© April 2026 | IJIRT | Volume 12 Issue 11 | ISSN: 2349-6002

Consumer Behavior Segmentation using Variational Autoencoder on RFM Data

Samtham Hari Priya Devi[1], Polisetty Durga Harshitha[2], Chinthapatla Siddhartha[3], Lakkamaraju Sai Sohith Varma[4], K. V. D. Kiran[5]

[1,2,3,4,5] *Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, 522502, Andhra Pradesh, India*

***Abstract*—Consumer Behavior analysis plays an essential part in modern e-commerce systems as it ensures the preservation of the target audience and the right marketing campaigns. At present, customer segments are created on the basis of the RFM model, which consists of Recency, Frequency, and Monetary value, but it does not take into account the complexities of consumer behavior. The purpose of this paper is to propose a technique for consumer segmentation via the application of the Variational Autoencoder algorithm to build RFM embedding space. First, it is supposed to train the VAE algorithm for obtaining embeddings based on RFM values. For embedding generation, one may project consumer RFM values on the manifold latent space. Besides, the reconstruction error resulting from the training process can be utilized as a criterion for measuring embedding quality. Clustering can be applied for classifying consumers according to their behavior into Loyal, Potential, Fresh, and At-Risk segments. Finally, a dashboard is created to perform clustering through visual analytics. As follows from the results of the experiment, the technique proposed for consumer segmentation based on the generation of embeddings via RFM proved its efficiency.**

***Index Terms*—Consumer Behavior Analytics, Deep Learning, E-Commerce, RFM Model, Variational Autoencoder, Customer Segmentation.**

I. INTRODUCTION

With changing trends in the world of the internet, a great deal of data is collected through transactions that can be applied in studying the behavior of consumers. The behavior of consumers at different levels such as purchase or browsing level indicates how they would respond to the website. Studying the behavior of consumers helps in enhancing marketing strategies and designing products for them. RFM is one of the commonest methods used to help in understanding behavioral patterns of consumers who adopt certain behavior for marketers to conduct customer segmentation. RFM technique helps in grouping customers' behavior according to the following three factors: recency (R), frequency (F) and monetary value (M). Recency is the period of time since the consumer made a transaction, frequency is the number of times the consumer made the transaction and monetary value is the total amount spent by the consumer when making purchases. This method makes it easy for customers' segmentation based on behavior.

Though the above-mentioned method seems simple, the traditional methods such as statistical and rule-based techniques may not be able to comprehend the complexity of the behavioral pattern of customers due to its non-linear nature. With a lot of algorithms in deep learning having been proposed, it becomes possible for scientists to propose the development of such a solution that would help identify underlying structures in huge datasets. Among the most effective techniques of representation learning in the unsupervised way are variational autoencoders. They make it possible to encode information from high-dimensional spaces into low-dimensional spaces while retaining all information related to interdependence of the given pieces of data. In this way, the machine learning algorithm learns the underlying distribution of the data and is able to detect various patterns within them. In this paper, we provide an overview of the approach we followed when applying VAEs for behavioral analysis. The task that we have is encoding of the given input RFM vector by utilizing the suggested technology. It will help us reveal some common features between those customers. As a result, latent behavioral profiles of the customers will be generated. Using those behavioral customer profiles, it will be possible for us to categorize these customers into four different classes: loyal customers, potential buyers, new customers, and risky customers. In addition, the validity of customer behavior can be measured using reconstruction errors of our VAEs. For ensuring that the information collected from them is meaningful, the proposed methodology involves the usage of an analytics dashboard that uses the technology of artificial intelligence for developing graphics in segmentations, behavior analysis, and KPIs. The implementation of the analytics dashboard proposed in the research will ensure that there is interaction in real time between the firms and the customer intelligence, alongside developing marketing strategies on the basis of customer evaluation.

II. LITERATURE REVIEW

Customer behavior analysis has developed into one of the primary fields of interest in the study conducted for e-commerce. The area of interest is very important in retaining customers and making profit. As the increase in e-commerce websites leads to increasing amounts of data from different firms, efficient tools are needed to analyze such datasets. Some of the ways that have been applied to customer segmentation are RFM segmentation among others. They have proven to be quite successful through various studies [3]. However, these ways work under some assumptions that do not allow for effective customer segmentation in cases where such assumptions are not met.

Some machine learning techniques can help overcome the problem. Clustering approaches have been widely applied for customer segmentation on the basis of customer behavior within businesses. Among the many ways, some of the approaches that have been applied are k-means algorithm [10] and DBSCAN [11]. In comparison to other ways, such approaches are more flexible and effective. Flexibility and efficiency depend largely on inputs and distances. The impact of data mining techniques on the extraction of useful patterns from large databases cannot be ignored. The framework presented in [7], [13], [27], and [28] has provided various methods for the discovery of hidden relationships and trends in the database of the customers. This has helped businesses gain more insight into the nature of the customers, thus making wise business decisions. In addition, the concept of association rule mining [18] provides businesses with the opportunity to discover various relationships concerning the products and the nature of the customers.

The importance of patterns in the database should not be undermined. There is the use of different methods in the model described by [7], [13], [27], and [28]. These techniques have been very important for businesses in understanding how the clients relate with trends in the customer databases. It has also been used in assisting organizations in decision-making through helping them understand how their customers behave. Besides, another important technique of analyzing customers is association rules mining [18]. With the development of portals for e-commerce business, big data analysis has come into existence. It is important to look into the literature review of the role played by intelligence systems in big data analysis based on client behavior [4] and [21]. This kind of data has been very essential in decision-making. Another aspect is that of customer data analytics [17]. The deep learning method has also been employed to improve the process of complex data analysis. With regards to [2] and [24], it is clear that the deep neural network method has been used to improve complex data analysis. When carrying out complex data analysis, the process of representation learning method has proven to be useful in solving problems of complex data analysis. With the use of the process of representation learning method, it becomes easy to determine the features of the complex data.

Moreover, LSTM-based neural network methods can be used in modeling complex data and the behavior of customers. Moreover, the use of neural networks can be employed in making predictions on the future behaviors of customers. The factorization model can be used in analyzing the relationship between the users and products. This is where some of the algorithmic models to be used include recommendation system algorithms such as collaborative filtering [25], matrix factorization [5], and neural collaborative filtering [6]. There does not seem to be any particular approach that helps in segmenting customers using this entire process. On the other hand, the application of VAE as indicated in [1] appears to be among the best ways of ensuring the entire process of learning takes place. The performance of this algorithm is measured based on how efficiently the entire process can convert all the information into a structural format. This process makes it one of the best ways of identifying patterns in the clients. This way, optimization of the deep learning algorithm will become possible as indicated in [22] and [23].

Another alternative solution to the above-presented issue can be a combination of machine learning and deep learning techniques, also known as hybrid learning. This concept has been considered by other researchers [19, 20], and the authors have shown that the integration of the two methods significantly improves efficiency. In this case, the benefits of both approaches are used, thus allowing identifying linear and nonlinear relationships. In addition, the importance of scalability is vital because of the increasing volume of generated data in the e-commerce system. As claimed in [29, 30], the suggested framework along with the optimal technique will ensure efficient functioning of the system. At the same time, it is important to mention that there are some drawbacks between the presented solution and currently available ones. While there are some instances of resolving the issue, it is still not enough since the solutions are not integrative. It is necessary to fill the gaps since both e-commerce system and VAE were considered in the proposed solution.

III. METHODOLOGY

The suggested methodology will be designed in such a way to predict consumer behavior concerning online purchasing through machine learning and variational autoencoders. These will be different procedures involved in coming up with the suggested methodology. Some of the processes include data acquisition, data pre-processing, feature extraction, latent feature extraction, clustering, and visualization. Through all of these procedures, one can identify latent consumer behavior along with consumer clustering.

A. Data Collection

The data source is the application that has been created to run as an entire e-commerce stack. Using this application, consumers will create their accounts, look for available items and review products based on the products purchased. However, on the other side, the shopkeepers will have the entire control over their items. As an administrator, the consumer behavior and revenues will be analyzed.

There are several transactions in the data source, including consumer identification, buying dates, transaction frequencies, and expenditures. This data source gives a hint about consumer behavior, which can be used to analyze other datasets.

B. Data Preprocessing

In order to enhance the quality of the data and increase the data integrity level, data preprocessing will be conducted before any analysis. This will involve getting rid of all unnecessary data points. At this stage, only necessary data points for the analysis of consumer behavior will be selected.

The dataset is filtered to ensure only valid customers with positive transaction frequency and amount are considered. This is necessary to ensure that the model is not trained on inappropriate customer information.

C. Feature Extraction using RFM Model

Customer’s purchase behavior is based on the following criteria:

- Recency (R): Days from the date of the last purchase made by the customer.
- Frequency (F): Number of purchases made by the customer.
- Monetary (M): Monetary value of purchases made by the customer.

RFM characteristics will be adjusted through the following method: g formula:

$$X\_scaled = (X - \mu) / \sigma$$

This step ensures that all features contribute equally during model training and prevents bias.

D. Latent Feature Learning using VAE

The Variational Auto Encoder (VAE) technique is applied to identifying the intrinsic pattern within the RFM feature set. The VAE framework includes two primary components, in this case, the Encoder paired with the Decoder. The encoder's key function is to transform the data into a a reduced dimensionality, while the main duty of the decoder is to reconstruct the data back to its original form.

The method to recognize the underlying variables is shown below:

$$Z = \mu + \sigma \epsilon$$

Using the above formula, complicated relationships among the customers.

E. Clustering

Latent variables were processed through K-Means clustering. This allows for the possibility of clustering customers depending on their behavior pattern. Customers may thus be segmented based on their cluster category (Loyal, Potential, Fresh, At-Risk, etc.).

F. Reliability Score

The reliability score shall be calculated as the measurement of the efficiency of prediction made by the system. The higher the score, the more accurate the segmentation process would turn out. In this case, the reliability score is dependent on the representation pattern of the system's customers.

Data has been retrieved from the full-stack e-commerce website, which is one of the real-life uses of the concept. Through this website, its users can register, purchase, and provide reviews about their purchases. On the other hand, the owner/administrator of the website shall be able to manage products sold and analyze sales data.

G. Visualization

These have been accomplished using the dashboard in regard to segmentation and KPIs.

Data has been obtained from transactional data, which includes identifying the customers, days they make their purchases, how often they purchase goods, and the amount of money they spend when purchasing. Such type of data can be very important in establishing how customers behave, hence behavioral data. In the first step, the behavioral data of the customer is transformed to RFM feature vectors. This implies that the data has been simplified as simple behavioral vectors. Recency entails the days elapsed since the customer made his/her last transaction. Frequency represents the number of transactions made by the customer during a certain period of time. Monetary is concerned with the total money spent by the customer. The behavioral vectors include these variables.

Customer Behavioral Vector:

$$X = (R, F, M)$$

Whereas R, F, and M are recency, frequency, and monetary variables respectively. Standardization of the variables has taken place.

The collected data will go through the data preprocessing stage. It is important for us to ensure that we will have high-quality data after collecting them. There will be some data cleaning where all the unnecessary information such as noise and incorrect data will be removed from the collected data. Normalization will take place where all the input variables that will be used in the analysis will be converted into the same scale. Lastly, the patient data will be anonymized.

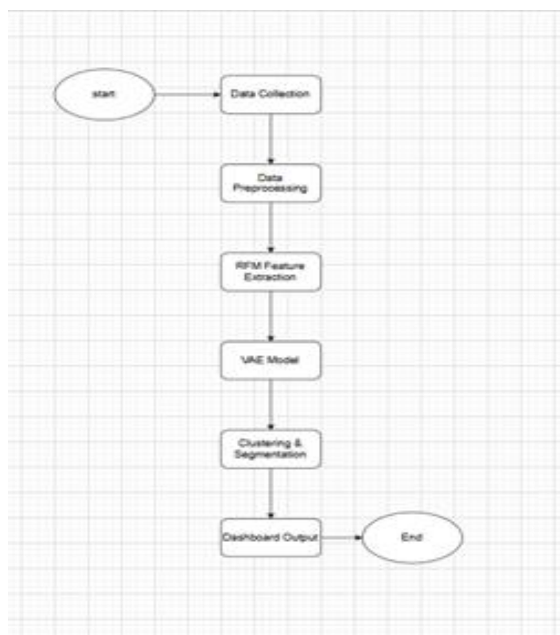

Fig (1): Flowchart representation of customer behavior prediction using RFM and VAE model.

IV. RESULT

Proposed E-commerce Analytics Model can be used for predicting behavior of the customers with the help of transactional data. With the help of RFM attributes combined with VAE approach, it becomes possible for the proposed system to recognize trends and classify customers into loyal, potential, fresh, and at-risk customers. Therefore, proposed system would definitely prove to be beneficial for organizations in enhancing their marketing, retention, and revenue performance.

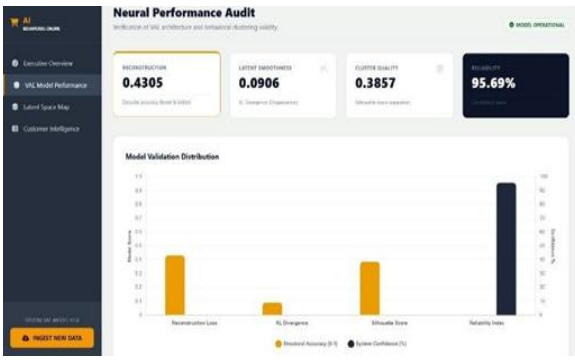

Fig (2): Neural Performance Dashboard – AI Behavioral Engine

Performance evaluation of proposed model based on different parameter values has been shown in Fig. (2). For instance, 0.4305 reconstruction loss value indicates high efficiency in data representation. A KL Divergence value of 0.0906 indicates high structural efficiency in the feature space while Silhouette Score value of 0.3857 reveals high efficiency in segmentation of customers in clusters. Reliability of proposed system has been determined to be 95.69%.

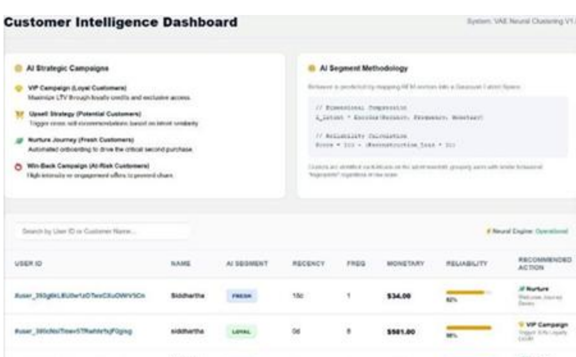

Fig (3): Customer Intelligence Dashboard – Segmentation Output.

Results obtained from the customer segmentation process have been illustrated in Fig. (3). Details regarding customer segmentation process including RFM, customer segments, and reliability of customer segments have been revealed in the figure. Classification of customers into different categories using proposed system has been done in order to illustrate the usage of proposed system in marketing activities.

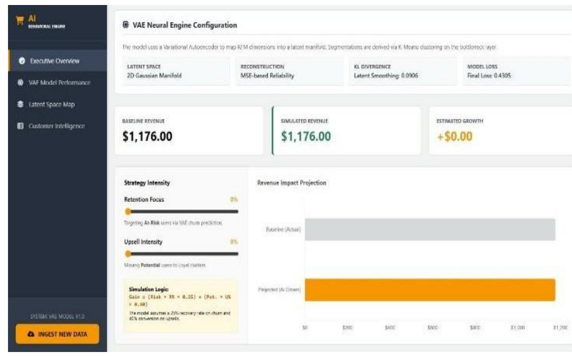

Fig (4): Executive Overview Dashboard – AI Behavioral Engine

Fig (4): The figure shows the dashboard for the predictive model that is to be used to predict consumer behavior.

The figure above shows the model for predicting the consumer behavior analysis using the effectiveness of the Variational Autoencoder Model. There are various assumptions of the factors to be used in the predictive model. These include the following:

V. CONCLUSION

The proposed recommendation framework can be defined as a practical approach to making predictions of consumer behavior in e-commerce sites through the use of the RFM method combined with variational autoencoders and clustering. Consumer behaviors within the e-commerce website will be analyzed to predict future consumer actions through hidden patterns in the transactions. Based on the proposed recommendation framework, consumers will be divided into various categories including loyal, potential, fresh, and risky consumers.

Reliability index included in the recommendation framework will be instrumental in ensuring the high level of reliability of the predictions generated using the framework. The inclusion of the dashboard for the web interface is considered as one of the most effective ways of evaluating the practicality of the recommendation framework. The recommended recommendation framework is considered as an efficient model to predict consumer behavior in e-commerce platforms.

VI. FUTURE ENHANCEMENT

In addition, the model can incorporate the behavior of consumers or other variables that might form part of the system to create prediction models that provide accurate results. Also, there could be incorporation of machine learning algorithms aimed at increasing efficiency of segmentation of the market segments. Last but not least, real time analysis would be incorporated as well for purposes of creating real-time decision-making abilities.

Looking towards the future improvement of this system, it is advisable that the recommendation engine be incorporated into this system. This module will play an important role of recommending strategies for marketing based on the findings of the segmentations performed.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv preprint arXiv:1312.6114, Dec. 2013. Presented at the 2nd Int. Conf. Learning Representations (ICLR), Banff, AB, Canada.

[2] Goodfellow, Y. Bengio, and A. Courville, Deep Learning, 1st ed. Cambridge, MA, USA: MIT Press, 2016.

[3] P. S. Fader, B. G. S. Hardie, and K. L. Lee, "RFM and CLV: Using Iso-Value Curves for Customer Base Analysis," J. Marketing Res., vol. 42, no. 4, pp. 415–430, Nov. 2005, doi: 10.1509/jmkr.2005.42.4.415.

[4] S. Akter and S. F. Wamba, "Big Data Analytics in E-Commerce: A Systematic Review and Agenda for Future Research," Electron. Markets, vol. 26, no. 2, pp. 173–194, Jun. 2016, doi: 10.1007/s12525-016-0219-0.

[5] Y. Koren, R. Bell, and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," Computer, vol. 42, no. 8, pp. 30–37, Aug. 2009, doi: 10.1109/MC.2009.263.

[6] X. He et al., "Neural Collaborative Filtering," in Proc. 26th Int. World Wide Web Conf. (WWW), Perth, Australia, Apr. 2017, pp. 173–182, doi: 10.1145/3038912.3052569.

[7] Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[8] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[9] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.

[10] MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, Berkeley, CA, USA, 1967, pp. 281–297.

[11] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD)*, Portland, OR, USA, Aug. 1996, pp. 226–231.

[12] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013, doi: 10.1109/TPAMI.2013.50.

[13] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[14] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[15] S. Rendle, "Factorization Machines," in *Proc. 10th IEEE Int. Conf. Data Mining (ICDM)*, Sydney, Australia, Dec. 2010, pp. 995–1000, doi: 10.1109/ICDM.2010.127.

[16] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005, doi: 10.1109/TKDE.2005.99.

[17] H. H. Olsson and J. Bosch, "Customer Data Analytics for Product Development: A Case Study," *J. Syst. Softw.*, vol. 137, pp. 229–239, Mar. 2018, doi: 10.1016/j.jss.2017.12.013.

[18] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in *Proc. 20th Int.*

Conf. Very Large Data Bases (VLDB), Santiago, Chile, Sep. 1994, pp. 487–499.
[19] S. Sahoo, S. Kumar, and R. K. Singh, "Customer Segmentation Using Machine Learning Techniques," *Int. J. Eng. Res. Technol.*, vol. 9, no. 5, pp. 1–5, May 2020.
[20] S. Mohanty and S. Chatterjee, "Customer Purchase Behavior Analysis Using Data Mining Techniques," *Int. J. Comput. Appl.*, vol. 182, no. 10, pp. 1–5, Jul. 2018, doi: 10.5120/ijca2018917398.
[21] H. Chen, R. H. Chiang, and V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quart.*, vol. 36, no. 4, pp. 1165–1188, Dec. 2012.
[22] A. Ngiam *et al.*, "Multimodal Deep Learning," in *Proc. 28th Int. Conf. Machine Learning (ICML)*, Bellevue, WA, USA, Jun. 2011, pp. 689–696.
[23] S. Ruder, "An Overview of Gradient Descent Optimization Algorithms," *arXiv preprint* arXiv:1609.04747, Sep. 2016.
[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[25] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," in *Proc. 10th Int. World Wide Web Conf. (WWW)*, Hong Kong, May 2001, pp. 285–295, doi: 10.1145/371920.372071.
[26] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2010.
[27] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, 2nd ed. Hoboken, NJ, USA: Wiley, 2011.
[28] C. C. Aggarwal, *Data Mining: The Textbook*. Cham, Switzerland: Springer, 2015, doi: 10.1007/978-3-319-14142-8.
[29] J. Brownlee, *Machine Learning Mastery with Python*. Melbourne, Australia: Machine Learning Mastery, 2016.
[30] L. Breiman, "Bagging Predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.