer Churn Predictive Model," in *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*, IEEE, Dec. 2018, pp. 85–88. doi: 10.1109/SCIS-ISIS.2018.00024.

[13] H. Amnur, "Customer Relationship Management and Machine Learning Technology for Identifying the Customer," *JOIV Int. J. Informatics Vis.*, vol. 1, no. 1, pp. 12–15, Mar. 2017, doi: 10.30630/joiv.1.1.10.

[14] Y. Kaneko and K. Yada, "A Deep Learning Approach for the Prediction of Retail Store Sales," *IEEE Int. Conf. Data Min. Work. ICDMW*, vol. 0, pp. 531–537, 2016, doi: 10.1109/ICDMW.2016.0082.

[15] A. J. Gallego, J. Calvo-Zaragoza, J. J. Valero-Mas, and J. R. Rico-Juan, "Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation," *Pattern Recognit.*, vol. 74, pp. 531–543, 2018, doi: 10.1016/j.patcog.2017.09.038.

[16] A. Generosi, S. Ceccacci, and M. Mengoni, "A deep learning-based system to track and analyze customer behavior in retail stores," in *2018 IEEE 8th International Conference on Consumer Electronics - Berlin (ICCE-Berlin)*, IEEE, Sep. 2018, pp. 1–6. doi: 10.1109/ICCE-Berlin.2018.8576169.

[17] A. P. Patil, M. P. Deepshika, S. Mittal, S. Shetty, S. S. Hiremath, and Y. E. Patil, "Customer churn prediction for retail business," *2017 Int. Conf. Energy, Commun. Data Anal. Soft Comput. ICECDS 2017*, pp. 845–851, 2018, doi: 10.1109/ICECDS.2017.8389557.

1. Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Big Text Data Analysis for Sentiment Classification in Product Reviews Using Advanced Large Language Models. *International Journal of AI, BigData, Computational and Management Studies*, *2*(2), 55-65.

2. Gangineni, V. N., Tyagadurgam, M. S. V., Chalasani, R., Bhumireddy, J. R., & Penmetsa, M. (2021). Strengthening Cybersecurity Governance: The Impact of Firewalls on Risk Management. *International Journal of AI, BigData, Computational and Management Studies*, *2*, 10-63282.

3. Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., & Gangineni, V. N. (2021). An Advanced Machine Learning Models Design for Fraud Identification in Healthcare Insurance. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(1), 26-34.

4. Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., & Polam, R. M. (2021). Advanced Machine Learning Models for Detecting and Classifying Financial Fraud in Big Data-Driven. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(3), 39-46.

5. Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2021). Enhancing IoT (Internet of Things) Security Through Intelligent Intrusion Detection Using ML Models. *International Journal of Emerging Research in Engineering and Technology*, *2*(1), 27-36.

6. Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2021). Smart Healthcare: Machine Learning-Based Classification of Epileptic Seizure Disease Using EEG Signal Analysis. *International Journal of Emerging Research in Engineering and Technology*, *2*(3), 61-70.

7. Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., & Kamarthapu, B. (2021). Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(2), 26-34.

8. Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., & Pabbineedi, S. (2021). Next-Generation Cybersecurity: The Role of AI and Quantum Computing in Threat Detection. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(4), 54-61.

9. Polu, A. R., Vattikonda, N., Gupta, A., Patchipulusu, H., Buddula, D. V. K. R., & Narra, B. (2021). Enhancing Marketing Analytics in Online Retailing through Machine Learning Classification Techniques. *Available at SSRN 5297803*.

10. Polu, A. R., Buddula, D. V. K. R., Narra, B., Gupta, A., Vattikonda, N., & Patchipulusu, H. (2021). Evolution of AI in Software Development and Cybersecurity: Unifying Automation, Innovation, and Protection in the Digital Age. *Available at SSRN 5266517*.

11. Polu, A. R., Vattikonda, N., Buddula, D. V. K. R., Narra, B., Patchipulusu, H., & Gupta, A. (2021). Integrating AI-Based Sentiment Analysis With Social Media Data For Enhanced Marketing Insights. *Available at SSRN 5266555*.

12. Buddula, D. V. K. R., Patchipulusu, H. H. S., Polu, A. R., Vattikonda, N., & Gupta, A. K. (2021). INTEGRATING AI-BASED SENTIMENT ANALYSIS WITH SOCIAL MEDIA DATA FOR ENHANCED MARKETING INSIGHTS. *Journal Homepage: http://www. ijesm. co. in*, *10*(2).

13. Gupta, A. K., Buddula, D. V. K. R., Patchipulusu, H. H. S., Polu, A. R., Narra, B., & Vattikonda, N. (2021). An Analysis of Crime Prediction and Classification Using Data Mining Techniques.

14. Rajiv, C., Mukund Sai, V. T., Venkataswamy Naidu, G., Sriram, P., & Mitra, P. (2022). Leveraging Big Datasets for Machine Learning-Based Anomaly Detection in Cybersecurity Network Traffic. *J Contemp Edu Theo Artific Intel: JCETAI/102*.

15. Sandeep Kumar, C., Srikanth Reddy, V., Ram Mohan, P., Bhavana, K., & Ajay Babu, K. (2022). Efficient Machine Learning Approaches for Intrusion Identification of DDoS Attacks in Cloud Networks. *J Contemp Edu Theo Artific Intel: JCETAI/101*.

16. Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. (2020). Big Data-Driven Time Series Forecasting for Financial Market Prediction: Deep Learning Models. *Journal of Artificial Intelligence and Big Data*, *2*(1), 153–164.DOI: 10.31586/jaibd.2022.1341

17. Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2022). Advance of AI-Based Predictive Models for Diagnosis of Alzheimer's Disease (AD) in Healthcare. *Journal of Artificial Intelligence and Big Data*, *2*(1), 141–152.DOI: 10.31586/jaibd.2022.1340

18. Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2022). Designing an Intelligent Cybersecurity Intrusion Identify Framework Using Advanced Machine Learning Models in Cloud Computing. *Universal Library of Engineering Technology*, (Issue).

19. Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2022). Leveraging Artificial Intelligence Algorithms for Risk Prediction in Life Insurance Service Industry. *Available at SSRN 5459694*.

20. Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Data Security in Cloud Computing: Encryption, Zero Trust, and Homomorphic Encryption. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(3), 70-80.

21. Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Tyagadurgam, M. S. V. Efficient Framework for Forecasting Auto Insurance Claims Utilizing Machine Learning Based Data-Driven Methodologies. *International Research Journal of Economics and Management Studies IRJEMS*, *1*(2).

22. Vattikonda, N., Gupta, A. K., Polu, A. R., Narra, B., Buddula, D. V. K. R., & Patchipulusu, H. H. S. (2022). Blockchain Technology in Supply Chain and Logistics: A Comprehensive Review of Applications, Challenges, and Innovations. *International Journal of Emerging Research in Engineering and Technology*, *3*(3), 99-107.

23. Narra, B., Vattikonda, N., Gupta, A. K., Buddula, D. V. K. R., Patchipulusu, H. H. S., & Polu, A. R. (2022). Revolutionizing Marketing Analytics: A Data-Driven Machine Learning Framework for Churn Prediction. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(2), 112-121.

24. Polu, A. R., Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., Vattikonda, N., & Gupta, A. K. BLOCKCHAIN TECHNOLOGY AS A TOOL FOR CYBERSECURITY: STRENGTHS, WEAKNESSES, AND POTENTIAL APPLICATIONS.

25. Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. (2022). Big Data-Driven Time Series Forecasting for Financial Market Prediction: Deep Learning Models. *Journal of Artificial Intelligence and Big Data*, *2*(1), 153–164.DOI: 10.31586/jaibd.2022.1341

26. Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2022). Advance of AI-Based Predictive Models for Diagnosis of Alzheimer's Disease (AD) in Healthcare. *Journal of Artificial Intelligence and Big Data*, *2*(1), 141–152.DOI: 10.31586/jaibd.2022.1340

License