INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 04, April 2023, pp : 983-988

e-ISSN : 2583-1062

Impact Factor : 5.725

CUSTOMER BEHAVIOUR ANALYSIS FOR EVALUATING BUYING PATTERNS USING SVM, LOGISTIC REGRESSION AND KNN

Roshni Mandli[1], Jaiminee Patel[2]

[1]PG Student, Computer Engineering, KITRC, Kalol, Gujarat, India

[2]Assistant Professor, Computer Engineering, KITRC, Kalol, Gujarat, India

ABSTRACT

Every business venture relies on customers, so it is crucial to comprehend their behaviour if businessmen want to be successful. The purpose of this study is to examine buying patterns by analyzing consumer behaviour. While existing systems for customer data analysis have produced valuable results, they lack the ability to categorize data into different parameters and often do not combine product and customer clusters in their analysis. This leads to ineffective targeting of consumers, resulting in wasted resources and time on inactive audiences. This study includes all important variables to produce the best results. The customer dataset is mined for useful insights using natural language processing and exploratory data analysis. In order to categorize clients and products on the basis of purchasing patterns in product categories, k-means clustering is utilized. Each client is given a cluster, which is determined by machine learning classifiers like logistic regression, SVM, KNN, and XGBoost. The top-performing classifier among these is the logistic regression classifier, which is further enhanced by using it in conjunction with the KNN approach and adding outlier detection to it. Businessmen can efficiently target particular customers by using segmented customer clusters to analyze their behaviour and purchasing patterns.

Keywords: Clustering, Analysis, Customer behaviour, Classification, Regression, Purchasing Pattern.

1. INTRODUCTION

Customers have a significant role in a company's success, so understanding consumer behaviour is essential to corporate strategy. Analysis of customer data is frequently used to learn more about consumer behaviour and buying habits. Although the system used for analyzing customer data has produced a great deal of useful research, it does not categorize the data into different parameters [17]. Other than this, there have been few studies that have included clusters of customers and clusters of products as components of the analysis. Therefore, precise knowledge about consumer purchasing habits can't be widely recognized, and businesses are unable to successfully target their clients, wasting resources and time on promoting to inactive consumers. This system properly analyses buyer behaviour to determine patterns in purchasing. Advanced approaches are used to analyse various dataset parameters in order to accomplish this. The customer dataset is mined for useful insights using natural language processing and exploratory data analysis. We used the k-means clustering technique to categorize products and customers based on their purchasing patterns across various categories of products [19] [20]. Additionally, a variety of machine learning models are built, and the accuracy of the best classifier among these is improved even more in order to accurately categorize clients into various product categories. Finally, the system generates segmented customer clusters, which become helpful in understanding the behaviour of customers, allowing them to target customers effectively and increase profitability.The goal of this research work is to create a system that is able to group consumers according to how they often buy across various product categories [1][5]. Companies can better serve their consumers by gaining more accurate and insightful information about consumer behaviour. Customers will also profit from this method because they will get more appropriate and tailored product suggestions based on their interests and previous buying activity. Eventually, employing this model will allow businesses to more accurately understand the buying habits and preferences of their clients, allowing them to better target their particular customer base.

2. METHODOLOGY

This section presents the systematic approach used to analyze customer behaviour. The goal of this research is to provide businesses with valuable insights into their customers' behaviour, enabling them to effectively target specific customers. To achieve this objective, the methodology employed exploratory data analysis, natural language processing, clustering and machine learning algorithms. The subsections below describe these methodologies in brief.

1.1 Data Collection

We acquired a vast purchasing dataset [16] from the Kaggle repository, which included every transaction for a UK-based and registered non-store online retailer concentrating on unique multi-occasion gifts over an entire year. The dataset was checked for missing values and duplicates before analysis, resulting in clean data that was prepared for exploratory data analysis.

1.2 Exploratory Data Analysis

We analyzed the customer dataset by examining the number of orders by country, identifying unique products and customers, managing cancelled orders, exploring basket pricing and analyzing monthly order volume. We also determined the top ten products and customers based on amount sold and basket price, and generated a pie chart to visualize the distribution of order amounts. These insights laid the groundwork for further investigation and a better understanding of customer behavior [19].

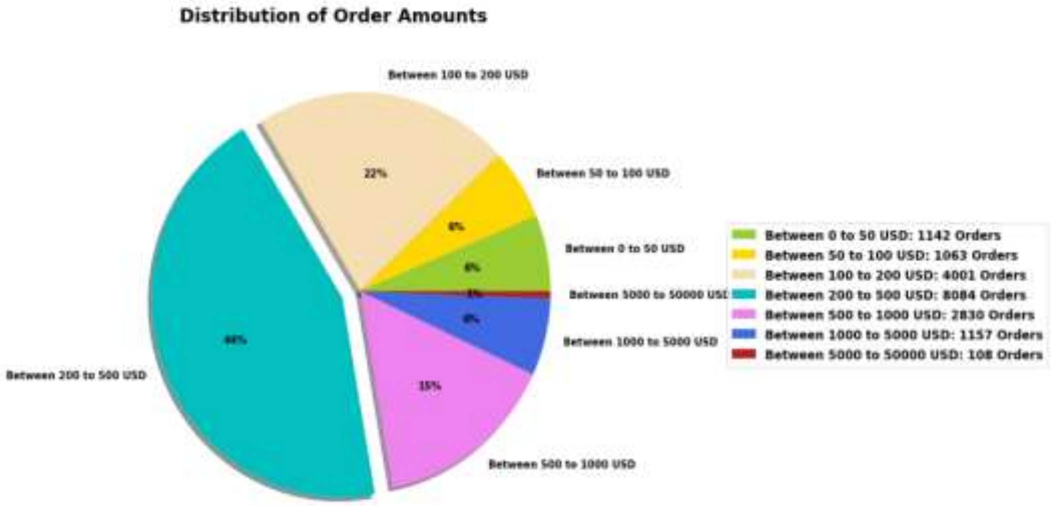


Figure 1: Distribution of Order Amounts

1.3 Natural Language Processing

We gathered useful information from the product description column using natural language processing techniques. We used the 'keywords_inventory' function, which makes use of the nltk library's snowBallStemmer and word_tokenize functions to generate keywords, their roots, and frequency of occurrence. This provided us with the most frequently graph of most occurring keywords and gave us insights on the kinds of products that clients desire, as well as the ability to generate word clouds for improved visual representation. These discoveries aided our understanding of client behaviour and preferences.

1.4 Clustering

To cluster products and customers, we used the K-Means clustering algorithm and optimised its hyperparameters[19][20]. Using the silhouette score, we established the most appropriate number of clusters. We classified products into five categories based on NLP findings and price ranges. Customers were classified into ten categories based on their similar buying habits across product categories. These clusters enable better knowledge of customer preferences and behaviour, resulting in more effective marketing and sales strategies.

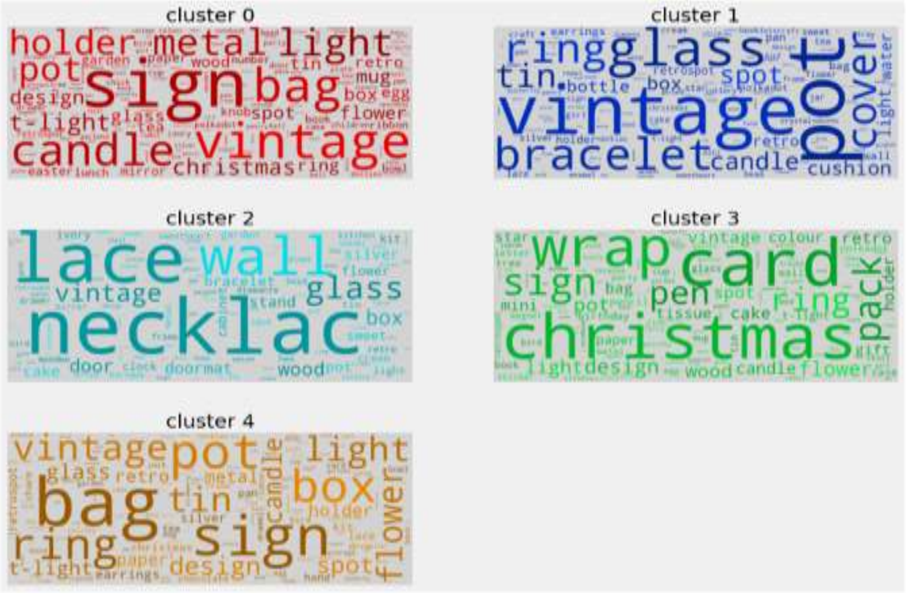


Figure 2: Word Clouds

1.5 Machine learning algorithms

We trained several algorithms under this portion, including support vector machine, logistic regression, k-nearest neighbour, and XGBoost classifier[3]. We used the GridSearchCV method to optimise the hyperparameters of these models. We created a learning curve and a confusion matrix to evaluate the model's quality. We also calculated the RMSE and accuracy for each model and discovered that logistic regression performed best. Furthermore, we used outlier detection using the interquartile range (IQR) method and a hybrid approach with K-NN to improve the performance of logistic regression. This innovative and creative approach resulted in an improved performance of the traditional logistic regression algorithm. The confusion matrix and learning curve for the hybrid logistic regression model, which was developed by enhancing the traditional logistic regression algorithm with outlier detection and incorporating k-nearest neighbour, are presented below.

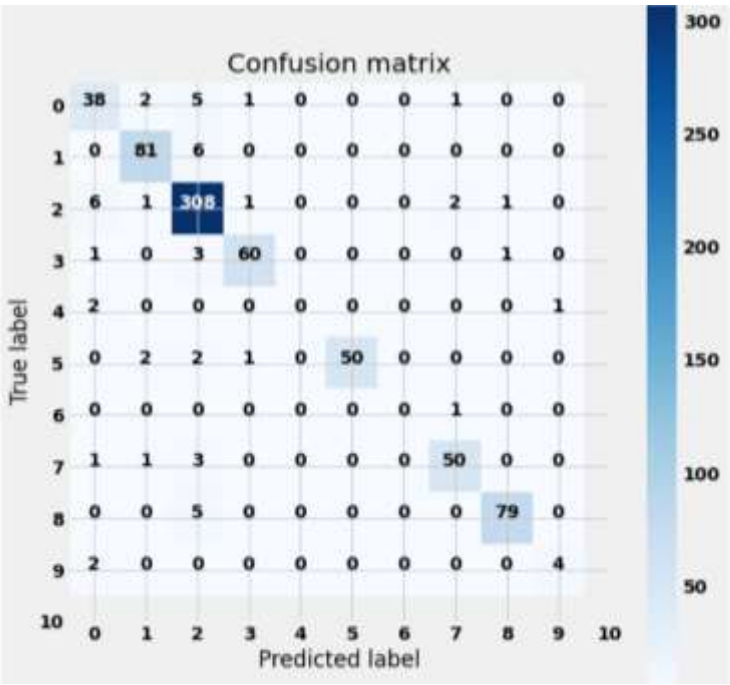


Figure 3: Confusion Matrix of Hybrid Approach

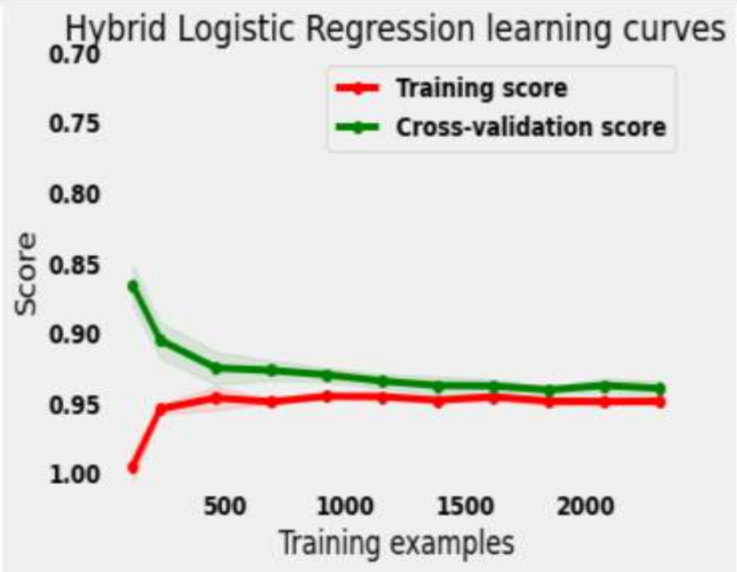


Figure 4: Learning curve for Hybrid approach

3. MODELING AND ANALYSIS

In this section, a comprehensive approach for analyzing customer behaviour to evaluate buying patterns is presented. The proposed procedure involves a series of steps, which are illustrated in the accompanying flow diagram. This systematic approach ensures a thorough analysis of customer data and provides valuable insights to enable effective targeting of specific customers.

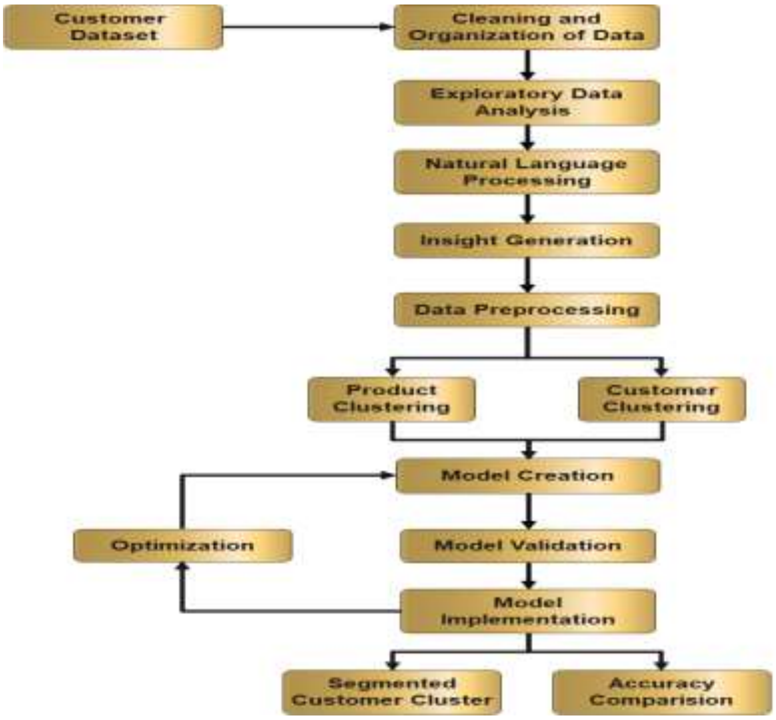


Figure 5: Proposed system

Step 1: Import the customer transaction dataset into the system.

Step 2: Prepare the customer dataset for accurate analysis by cleaning and organizing.

Step 3: Conduct exploratory data analysis to learn more about customer behaviour.

Step 4: Use natural language processing to extract useful information from consumer data.

Step 5: Generate insights that will inspire business decision-making.

Step 6: Prepare the data for clustering by pre-processing it.

Step 7: Group products and customers to better understand their spending habits.

Step 8: Build models with machine learning methods like SVM, logistic regression, KNN, and XGBoost.

Step 9: Validate models for reliable performance.

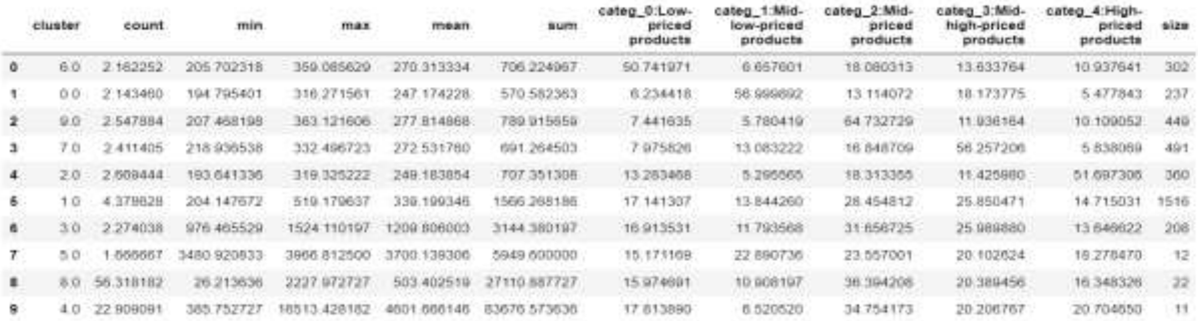
Step 10: Improve the accuracy of the best-performing model.

Step 11: Make use of the models to separate consumers into clusters according to their purchasing patterns in various categories of products to get segmented customer clusters.

Step 12: Review the accuracy of every machine learning model to figure out which one is the most accurate.

4. RESULTS AND DISCUSSION

Our system's output shows a clustering of customers based on their spending habits in various product categories, allowing businesspeople to understand their customers' behaviour and purchasing patterns. The table shows information about the clusters, such as the average number of visits and total amount spent by consumers in each cluster, as well as the percentages of total amount spent in various product categories. The findings indicate that there are considerable differences in purchasing patterns among different consumer groups, implying that efficient targeting of certain clients might help organizations maximize revenues.



	cluster	count	min	max	mean	sum	categ_0:Low-priced products	categ_1:Mid-low-priced products	categ_2:Mid-priced products	categ_3:Mid-high-priced products	categ_4:High-priced products	size
0	6.0	2.162252	205.702318	359.085629	270.313334	706.224967	50.741971	6.657601	18.080313	13.633764	10.937641	302
1	0.0	2.143460	194.795401	316.271561	247.174228	570.582363	6.234418	56.999892	13.114072	18.173775	5.477843	237
2	9.0	2.547884	207.468198	363.121606	277.814868	789.915659	7.441635	5.780419	64.732729	11.936164	10.109052	449
3	7.0	2.411405	218.936538	332.496723	272.531780	691.264503	7.975826	13.083222	16.848709	56.257206	5.838069	491
4	2.0	2.669444	193.641336	319.325222	249.183854	707.351308	13.283468	5.295565	18.313365	11.425880	51.697306	360
5	1.0	4.379828	204.147672	519.179637	339.199346	1566.268186	17.141307	13.844260	28.454812	25.850471	14.715031	1516
6	3.0	2.274038	976.465529	1524.110197	1209.806003	3144.380197	16.913531	11.793568	31.656725	25.989880	13.646622	208
7	5.0	1.666667	3480.920833	3866.812500	3700.139306	5949.600000	15.171168	22.890736	23.557001	20.102624	18.278470	12
8	8.0	56.318182	26.213636	2227.972727	503.402519	27110.887727	15.974691	10.908197	36.394208	20.389456	16.348326	22
9	4.0	22.909091	365.752727	18513.428182	4601.666146	83676.573636	17.813890	6.520520	34.754173	20.206767	20.704650	11

Figure 6: Segmented customer clusters

Below table provides a comparison of the accuracy and root mean squared error (RMSE) of five different machine learning algorithms: support vector machine (SVM), logistic regression, k-nearest neighbors (KNN), XGBoost classifier, and a Logistic Regression with Outlier Detection and KNN Hybridization **(**LODK Hybrid Algorithm**)**.

Table 1: Performance Comparison

Algorithm	Accuracy	RMSE
Support Vector Machine	81.30 %	2.18
Logistic Regression	88.91 %	1.82
K-Nearest Neighbour	82.67 %	2.18
XGBoost Classifier	88.00 %	1.89
Logistic Regression with Outlier Detection and KNN Hybridization (LODK Hybrid Algorithm)	89.30 %	1.73

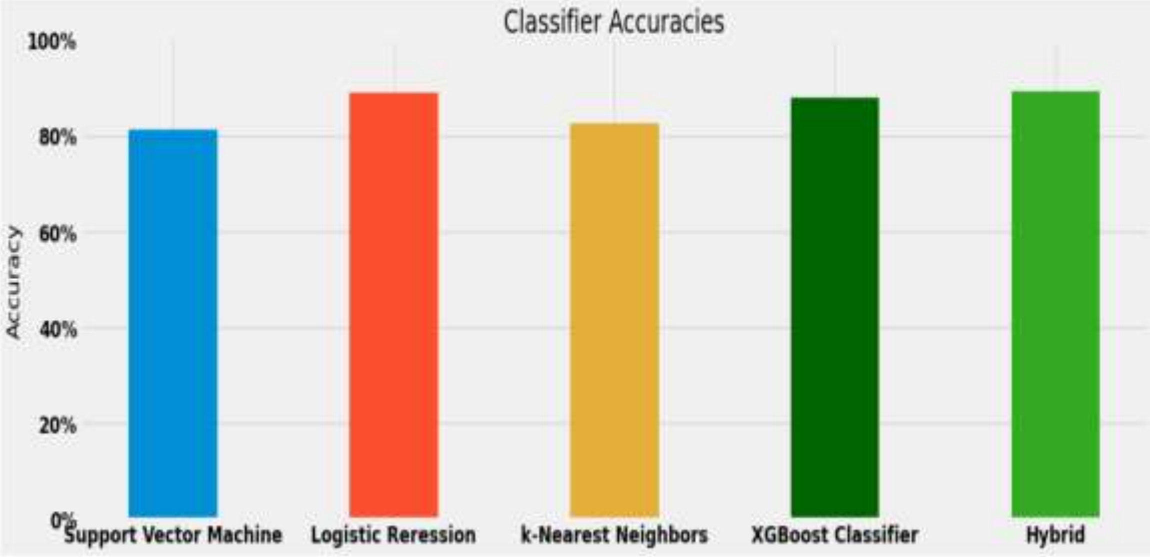


Figure 7: Accuracy Comparison

After comparing the accuracy of the support vector machine, k-nearest neighbour, XGBoost classifier, and logistic regression, it was discovered that logistic regression was the most accurate. Building on this achievement, an effort was undertaken to improve the traditional logistic regression model's performance. To do this, the interquartile range (IQR) method was used to detect outliers. A hybridization with the K-nearest neighbour was also used to reduce false positive and false negative values. By combining these strategies, the modified logistic regression model displayed improved predictive capabilities, offering precise analysis and customer behaviour prediction.

5. CONCLUSION

Our system provides valuable insights into customer behaviour and buying patterns, allowing business owners to target specific customers effectively. By segmenting customers into clusters based on their spending habits, we can identify significant differences in purchasing patterns among different consumer groups, enabling businesses to maximize revenue. We have compared the accuracy and RMSE of five different machine learning algorithms and found that the modified logistic regression algorithm outperforms the others. The addition of outlier detection and hybridization with K-nearest neighbor algorithms has further improved the logistic regression model's performance. Our system provides an efficient and accurate approach for businesses to understand their customers' behaviour and buying patterns, allowing them to optimize their advertising and increase their revenue.

6. REFERENCES

[1] Spenrath, Y., Hassani, M., van Dongen, B.F. (2022). Online Prediction of Aggregated Retailer Consumer Behaviour. In: Munoz-Gama, J., Lu, X. (eds) Process Mining Workshops. ICPM 2021. Lecture Notes in Business Information Processing, vol 433. Springer, Cham.

[2] KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2017, Pages 1753–1762Gyusoo Kim and Seulgi Lee, “2014 Payment Research”, Bank of Korea, Vol. 2015, No. 1, Jan. 2015.

[3] Hassani, M., & Habets, S. (2021). Predicting Next Touch Point In A Customer Journey - A Use Case In Telecommunication. In 35th ECMS INTERNATIONAL CONFERENCE ON MODELLING AND SIMULATION (1 ed., Vol. 35, pp. 48-54). (Proceedings - European Council for Modelling and Simulation, ECMS).

[4] H C. D. Francescomarino, M. Dumas, F. M. Maggi and I. Teinemaa, "Clustering-Based Predictive Process Monitoring," in IEEE Transactions on Services Computing, vol. 12, no. 6, pp. 896-909, 1 Nov.-Dec. 2019, Doi: 10.1109/TSC.2016.2645153.

[5] X. Chen, Y. Fang, M. Yang, F. Nie, Z. Zhao and J. Z. Huang, "PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 3, pp. 559-572, 1 March 2018.

[6] X. Chen, W. Sun, B. Wang, Z. Li, X. Wang and Y. Ye, "Spectral Clustering of Customer Transaction Data With a Two-Level Subspace Weighting Method," in *IEEE Transactions on Cybernetics*, vol. 49, no. 9, pp. 3230-3241, Sept. 2019

[7] De Medeiros, A.K.A. *et al.* (2008). Process Mining Based on Clustering: A Quest for Precision. In: ter Hofstede, A., Benatallah, B., Paik, HY. (eds) Business Process Management Workshops. BPM 2007. Lecture Notes in Computer Science, vol 4928. Springer, Berlin, Heidelberg.

[8] Spenrath, Y., Hassani, M., Dongen, B.v., Tariq, H. (2021). Why Did My Consumer Shop? Learning an Efficient Distance Metric for Retailer Transaction Data. In: Dong, Y., Ifrim, G., Mladenić, D., Saunders, C., Van Hoecke, S. (eds) Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track. ECML PKDD 2020. Lecture Notes in Computer Science(), vol 12461. Springer, Cham

[9] ACM SIGMOD Record, Volume 25, Issue 2, June 1996, pp 103–114

[10] Tariq, M.U., Babar, M., Poulin, M. and Khattak, A.S. (2022), "Distributed model for customer churn prediction using convolutional neural network", Journal of Modelling in Management, Vol. 17 No. 3, pp. 853-863.

[11] Vijayaraman Balakumar et al., Asian Journal of Research in Social Sciences and Humanities, 2016

[12] Tax, N., Verenich, I., La Rosa, M., Dumas, M. (2017). Predictive Business Process Monitoring with LSTM Neural Networks. In: Dubois, E., Pohl, K. (eds) Advanced Information Systems Engineering. CAiSE 2017. Lecture Notes in Computer Science(), vol 10253. Springer, Cham

[13] Folino, F., Guarascio, M., Pontieri, L. (2012). Discovering Context-Aware Models for Predicting Business Process Performances. In: , et al. On the Move to Meaningful Internet Systems: OTM 2012. OTM 2012. Lecture Notes in Computer Science, vol 7565. Springer, Berlin, Heidelberg.

[14] A. Metzger et al., "Comparing and Combining Predictive Business Process Monitoring Techniques," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 45, no. 2, pp. 276-290, Feb. 2015

[15] Günesen, S.N., Şen, N., Yıldırım, N., Kaya, T. (2021). Customer Churn Prediction in FMCG Sector Using Machine Learning Applications. In: Mercier-Laurent, E., Kayalica, M.Ö., Owoc, M.L. (eds) Artificial Intelligence for Knowledge Management. AI4KM 2021. IFIP Advances in Information and Communication Technology, vol 614. Springer, Cham.

[16] VIJAYKUMAR UMMADISETTY (2017). Online Retail Data Set
https://www.kaggle.com/datasets/vijayuv/onlineretail

[17] Abdolreza Mosaddegh, Amir Albadvi, Mohammad Mehdi Sepehri, Babak Teimourpour, Dynamics of customer segments: A predictor of customer lifetime value, Expert Systems with Applications,Volume 172, 2021,114606,ISSN 0957-4174.

[18] Tao, Fu & Wang, Xindi. (2020). Rental Customer Segmentation Based on Length, Recency, Frequency, Average-Monetary and Satisfaction Value Model and Cluster Analysis.

[19] S. Wu, W. -C. Yau, T. -S. Ong and S. -C. Chong, "Integrated Churn Prediction and Customer Segmentation Framework for Telco Business," in IEEE Access, vol. 9, pp. 62118-62136, 2021

[20] R. Punhani, V. P. S. Arora, S. Sabitha and V. Kumar Shukla, "Application of Clustering Algorithm for Effective Customer Segmentation in E-Commerce," 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, 2021, pp. 149-154