Analyze and Predict Potential Customers Based on Customer Clustering

Ta Cong Binh, Ngo Chi Trung, Hoang Anh Tu, and Phan Duy Hung(✉)

FPT University, Hanoi, Vietnam
{binhtche186122,trungnche180620,tuhahe182038}@fpt.edu.vn,
hungpd2@fe.edu.vn

Abstract. Customer segmentation plays a vital role in designing effective marketing campaigns to drive business growth and maximize revenue. In this study, we apply Principal Component Analysis and Autoencoder to extract meaningful features, reducing data complexity while preserving essential information. To enhance clustering accuracy, we first utilize DBSCAN to detect and remove noise before employing the K-Means algorithm for customer segmentation. This approach helps identify distinct customer groups, providing valuable insights into consumer behavior. Furthermore, we analyze potential customers who have the possibility of being upgraded from loyal customers to VIP status. Finally, we implement a Random Forest Classification model to predict potential new customers based on fundamental customer information, enabling businesses to develop proactive strategies for customer acquisition and retention.

Keywords: Customer Clustering · Predicting Potential Customers · PCA · Autoencoder · K-Means Clustering · DBSCAN · Random Forest

1 Introduction

In the era of increasingly strong digital transformation, online shopping has greatly changed the shopping behavior of consumers. Businesses are increasingly focusing on consumer purchasing behavior to meet the growing trend of online shopping to improve service provision, enhance customer satisfaction, encourage customers to return to purchase and ultimately increase profits. Therefore, customer segmentation and prediction of potential customers are extremely important and to address the limitations in data mining on consumers' online shopping behavior, this paper aims to cluster and predict consumer behavior using machine learning techniques, thereby supporting companies in resource allocation and decision making.

Customer segmentation and predictive modeling have been extensively studied in data-driven marketing and customer management. Traditional segmentation methods have used techniques such as RFM (Recency, Frequency, Monetary) analysis [1] to segment customers based on their purchasing behavior. While the RFM has the advantage of simplicity and easy cheesiness, it does not offer the flexibility to accommodate complex behavioral patterns in high-dimensional datasets. K-Means clustering [2, 3]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15888, pp. 75–87, 2026.
https://doi.org/10.1007/978-3-031-97596-7_5

has been used to cluster customers according to their behavior characteristics. K-Means does however not always fit the distribution of data in the real world, as it is presuming spherical clusters and a fixed number of clusters. To overcome the limitations of K-Means, consumer segmentation works found global optimality by using hierarchical clustering [4, 5] and density-based clustering (DBSCAN) [6]. DBSCAN is good for noise filtering and allows us to find clusters of arbitrary shape. Hierarchical clustering allows us a more flexible cluster architecture. However, due to their computational complexity, these approaches often struggle when applied to large-scale consumer databases. Predictive analytics is also useful in identifying potential customers who are likely to do repeat purchases in the future. Initial works used a combination of Bayesian and logistic regression models to estimate the probability of customer conversion [7], such models are, however, still limited when attempting to account for consumer data exhibiting non-linear correlations. To get over this limitation, machine learning methods like Random Forest, Gradient Boosting Machines (GBM), and Neural Networks have been extensively employed and have produced classification results that are more accurate [7, 8]. A significant quantity of both organized and unstructured data may be processed by these models, which also make it possible to identify special characteristics in consumer behavior. For example, XGBoost has been deployed to predict Customer Lifetime Value (CLV) which helps to improve the effectiveness of marketing strategies [9]. Customer segmentation and predictive modeling are important aspects of data-driven marketing. The use of RFM, K-means, DBSCAN models are traditional methods but they are quite simple and ineffective for complex and non-linear data or very large scale. Models like Random Forest, Boosted Gradient Machine (GBM) and Artificial Neural Networks perform better than traditional methods in handling large, complex data in terms of speed and accuracy.

The objective of this paper is to present a robust and suitable approach for customer segmentation and lead prediction by combining powerful algorithms into a common framework. This approach helps to provide a performance evaluation of the model which helps to optimize the algorithm selection process and improve computational efficiency. Before segmenting and predicting the leads, we use Principal Component Analysis (PCA) techniques and if the data is non-linear, dimensionality reduction using Autoencoder will help to reduce dimensionality more effectively than PCA. DBSCAN is used to detect customers with unusual behavior. Finally, K-means algorithm is applied to identify customer segments in the dataset and analyzed for each customer group to identify the behavioral characteristics of each group to come up with a good solution. Integrating this approach not only provides deeper insights into customer segmentation but also optimizes future predictions. For potential customer prediction, the Random Forest model is implemented to identify customers who are likely to become loyal or high-value customers for the business. By leveraging customer data, we can recommend solutions, helping businesses optimize strategies and allocate resources more effectively.

2 Methodology

2.1 Dataset

The data for the research paper was taken from Kaggle [10], which includes 2240 customers and 29 different characteristics collected from a shopping center. This database includes essential information about customers, and the main features of the data include: Personal information, Financial data, Family information, Purchase history, Shopping channels, Interaction with marketing campaigns, Web access behavior, feedback and complaints.

The statistical data includes the following information:

- The average income of 2216 people is 52,247.25 with a standard deviation of 25,173.08, ranging from 1,730 to 666,666.
- The average number of children and teenagers per household is 0.44 and 0.51, respectively.
- The average recent time is 49.11 days with a standard deviation of 28.96 days.
- The average amount spent on wine, fruit, meat products, fish products, sweet products, and gold products is 303.94, 26.30, 166.95, 37.53, 27.06, and 44.02, respectively.
- The average number of online, catalog, and in-store purchases is 4.08, 2.66, and 5.79, respectively.
- The average number of website visits per month is 5.32 with a standard deviation of 2.43.

2.2 Data Preprocessing

Before performing data analysis, preprocessing is a crucial step to ensure the accuracy and stability of the model. This process begins with data cleaning, which includes handling missing values by removing them or imputing them using mean or median values. After cleaning, data needs to be normalized to ensure that features are on the same scale, improving model performance. What makes this dataset difficult to apply customer clustering techniques to is that it is quite noisy (Table 1) and we cannot completely remove it, so RobustScaler is applied to this noisy dataset to reduce the impact of outliers. Additionally, Feature Engineering plays a vital role in creating and selecting useful features to reduce redundant variables and prevent overfitting. To reduce dimensionality while retaining essential information, methods such as Principal Component Analysis (PCA) and Autoencoder are used. PCA identifies principal components that carry the most information, whereas Autoencoder, with its ability to learn nonlinear representations, reduces dimensionality while allowing data to be reconstructed as accurately as possible (Fig. 1).

2.3 Model Selection

To perform customer segmentation, the DBSCAN and K-Means models are applied to group customers based on similar characteristics. First, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is used to remove customers with unusual purchasing behaviors, cleaning the dataset by filtering out noise (outliers). The cleaned

Table 1: Number of exceptions detected by feature (1285 outliers / 2240 values)

Feature	Outlier Count	Feature	Outlier Count
Marital_Status	0	NumWebPurchases	3
Income	8	NumCatalogPurchases	23
Kidhome	0	NumStorePurchases	0
Teenhome	0	NumWebVisitsMonth	8
Recency	0	Family_Size	31
MntWines	35	MntFishProducts	222
MntFruits	246	MntSweetProducts	246
MntMeatProducts	174	MntGoldProds	205
NumDealsPurchases	84		

dataset is then clustered using the K-Means approach, where the optimal number of clusters is determined using methods like the Silhouette Score and the Elbow method. In addition to improving consumer segmentation efficacy, the combination of the two models helps in precisely identifying unusual behaviors.

The next step after clustering is to create a predictive model that helps identify potential customers. Based on the distance measurements, we conduct an analysis and set a threshold to evaluate loyal consumers and promote them to potential VIP status when they are close to qualifying. This not only helps to evaluate customers better but also helps businesses to develop strategies to promote them to become VIP in the future. To accurately and reliably classify new potential customers based on their purchasing characteristics, we finally deploy the Random Forest algorithm.

2.4 Evaluation

The Elbow method will look at the proportion of variance explained as a function of the number of clusters (k) to find the ideal number of clusters.The best number of clusters can be determined from WCSS (within-cluster sums of squares), the point at which the WCSS slows down significantly, forming an “elbow”. This point is the optimal number of clusters. Silhouette Score measures how well data points are clustered based on its cohesion inside the assigned cluster and separation from other clusters. The Silhouette Score, which ranges from -1 to 1, where near 1 means points were clustered well, near 0 that points are on the line between two natural clusters, and negative value means the points were clustered incorrectly.

To assess the quality of predictive models, several key metrics are used to measure accuracy. Precision calculates the percentage of predicted positive samples that are truly positive, while Recall measures the percentage of actual positive samples correctly identified by the model. Additionally, ROC-AUC provides an overall evaluation of classification performance. The ROC Curve visualizes the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR), while AUC (Area Under Curve)

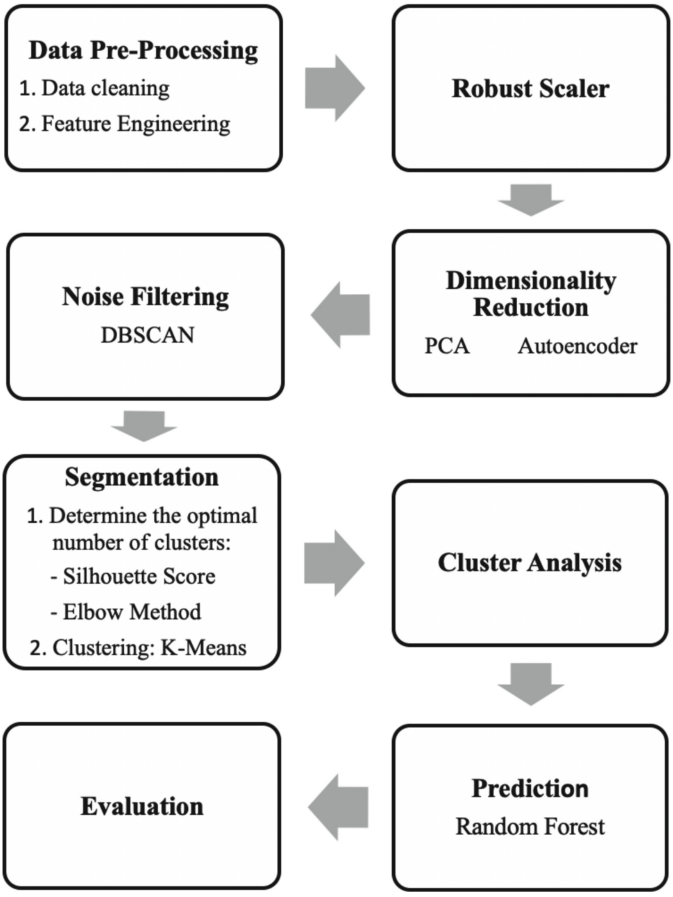

Fig. 1. Flow diagram of the study

quantifies the area under the ROC curve, ranging from 0 to 1. A higher AUC indicates a more effective classification model.

3 Implementation and Results Analysis

3.1 Dimensionality Reduction and Clustering

Before proceeding with customer clustering and classification, two dimensionality reduction techniques, PCA and Autoencoder, are applied to determine which method yields the most optimal results. One of the key issues to address is identifying the appropriate number of dimensions for these algorithms to function effectively before performing clustering. To achieve this, we evaluate the optimal number of components for PCA using the explained variance ratio (Fig. 2) and determine the appropriate latent space dimension for the Autoencoder based on reconstruction error minimization (Fig. 3). We decided to select 7 dimensions for both techniques. As shown in Fig. 2, retaining 7

principal components in PCA preserves approximately 83.5% of the original data variance, which is an acceptable level in practical applications where reducing complexity is necessary while maintaining essential information. For the Autoencoder, as illustrated in Fig. 3, encoding the data into 7 dimensions results in a relatively low reconstruction loss (approximately 0.51), reflecting the model's ability to effectively learn meaningful representations.

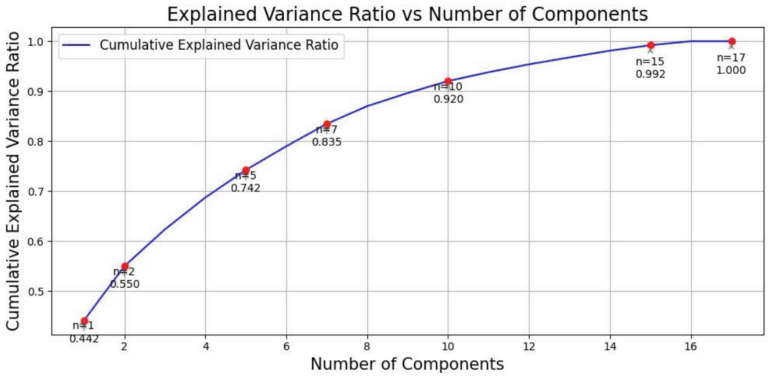

Fig. 2. Explained variance ratio of PCA

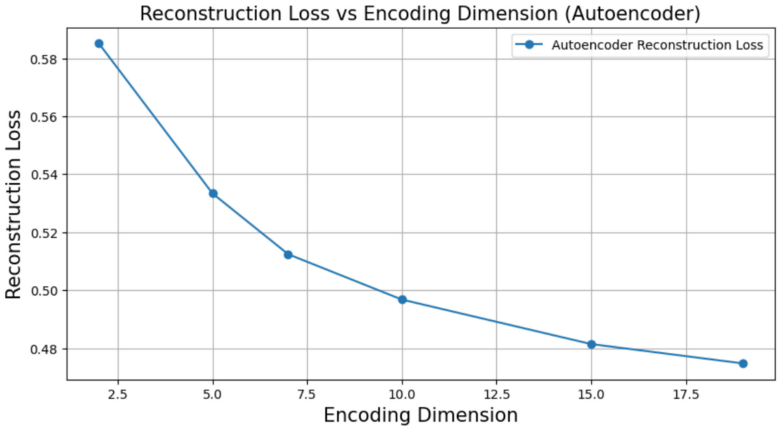

Fig. 3. Reconstruction Loss vs Encoding Dimension (Autoencoder)

Next, to enhance the performance of the K-Means algorithm, DBSCAN is applied to identify and remove unwanted outliers from the dataset. The effectiveness of DBSCAN relies heavily on two key parameters: the neighborhood radius (ε) and the minimum number of points (MinPts). Proper selection of these parameters significantly improves the accuracy of noise filtering, ensuring that the remaining data is well-suited for clustering. To optimize ε, the Silhouette Score can be used as a guiding metric. The optimal ε value is determined when the Silhouette Score reaches its peak, indicating that the

clustering structure is well-defined, with clear separation between clusters and minimal noise. This approach ensures that DBSCAN effectively refines the dataset before applying K-Means clustering. Figure 4 shows the variation of Silhouette Score with ε values ranging from 0.5 to 6.0, applied on two different data representations: data after dimensionality reduction by PCA and data encoded by autoencoder. On PCA data, Silhouette Score gradually increases and peaks in the ε range from 4.0 to 5.5 (maximum value is approximately 0.55), then decreases sharply. Therefore, $\varepsilon = 4.5$ is chosen as the optimal value for this case. For encoded data, Silhouette Score improves significantly after $\varepsilon = 3.0$ and stabilizes around the highest level (~0.38) at $\varepsilon = 5.5$–6.0, so $\varepsilon = 5$ is chosen.

Additionally, to assess whether DBSCAN effectively filters out outliers, the k-distance plot can be utilized to observe the difference in distance between the outlier group and the main cluster, helping to establish an appropriate threshold for noise removal (Fig. 5).

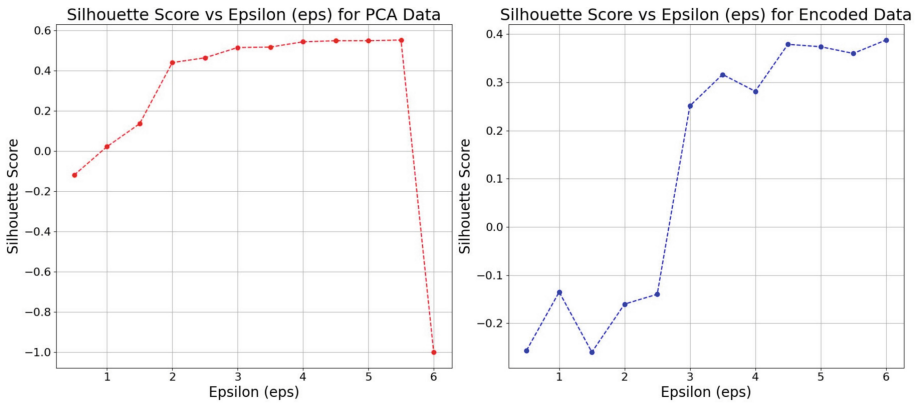

Fig. 4. Silhouette Score vs Epsilon (eps)

The K-Means algorithm is employed to cluster the data corresponding to each reduction method (Fig. 6) and evaluate the results based on features such as salary, total purchase amount, and total number of products purchased (Fig. 7, 8, 9). Finally, the obtained clusters are used in the customer classification task to identify VIP customers, potential customers, or those at risk of churning.

3.2 Clusters Analysis

When comparing two dimensionality reduction methods before applying K-Means, Autoencoder outperforms PCA based on three evaluation criteria: Silhouette Score, Inertia (WCSS), and Davies-Bouldin Index (DBI).

For the Silhouette Score, the chart shows that Autoencoder achieves a higher score than PCA when the number of clusters is 3. This indicates that Autoencoder enables K-Means to cluster more effectively, with greater separation between clusters and less overlap. Both methods show a decreasing trend in the Silhouette Score as the number of

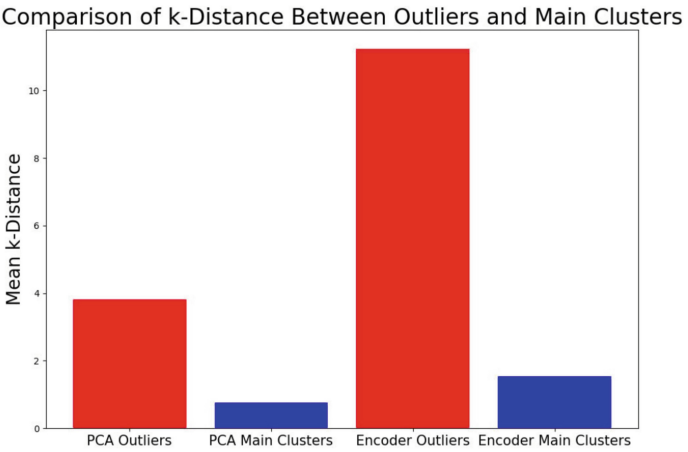

Fig. 5. Comparison of k-Distance Between Outliers and Main Clusters

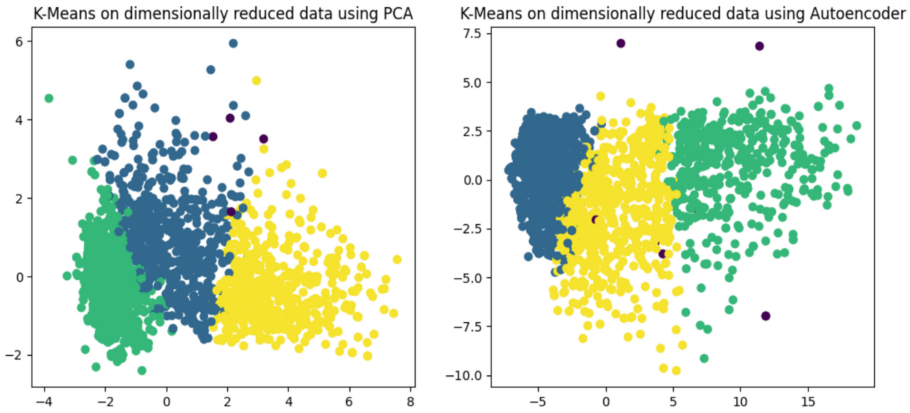

Fig. 6. Visualization of Customer Clusters

clusters increases, but Autoencoder consistently maintains a higher score in most cases, demonstrating its ability to enhance cluster separation.

Regarding Inertia (WCSS), the value for Autoencoder (42,740.07) is significantly higher than that for PCA (11,422.85). Inertia measures the total squared distance from each data point to its nearest cluster center, and the higher value obtained with Autoencoder suggests that the reduced-dimensional data retains better cluster separation, leading to larger distances between points and their cluster centers. This implies that Autoencoder preserves more important features for K-Means rather than simply compressing the data linearly, as PCA does.

Evaluated with the Davies-Bouldin Index (DBI), Autoencoder again demonstrates superiority with a lower value (1.2372) compared to PCA (1.4087). A lower DBI indicates more clearly separated clusters, meaning that Autoencoder enhances K-Means clustering efficiency by preserving the nonlinear structure of the data.

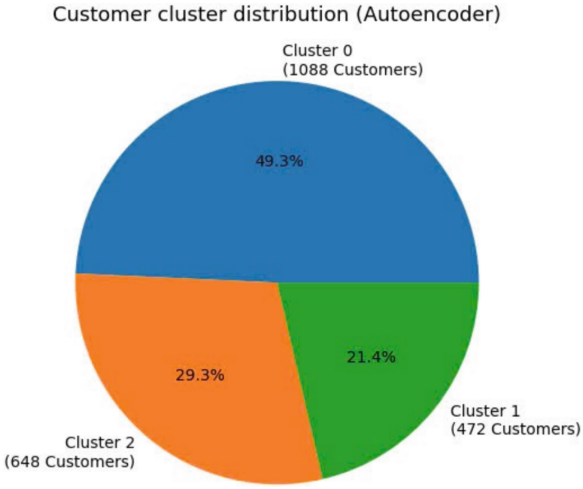

Fig. 7. Number of customers per cluster

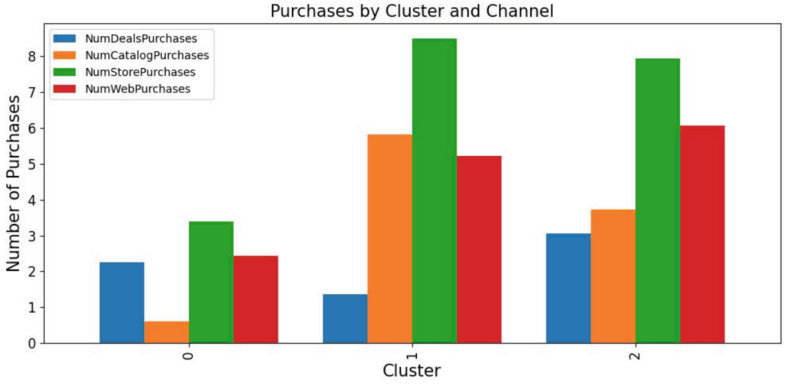

Fig. 8. Number of purchases per cluster

In conclusion, Autoencoder is a better dimensionality reduction method than PCA before applying K-Means. While PCA relies only on linear transformations, which may lose important information in complex data, Autoencoder can learn nonlinear features, maintaining cluster separability. Therefore, if the data has a nonlinear structure, Autoencoder is the preferred choice for improving K-Means clustering quality.

3.3 Potential Customers

In this study, we propose a method to identify potential VIP customers based on important characteristics that are similar between two groups of VIP and loyal customers by dynamic distance analysis. Specifically, the Loyal customer group is defined as those whose purchasing behavior closely resembles VIP customers but has not yet reached that

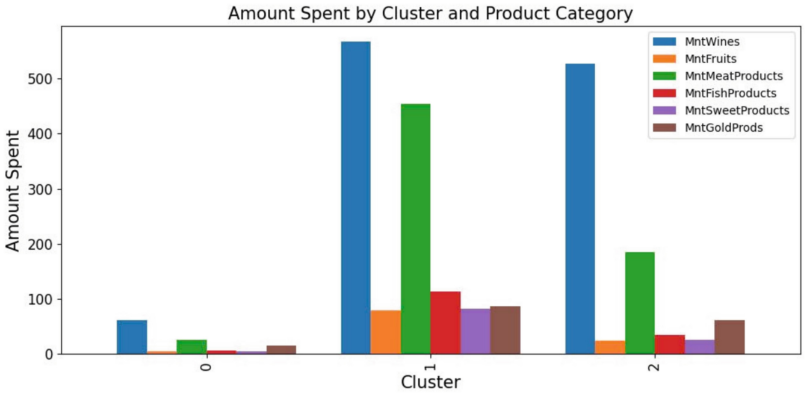

Fig. 9. Amount spent per cluster

level. To identify customers with the potential for status upgrade, we apply Min-Max Scaling to normalize customer features to a common range from 0 to 1, ensuring fairness and accuracy in comparisons. Next, we use Euclidean distance to measure the similarity between customers in the Loyal cluster and the centroid of the VIP cluster. This distance helps us determine which customer's behavior level is closest to the VIP group and from there we can determine the dynamic threshold to help classify the group of customers with the most potential to become VIP compared to the majority of customers in the loyalty group, then we need to calculate the average value (μ) and standard deviation (σ) of the distance from the loyal customer to the VIP center to determine this dynamic threshold. The dynamic threshold is defined as:

Threshold = mean (μ) – k. std (σ).

where k is an adjustment factor that helps us determine the number of corresponding customers we are interested in. Customers in the loyal group whose distance to the center of the VIP group is smaller than the distance to this threshold are identified as potential VIP customers.

The special feature of this method is that it can automatically adjust to new changes in data in the future and provides high flexibility. These results help businesses identify customers who need care, thereby optimizing customer care strategies and increasing VIP conversion rates.

Analyze the components of the chart (Fig. 10):

- X-Axis (Composite Behavior): Some customer behaviors after collection and dimension reduction for visualization purposes ('Income', 'Recency', 'Total Amounts', 'Total Purchase', 'Family_Size').
- Y-Axis (Distance to VIP): The distance of customers from the VIP group, which may be calculated based on a clustering model or a similar metric.
- Blue Dots (Cluster 2 Customers): Customers belonging to the second cluster (possibly loyal or regular customers).
- Red Dots (Potential VIP Customers): Customers with the potential to become VIP, meaning their distance to the VIP threshold is small.

- Green Dotted Line (Vip potential threshold): The distance threshold, below which customers have the potential to become VIP
- Black Dotted Line (Vip Threshold): This is the real VIP threshold when there are customers who reach this levels or higher then the customer has the ability to become a VIP in the future.

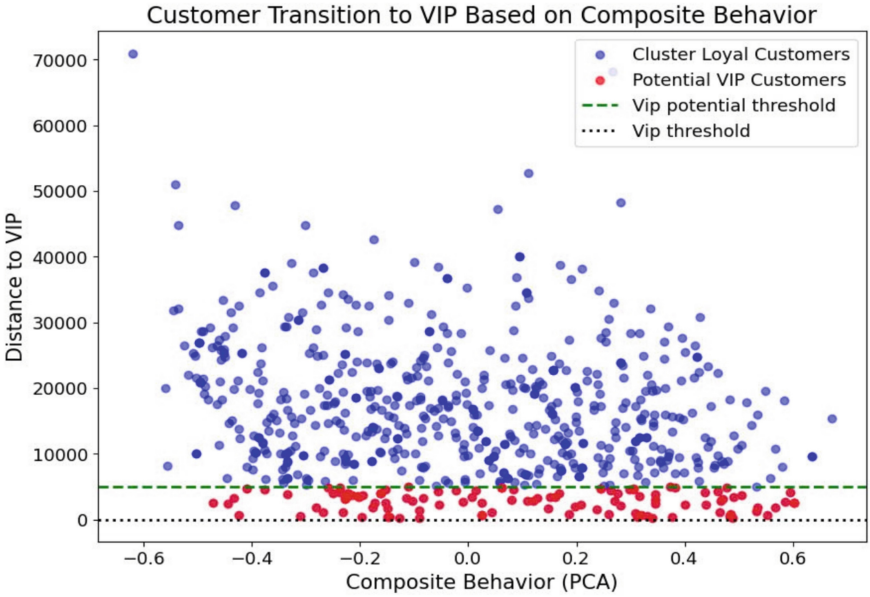

Fig. 10. Conversion rate among loyal customers with the potential to become

VIPAfter analyzing the data and identifying the group of customers with the potential to become VIPs, we calculated that the proportion of loyal customers with the potential to be upgraded to VIP status is 16.69%. This result indicates that a significant portion of loyal customers has the potential to be upgraded to VIP if approached and nurtured appropriately. Optimizing customer retention strategies based on this insight can help businesses enhance their customer engagement and increase Customer Lifetime Value (CLV).

Finally, we apply the Random Forest Classification model to predict which new customers have the potential to become loyal customers or VIP customers. New customers have identifiable input information such as income (Income), number of children in the household (Kidhome), family size (Family_Size), marital status (Marital_Status), and the amount spent on a single purchase (Amount_1_Time). First-time shoppers will be classified into the normally customer group based on their total spending and total quantity of items purchased. Then, we compare their spending characteristics with those of existing customer groups to determine which group they resemble most closely. If their spending behavior aligns with that of the VIP or loyal customer groups, they will be labeled as potential new customers. This approach allows us to early identify high-value

customers, enabling the deployment of tailored marketing and customer care strategies to maximize their lifetime value (Fig. 11).

```
Customer Data:
     Income  Kidhome  Family_Size Marital_Status  Amount_1_Time
0  69245.0        2            3        Married           40.0
Model predictions: Potential
```

Fig. 11. Example of predicting new customers as potential

4 Conclusion and Future Works

This paper presents an optimized approach for analyzing and predicting customer segmentation behavior using machine learning techniques and analytical methods. By applying these techniques, we can help businesses better understand customer behavior based on shopping characteristics, helping them predict customer segments that are likely to convert and have more appropriate strategies. A key aspect of this study is the comparison of dimensionality reduction and noise reduction techniques to enhance clustering model performance. In this project, we found that Autoencoder performed more effectively than PCA on nonlinear datasets. By improving data preprocessing, we aim to achieve more accurate and meaningful segmentation results. The use of DBSCAN for noise filtering also enhances the clustering performance of PCA. Furthermore, this study explores customer cluster analysis to identify potential customers within the loyal customer who may be upgraded to VIP customer. Finally, we built a predictive model that helps predict which future new customers are potential customers that businesses should pay attention to and from there suggest more specific strategies for their businesses. The work can be applied when dealing with time-series data [11, 12], market segmentation [13, 14], etc.

References

1. Hosseini, M., Shabxani, M.: New approach to customer segmentation based on changes in customer value. J. Market. Anal. **3**(3), 110–121 (2015). https://doi.org/10.1057/jma.2015.10
2. Khan, R.H., Dofadar, D.F., Alam, M.G.R.: Explainable customer segmentation using K-means clustering. In: Proceedings of the IEEE International Conference on Data Science and Business Analytics (2021). https://doi.org/10.1109/UEMCON53757.2021.9666609
3. Aslantaş, G., Gençgül, M., Rumelli, M., Özsaraç, M., Bakırlı, G.: Customer segmentation using K-Means clustering algorithm and RFM model. In: Proceedings of the International Conference on Artificial Intelligence and Data Science, 2023, vol. 25, no. 74, pp. 491–503 (2023). https://doi.org/10.21205/deufmd.2023257418
4. Afzal, A., et al.: Customer segmentation using hierarchical clustering. In: Proceedings of the International Conference on Machine Learning and Data Analytics (2024). https://doi.org/10.1109/I2CT61223.2024.10543349

5. Abdulhafedh, A.: Incorporating K-means, hierarchical clustering and PCA in customer segmentation. In: Proceedings of the IEEE International Conference on Big Data and Analytics (2021). https://doi.org/10.12691/jcd-3-1-3
6. John, J.M., Shobayo, O., Ogunleye, B.: An Exploration of clustering algorithms for customer segmentation in the UK retail market. Analytics. **2**(4), 809–823 (2023). https://doi.org/10.3390/analytics2040042
7. Balaji, S., Srivatsa, S.K.: Customer segmentation for decision support using clustering and association rule-based approaches. In: Proceedings of the International Conference on Computer Science and Information Technology (2012). https://www.ijcset.com/docs/IJCSET12-03-11-019.pdf
8. Tanhaei, H.G., Boozary, P., Sheykhan, S., Rabiee, M., Rahmani, F., Hosseini, I.: Predictive analytics in customer behavior: anticipating trends and preferences. In: Proceedings of the IEEE Conference on Business Intelligence and Data Analytics (2024). https://doi.org/10.1016/j.rico.2024.100462
9. Abidar, L., Zaidouni, D., El Asri, I., Ennouaary, A.: Predicting customer segment changes to enhance customer retention: a case study for online retail using machine learning. In: Proceedings of the IEEE International Conference on Machine Learning and Applications (2023). https://doi.org/10.14569/IJACSA.2023.0140799
10. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis
11. Dat, D.Q., Hung, P.D.: Clustering of time-series balance history data streams using apache spark. In: Luo, Y. (eds) Cooperative Design, Visualization, and Engineering. CDVE 2020. LNCS, vol 12341. Springer, Cham (2020)
12. Dat, D.Q., Hung, P.D.: Improvement for time series clustering with the deep learning approach. In: Luo, Y. (eds) Cooperative Design, Visualization, and Engineering. CDVE 2021. LNCS, vol 12983. Springer, Cham (2021)
13. Phan, D.H., Do, Q.D.: Analysing effects of customer clustering for customer's account balance forecasting. In: Nguyen, N.T., Hoang, B.H., Huynh, C.P., Hwang, D., Trawiński, B., Vossen, G. (eds) Computational Collective Intelligence. ICCCI 2020. LNCS, vol 12496. Springer, Cham. (2020). https://doi.org/10.1007/978-3-030-63007-2_20
14. Phan Duy Hung, Nguyen Duc Ngoc, and Tran Duc Hanh.: K-means clustering using R a case study of market segmentation. In Proceedings of the 2019 5th International Conference on E-Business and Applications (ICEBA 2019). Association for Computing Machinery, New York, NY, USA, 100–104 (2019)