Customer Segmentation and Prediction with Logistic Regression and Flask Technology

Vijay Mane
Electronics and Telecommunication Engineering.
Vishwakarma Institute of Technology, Pune.
Pune,India
vijay.mane@vit.edu

Ruturaj Patil
Electronics and Telecommunication Engineering.
Vishwakarma Institute of Technology, Pune.
Pune,India
ruturaj.patil21@vit.edu

Shantanu Pawar
Electronics and Telecommunication Engineering.
Vishwakarma Institute of Technology, Pune,India
shantanu.pawar21@vit.edu

Nagesh Pujari
Electronics and Telecommunication Engineering
Vishwakarma Institute Of Technology, Pune,India
nagesh.pujari21@vit.edu

***Abstract* — The project centers on predicting customer churn using a dataset of 7,044 entries and 21 columns, integrated into a user-friendly Flask application. The system allows users to input data manually or via CSV file uploads, facilitating easy interaction and analysis. Employing machine learning algorithms such as logistic regression and correlation analysis, the model achieves high accuracy in forecasting churn, with a precision of 95.57% and an overall accuracy of 92.08%. The application's intuitive interface and predictive capabilities enable businesses to anticipate customer attrition, optimize retention strategies, and enhance decision-making processes. By offering real-time feedback and streamlined outputs, this research significantly contributes to customer relationship management, empowering organizations across various industries to mitigate churn risks and drive sustainable growth. This system not only improves operational efficiency but also provides actionable insights that are critical for maintaining competitive advantage. Continuous monitoring and model refinement ensure that the tool remains effective in adapting to changing business environments, further enhancing its value for long-term success.**

Keywords— Machine Learning, Business Optimization, Logistic Regression, Customer Churn.

I. Introduction

In today's competitive business landscape, customer retention is paramount for sustaining growth and profitability. Customer churn, the phenomenon of customers ceasing their relationship with a company, poses a significant challenge across various industries. Identifying potential churners and implementing proactive retention strategies are essential for mitigating revenue loss and maintaining a loyal customer base. Traditional methods relying on manual analysis often fall short in providing timely insights and accurate predictions.

To address this challenge, our project focuses on developing a robust customer churn prediction system leveraging machine learning techniques. By analyzing historical customer data, including contract details, service usage, and tenure, our model aims to forecast the likelihood of churn for individual customers. Through this predictive analytics approach, businesses can prioritize retention efforts, allocate resources efficiently, and enhance customer satisfaction. The integration of a user-friendly Flask application empowers businesses to interactively explore churn predictions, enabling informed decision-making and proactive retention strategies.

II. Literature Review

A novel hybrid classification technique for customer churn prediction called the Logit Leaf Model (LLM) that blends logistic regression and decision trees. LLM works in two steps: first, it uses decision tree rules to segment the data, and then it applies logistic regression models to each segment. This strategy is comparable to more sophisticated techniques like random forests and logistic model trees, and it increases predictive performance as assessed by AUC and top decile lift, surpassing individual decision trees and logistic regression. Better comprehensibility is another benefit of the LLM, as shown by a case study[1].

All significant phases of churn prediction models and looks at 212 papers from 2015 to 2023 on machine learning techniques for customer churn prediction are covered. It emphasizes the value of combining many elements, such as consumer behavior and demography, and focuses on industries like online gaming, finance, and telecommunications. In order to improve client retention and profitability, the analysis points out a research gap in profit-based evaluation criteria and suggests utilizing ensembles, deep learning strategies, and explainable methodologies[2].

A Python-based supervised machine learning model used to forecast client attrition. It concludes that KNN is 2.0% more accurate than Logistic Regression after comparing the performance of K-Nearest Neighbors (KNN) and Logistic Regression on labeled data. When compared to logistic regression, KNN is a more accurate method of predicting client attrition, according to the examination of the confusion matrix[3].

A churn prediction model for the telecom sector utilizing Random Forest (RF) and clustering algorithms to detect churn consumers and the drivers behind their churn this research proposed. The model achieves 88.63% accuracy in correctly classifying churn events. It also identifies key churn factors to help CRM improve productivity and targeted marketing campaigns. The model's effectiveness is evaluated using metrics such as accuracy, precision, recall, f-measure, and ROC area, demonstrating superior performance in churn classification and customer profiling[4].

A combination of k-means clustering and support vector machine (SVM) approaches to introduce a customer churn prediction model for B2C e-commerce uses. In order to identify core consumer groups and increase forecast accuracy, the model divides its customer base into three categories. When comparing SVM with logistic regression, the study discovers that SVM predicts churn more accurately[5].

The application of data mining techniques to forecast customer attrition in the telecom sector is examined in this article. In order to detect possible churners, it emphasizes how crucial it is to process and segment the customer data that has been collected. In order to assess how well different classification algorithms predict churn, the study runs a number of trials. The results highlight how important precise churn prediction is to a business's ability to keep customers and keep them satisfied[6].

Artificial neural networks (ANNs) used to develop a customer churn prediction model for the banking industry. Through the use of forward propagation and cross-validation to change hyperparameters, the model is able to forecast customer attrition with an accuracy rate of 86%. The outcomes show that ANNs work better in this situation than logistic regression. The study offers valuable perspectives on leveraging machine learning to improve customer retention and lower attrition rates, allowing banks to proactively detect and hold onto high-risk customers[7].

Using AI approaches, this study investigates client segmentation, profile, and sales prediction in direct marketing. It divides clients into three groups using RFM analysis: new customers, best customers, and intermittent customers. To find the best clusters, the K-means algorithm is used in conjunction with the Elbow method, Silhouette coefficient, and gap statistics[8].

This work uses a Markov model and K-Means clustering to propose a consumer segmentation and prediction model for retail transactional data. Based on past transaction data, clients are segmented using K-Means, and the resulting clusters are utilized to create a transition matrix that the Markov Model uses to forecast future cluster movements. This method supports budgeting, consumer targeting, and strategic marketing decisions by estimating the future worth of client groups. Large-scale marketing in industries like retail, insurance, and e-commerce benefits greatly from the model since it makes it possible for companies to efficiently segment, profile, and evaluate the lifecycle value of their customers[9].

This study employs RFM (Recency, Frequency, Monetary) metrics to analyze client segmentation and forecast future purchases in the e-commerce industry. The study uses one-dimensional clustering on each RFM column separately rather than directly clustering the RFM table. This approach helps the company's strategic marketing and customer retention efforts by optimizing segmentation and customer relationship management with machine learning algorithms, including ensemble techniques[10].

In order to handle the continually changing behavior of consumers, this article looks at the importance of customer segmentation and highlights the role that machine learning techniques play in this process. Common machine learning algorithms including AdaBoost, Decision Trees, and Logistic Regression are covered, and their performance and underlying concepts are examined via experiments. The study proposes future research areas in machine learning-based customer segmentation and attempts to offer insights for academics in adjacent fields[11].

The importance of correlation management and customer preservation for organizational success is discussed in this paper, with a focus on sectors including banking, finance, and telecommunication. It highlights how crucial it is to estimate customer churn accurately in order to hold onto key clients and maintain market competitiveness. The findings show that the K-Means clustering technique improves the accuracy of churn prediction in the telecommunications sector when used in conjunction with decision trees[12].

Specifically, this paper discusses the importance of correlation management and customer preservation for the performance of organizations in the banking, finance, and telecommunications sectors. It highlights how crucial precise customer churn prediction is to keeping valuable clients and maintaining market competitiveness. Fuzzy C-Means, Possibilistic Fuzzy C-Means, and K-Means are three clustering techniques for consumer segmentation that are examined in this work. Next, it employs a hybrid learning system that integrates supervised and unsupervised techniques, employing decision trees for both training and testing clusters. The findings show that decision trees and the K-Means clustering algorithm improve the accuracy of churn prediction in the telecom sector[13].

This paper investigates how to enhance consumer segmentation in e-commerce through the application of predictive neural networks. Through the use of clustering algorithms and analysis of data including product evaluations, buying and viewing trends, and time-based segments, the study attempts to separate the target market into groups with comparable traits and preferences. Neural

networks are used to determine which products consumers like, whereas feature extraction analyzes unigrams, bigrams, and trigrams. In order to comprehend consumer perceptions of brands and products, the study also takes sentiment mining and classification into account. The most appropriate brands from the input data set are identified by assessing the prediction accuracy[14].

An integrated system for customer segmentation and churn prediction in the telecom sector is presented in this article. Data pre-processing, factor analysis, churn prediction, customer segmentation, and customer behavior analytics are the six components that make up the framework. K-means clustering is then utilized for segmentation after factor analysis using Bayesian Logistic Regression has been utilized to uncover key features for churn customer segmentation. By using more accurate retention strategies, this integrated strategy helps telecom operators better control customer attrition strategies[15].

This article offers a novel approach to customer segmentation that accounts for variations in customer value over time. This methodology considers the temporal aspect of value swings, while traditional segmentation strategies focus on the value of the client at a specific moment in time. Forecasts that are based on past customer behavior are therefore more accurate. POS customer transactions are used to accomplish this[16].

In order to segment customers, a modified clustering algorithm is presented in this paper and is appropriate for clustering attributes of categorical data. Once each segment has been mined, frequent pattern mining is used to infer rules for forecasting consumers' future purchases. Understanding purchasing dependencies enables products to be grouped together, enabling businesses to provide clients with enticing offers and boost total revenues. To assess the efficacy of the suggested method, real-time database tests are carried out, highlighting the significance of efficiently clustering categorical data for purchase prediction in customer segmentation[17]

III. METHODOLOGY

The methodology begins with data acquisition, where relevant datasets are collected from various sources or generated internally. These datasets undergo initial exploration to understand their structure and contents. Following this, a comprehensive data preprocessing step is initiated. This involves handling missing values, outlier detection, and feature engineering to extract meaningful insights. Subsequently, the preprocessed data is divided into training and testing sets. Model selection is a critical phase where various machine learning algorithms are evaluated to identify the most suitable one for the task at hand. Once the model is selected, it undergoes training using the training dataset. During training, hyperparameters are tuned to optimize model performance. Once the model is trained, it undergoes evaluation using the testing dataset to assess its generalization ability.

After model evaluation, the next step involves deploying the trained model into a production environment. This deployment phase may involve integrating the model into existing systems or developing a standalone application. User interface design is crucial to ensure ease of interaction with the deployed model. Depending on the application requirements, users may have the option to input data manually or upload files for processing. Post-deployment monitoring is essential to track the model's performance over time and ensure its continued effectiveness. Additionally, periodic model retraining may be necessary to adapt to changing data distributions or business requirements. Continuous improvement is emphasized, with feedback loops established to incorporate user feedback and enhance the model iteratively.

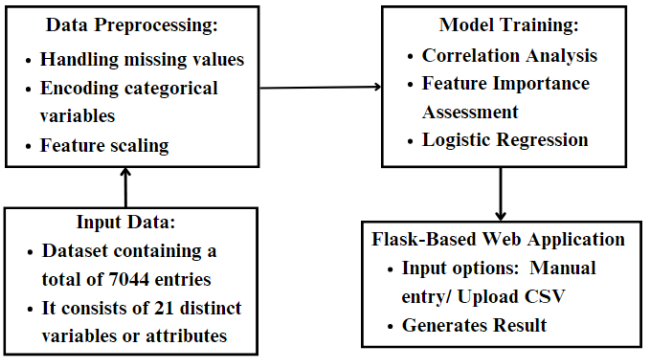


Fig. 1 Block Diagram

The fig.1 block diagram illustrates the step-by-step process involved in the project. It begins with the input stage, where raw data is fed into the system. In the project, the input data consists of customer-related information, such as contract details, online security status, and monthly charges. Following the input stage, the data undergoes preprocessing, which includes tasks like data cleaning, normalization, and feature extraction. This stage is crucial for preparing the data in a format suitable for model training. Once the preprocessing is complete, the data is fed into the model training phase, where machine learning algorithms are employed to build predictive models. In this project, correlation analysis and logistic regression are used as part of the model training process. Finally, the trained model is deployed in a user-friendly interface, allowing stakeholders to interact with the system easily. Users have the option to input data manually or upload CSV files for prediction, enhancing the accessibility and usability of the system. Overall, the block diagram provides a comprehensive overview of the project workflow, from data input to model deployment, highlighting the key steps involved in predicting customer churn.

IV. Results and Discussion

The Flask-based website has been created to show an intuitive interface designed to facilitate seamless interaction and efficient churn prediction. Featuring two primary options as shown in main page fig.2, users can choose between manual data entry as shown in fig.3 or CSV file upload for inputting customer data as shown in fig.4 . The manual data entry option offers a user-friendly form where users can input essential customer details such as contract type, online security status, tenure, and more. On the other hand, the CSV file upload functionality enables users to upload bulk data effortlessly, streamlining the prediction process. This dual approach enhances accessibility and accommodates varying user preferences, ensuring a smooth and tailored user experience.

Fig. 2 Homepage Page

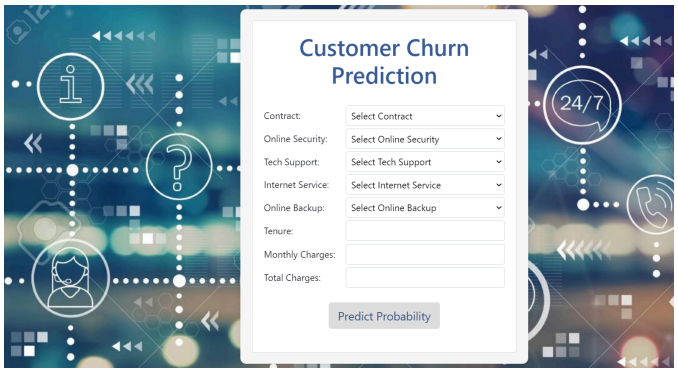


Fig. 3 Prediction via manual entry of data

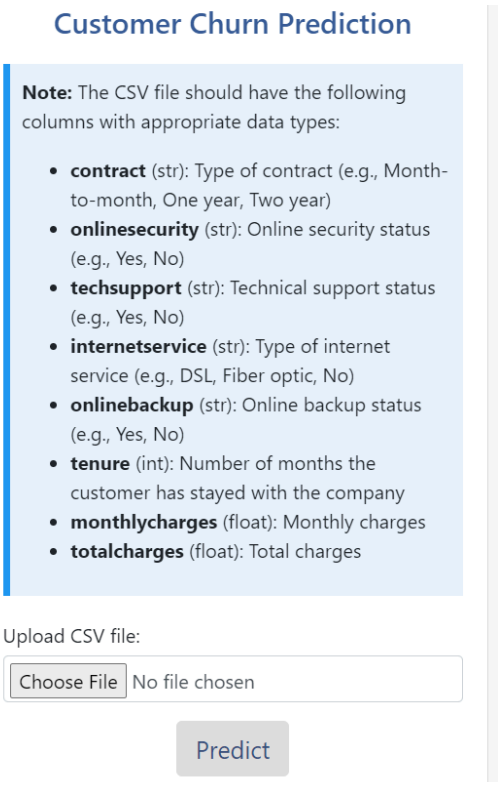


Fig. 4 Prediction via uploading csv file

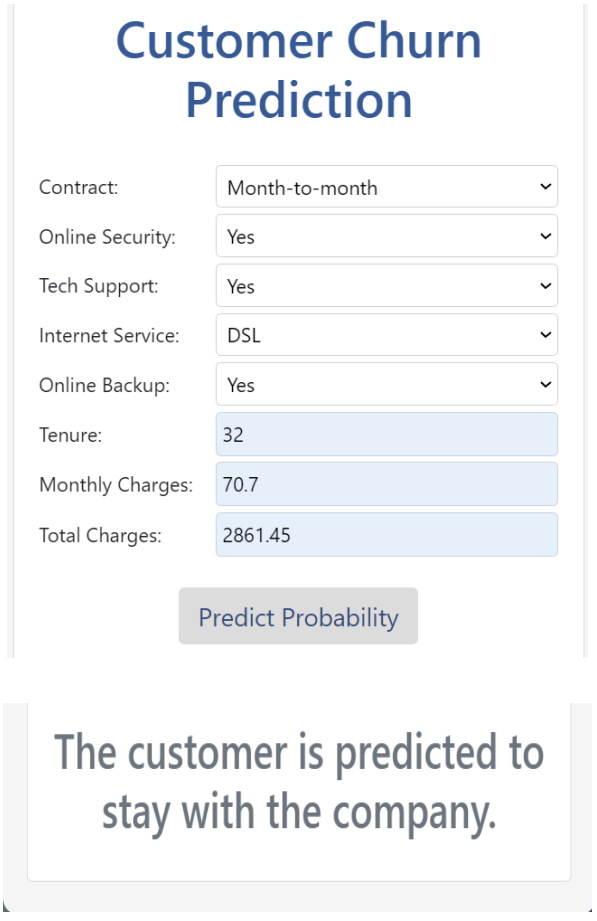
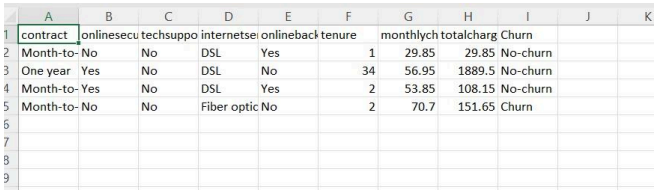


Fig. 5 Results

The screenshot of the interface is shown in fig.5. After entering the data manually into the form, the system processes the input and promptly displays the churn prediction result directly on the webpage. This seamless interaction provides users with instant feedback on the likelihood of churn based on the entered customer information. By eliminating the need for additional steps or downloads, this real-time prediction feature enhances user experience and facilitates quick decision-making for businesses. Additionally, users can interpret the prediction outcome immediately and take appropriate actions to mitigate potential churn risks, thereby optimizing customer retention strategies.

	A	B	C	D	E	F	G	H	I	J	K	L
1	contract	onlinesecu	techsuppo	internetse	onlineback	tenure	monthlych	totalcharges				
2	Month-to-	No	No	DSL	Yes	1	29.85	29.85				
3	One year	Yes	No	DSL	No	34	56.95	1889.5				
4	Month-to-	Yes	No	DSL	Yes	2	53.85	108.15				
5	Month-to-	No	No	Fiber optic	No	2	70.7	151.65				
6												
7												
8												
9												
10												
11												

Fig. 6 Input csv file



	A	B	C	D	E	F	G	H	I	J	K
1	contract	onlinesecu	techsuppo	internetse	onlineback	tenure	monthlych	totalcharg	Churn		
2	Month-to-	No	No	DSL	Yes	1	29.85	29.85	No-churn		
3	One year	Yes	No	DSL	No	34	56.95	1889.5	No-churn		
4	Month-to-	Yes	No	DSL	Yes	2	53.85	108.15	No-churn		
5	Month-to-	No	No	Fiber optic	No	2	70.7	151.65	Churn		
6											
7											
8											
9											

Fig. 7 Output csv file

After submitting the CSV file with multiple data entries as shown in fig.6 and clicking on the predict button, the Flask-based website initiates the churn prediction process. The uploaded dataset is processed through the trained churn prediction model, which analyzes each entry to forecast the likelihood of customer churn. Once the prediction process is completed, a new CSV file is generated or downloaded, containing the original dataset with an additional column appended to indicate the predicted churn status for each customer as shown in fig.7. This streamlined output format enables users to quickly assess the churn risk associated with individual customers and devise targeted retention strategies. By seamlessly integrating prediction outcomes with actionable insights, businesses can optimize customer retention efforts and enhance overall operational efficiency.

The machine learning component of the churn prediction system utilizes advanced techniques to achieve high predictive accuracy. Leveraging sophisticated algorithms, the model undergoes rigorous training on a comprehensive dataset comprising various customer features. Through iterative learning iterations, the model fine-tunes its parameters to optimize performance and enhance predictive capabilities. As a result, the churn prediction model achieves an impressive accuracy of 92.08%, reflecting its proficiency in distinguishing between customers likely to churn and those likely to remain.

Furthermore, the model's performance is evaluated using key metrics such as precision, recall, and F-measure. With a precision score of 95.57%, the model exhibits a high degree of accuracy in correctly identifying churn cases among the predicted positives. The recall score of 88.52% signifies the model's ability to capture the majority of actual churn instances from the dataset. Moreover, the F-measure, which combines precision and recall into a single metric, yields a robust score of 91.91%.These evaluation metrics demonstrate the effectiveness and reliability of the machine learning approach in churn prediction. By leveraging cutting-edge techniques and meticulous model evaluation, businesses can gain valuable insights into customer behavior and proactively mitigate churn risk.

To improve performance and accuracy in predicting customer churn, focus should be on feature selection, advanced algorithms like ensemble methods, and hyperparameter tuning techniques such as Grid Search. Balancing the dataset and regularizing models like logistic regression using L1 or L2 methods can further enhance accuracy. Preprocessing is crucial—cleaning data, encoding categorical variables, handling outliers, and imputing missing values using techniques like KNN or MICE will refine the dataset. Dataset details should include thorough descriptions of features and class distributions to provide clarity. Additionally, continuous model optimization through cross-validation and automation using machine learning pipelines improves performance. The high accuracy, precision, recall, and F-measure of the churn prediction model underscore its potential to drive actionable insights and inform strategic decision-making for sustainable business growth.

V. CONCLUSION

The development of a robust customer churn prediction system represents a significant advancement in enhancing business strategies and customer retention efforts. By harnessing the power of machine learning techniques, our system achieves high accuracy in forecasting churn, enabling businesses to proactively address customer attrition. The user-friendly interface, offering both manual data entry and CSV upload options, ensures accessibility and ease of use for stakeholders across various levels of technical expertise. Moreover, the seamless integration of predictive analytics into the decision-making process empowers businesses to optimize resources and prioritize efforts towards retaining valuable customers.

Overall, the implementation of this churn prediction system underscores the potential for leveraging data-driven insights to drive actionable outcomes and foster sustainable business growth. As organizations increasingly recognize the value of predictive analytics in anticipating customer behavior, this system serves as a valuable tool for staying ahead in a competitive landscape. Moving forward, continued refinement and adaptation of the system will further enhance its predictive capabilities, ultimately enabling businesses to foster stronger customer relationships and drive long-term success.

VI. References

1. De Caigny, Arno, Kristof Coussement, and Koen W. De Bock. "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees." *European Journal of Operational Research* 269, no. 2 (2018): 760-772..

2. A. Manzoor, M. Atif Qureshi, E. Kidney and L. Longo, "A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners," in *IEEE Access*, vol. 12, pp. 70434-70463, 2024, doi: 10.1109/ACCESS.2024.3402092.

3. A. Bhatnagar and S. Srivastava, "A Robust Model for Churn Prediction using Supervised Machine Learning," *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, Tiruchirappalli, India, 2019, pp. 45-49, doi: 10.1109/IACC48062.2019.8971494.

4. I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," in *IEEE Access*, vol. 7, pp. 60134-60149, 2019, doi: 10.1109/ACCESS.2019.2914999

5. Xiahou, Xiancheng, and Yoshio Harada. 2022. "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM" *Journal of Theoretical and Applied Electronic Commerce Research* 17, no. 2: 458-475. https://doi.org/10.3390/jtaer17020024.

6. L. F. Khalid, A. Mohsin Abdulazeez, D. Q. Zeebaree, F. Y. H. Ahmed and D. A. Zebari, "Customer Churn Prediction in Telecommunications Industry Based on Data Mining," *2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, Langkawi Island, Malaysia, 2021, pp. 1-6, doi: 10.1109/ISIEA51897.2021.9509988

7. B. Baby, Z. Dawod, M. S. Sharif and W. Elmedani, "Customer Churn Prediction Model Using Artificial Neural Network: A Case Study in Banking," *2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Sakheer, Bahrain, 2023, pp. 154-161, doi: 10.1109/3ICT60104.2023.10391374.

8. Kasem, M.S., Hamada, M. & Taj-Eddin, I. Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Comput & Applic* **36**, 4995–5005 (2024). https://doi.org/10.1007/s00521-023-09339-6

9. A.S. Harish and C. Malathy, "Customer Segment Prediction on Retail Transactional Data Using K-Means and Markov Model," *Intell. Automat. Soft Comput.* , vol. 36, no. 1, pp. 589-600. 2023. https://doi.org/10.32604/iasc.2023.032030

10. A. Patra, R. Khan and S. Vijayalakshmi, "Customer Segmentation and Future Purchase Prediction using RFM measures," *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Greater Noida, India, 2022, pp. 753-759, doi: 10.1109/ICAC3N56670.2022.10073993.

11. Wang, Zhiyue. "Customer Segmentation Based on Machine Learning Methods." *Highlights in Science, Engineering and Technology* 92 (2024): 126-132.

12. Sivasankar, E., and J. Vijaya. "Customer segmentation by various clustering approaches and building an effective hybrid learning system on churn prediction dataset." In *Computational Intelligence in Data Mining: Proceedings of the International Conference on CIDM, 10-11 December 2016*, pp. 181-191. Springer Singapore, 2017.

13. Khalili-Damghani, Kaveh, Farshid Abdi, and Shaghayegh Abolmakarem. "Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries." *Applied Soft Computing* 73 (2018): 816-828.

14. Singh, Himanshu, and Shubhangi Neware. "Improving customer segmentation in e-commerce using predictive neural network." *International Journal* 9, no. 2 (2020).

15. Wu, Shuli, Wei-Chuen Yau, Thian-Song Ong, and Siew-Chin Chong. "Integrated churn prediction and customer segmentation framework for telco business." *Ieee Access* 9 (2021): 62118-62136.

16. Hosseini, Monireh, and Mostafa Shabani. "New approach to customer segmentation based on changes in customer value." *Journal of Marketing Analytics* 3 (2015): 110-121.

17. Singh, Juhi, and Mandeep Mittal. "Customer's purchase prediction using customer segmentation approach for clustering of categorical data." *Management and Production Engineering Review* 12 (2021).