Customer profiling segmentation and sales prediction using machine learning

B. Mahitha[1], Y. Jaswanth[2], R. Sahasra[3], and B. Rohan[4]
Department of Computer science and engineering.
GITAM (Deemed to be) University
Hyderabad, India
Email:bobbamahitha@gmail.com

Abstract— Customer behaviour knowledge and forecasting sales trends accurately are the biggest challenges for organizations in today's competitive age. The current paper explores a machine learning-based approach of customer profiling, segmentation, and sales forecasting from customer reviews and historical sales data. The paper revolves around the use of clustering algorithms like K-Means and DBSCAN for effective customer segmentation, VADER-based sentiment analysis for customer review comprehension, and multiple regression models like Linear Regression, Decision Tree Regression, and Random Forest Regression for effective sales forecasting. Time-series forecasting models ARIMA and LSTM are also employed to predict sales trends over time. The system proposed not only identifies distinct customer segments but also predicts future sales with greater accuracy, allowing key insights for effective targeted marketing strategies and inventory management. Experimental results validate the effectiveness of using customer sentiment and profiling data in prediction models, which leads to more informed business decisions and optimal sales performance.

.

Index Terms— Customer behaviour, customer profiling, segmentation, sales forecasting, ARIMA, LSTM.

I. Introduction

In the contemporary commercial environments, knowledge of consumption patterns and forecasting of sales patterns is a prerequisite for competitiveness. With the evolution of digital user interfaces and web-based platforms, firms have been able to access extensive customer data, such as purchasing history, demographics, and user-generated content in the form of reviews and feedback. With machine learning (ML) platforms, this information offers the possibility of obtaining insight into customer opinion, segmenting markets more effectively, and predicting future sales trends. This study addresses the growing imperative for business organizations to integrate data-driven solutions that enhance customer interaction, segment markets more effectively, and enhance overall sales effectiveness.

Despite being exposed to large volumes of customer and sales data, most organizations fail to analyze the data. Traditional methods cannot identify complex customer patterns and make sound predictions regarding future sales. Challenges such as unstructured data, customer behavior heterogeneity, and evolving market dynamics hinder organizations from creating accurate customer profiles and reliable sales forecasts. Poor segmentation and inaccurate sales forecasts can lead to ineffective marketing campaigns, subpar customer experiences, and significant revenue losses. The current study attempts to overcome these challenges by designing a machine learning-based method for customer profiling, segmentation, sentiment analysis, and sales forecasting.

The specific objectives of the current study are:

1. To develop an efficient system for customer profiling and segmentation through clustering algorithms such as K-Means and DBSCAN.
2. To conduct sentiment analysis of customer reviews with VADER to analyze customer satisfaction and preferences.
3. To forecast future sales patterns using regression models (Linear Regression, Decision Tree Regression, Random Forest Regression) and time-series forecasting methodologies (ARIMA, LSTM).
4. To combine customer segmentation and sentiment analysis with sales forecasting models for better forecasting and targeted marketing campaigns.

This research focuses on using machine learning techniques to analyze customer data and forecast sales trends. It seeks to integrate customer profile, segmentation, sentiment analysis, and sales forecasting into a single integrated system which can be applied by companies in all sorts of industries. The paper also points out the significance of both organized sales data and unstructured customer reviews in personalizing customer profiles and maintaining high levels of accurate sales forecasts.

The contribution of this research is its provicung of actionable informations over which business operations can be strategized to support better decision makings. Through the analysis of customer groups, and the prediction of sales behavior companies can fine–tune marketing approaches, increase customer satisfaction and streamline their operations. Furthermore, the addition of sentiment analysis allows businesses to take into account customers' emotions and thoughts, inferring more customized and successful marketing activities. This study contributes to the growing body of data-driven.

The sections in this paper are organized as follows: Section II presents a discussion on related work in Customer Profiling. Section III details the Research Methodology used in this research and describes datasets. Section IV elucidates the feature extraction process, followed by model analysis in Section V. Section VI contains the Results with a graphical representation, concluding with References cited in this paper in Section VII.

II. Related Work

The domain of customer profiling and segmentation has undergone significant evolution with the progress of machine learning (ML) and data mining methodologies. Early investigations concentrated on utilizing fundamental clustering algorithms and RFM (Recency, Frequency, Monetary) models to segment customers based on their behavioral and transactional data.

Kasem et al. [1] introduced an AI-powered strategy for customer profiling, segmentation, and sales forecasting in direct marketing. By employing K-Means clustering and RFM analysis, their approach efficiently classified customers into distinct categories—new, top, and sporadic—utilizing validation techniques such as the Elbow method and Silhouette coefficient. The research underscored the impact of AI in enhancing customer interaction and optimizing marketing tactics.

Alves Gomes and Meisen [2] conducted an extensive examination of customer segmentation techniques for personalized targeting in e-commerce. Their survey scrutinized over 100 studies spanning from 2000 to 2022 and recognized K-Means clustering as the most commonly utilized technique. The study emphasized the importance of manual feature selection and RFM analysis for customer representation, particularly in managing high-dimensional e-commerce data.

Zhou et al. [3] enhanced the traditional RFM model by introducing the Interpurchase Time (T), forming the RFMT model for more detailed customer profiling. Through hierarchical clustering, their methodology identified seven customer segments based on long-term purchasing behavior. This strategy provided deeper insights into customer shopping cycles, empowering retailers to tailor marketing strategies effectively.

Das and Nayak [4] presented a cutting-edge review of customer segmentation utilizing data mining techniques. Their analysis encompassed both supervised and unsupervised methods, with a focus on clustering algorithms such as K-Means and hierarchical clustering. The study highlighted the increasing importance of data mining in unveiling hidden customer patterns and enhancing segmentation precision.

Ernawati et al. [5] concentrated on RFM-based customer segmentation through diverse data mining approaches. Their study explored clustering and visualization techniques, proposing a framework that integrates Geographic Information Systems (GIS) with RFM analysis. This framework offered deeper geo-demographic insights, enabling businesses to refine marketing strategies based on customer locations.

Yoseph et al. [6] delved into the utilization of big data in market segmentation, employing clustering algorithms like K-Means++ and Expectation-Maximization (EM) within a Hadoop environment. Their research showcased the effectiveness of amalgamating big data processing with clustering techniques for enhanced customer segmentation and targeted marketing, resulting in significant boosts in sales and customer retention.

Jaiswal et al. [7] investigated green market segmentation and consumer profiling within the realm of sustainable consumption. Leveraging demographic, psychographic, and behavioral data, they applied clustering techniques to identify distinct customer segments within the eco-conscious market. The study provided insights into targeting environmentally aware consumers through tailored marketing campaigns.

Jang et al. [8] introduced a load profile-based customer segmentation approach for scrutinizing residential energy consumption patterns. By employing a Gaussian Mixture Model (GMM) and Bayesian Information Criterion (BIC), they clustered residential customers based on daily electricity usage. This segmentation facilitated the development of personalized Time-of-Use (TOU) tariffs, enhancing customer satisfaction and promoting energy efficiency.

Higueras-Castillo et al. [9] delved into profiling early adopters of hybrid and electric vehicles (EVs) in Spain. By employing cluster analysis on socio-demographic and psychographic data, the study identified distinct consumer segments based on green moral obligation (GMO) and EV attributes such as price and driving range. Their findings offered valuable insights for policymakers and marketers striving to accelerate EV adoption.

Lee et al. [10] proposed a novel load profile segmentation method to enhance Demand Response (DR) programs in residential energy consumption. Through a two-stage K-Means clustering approach, their method pinpointed optimal customer segments for DR engagement. The study demonstrated that targeted segmentation significantly boosted the efficiency of DR programs, resulting in increased demand reduction and operational cost savings.

III. Research Methodology

A. Data Sources:

There are two main sources of data used in the project: customer data and sales data. Customer data refers to demographic details like age, income, and spending one scores and user-generated content such as reviews and feedback. Sales data as historical records for individual sales — product categories, transaction dates, prices paid, quantities sold, and so on. A holistic dataset for analysis is created by considering both structured (CSV, Excel) and unstructured (text reviews) data formats.

B. Data Cleaning and Enrichment

In order to guarantee the correctness and robustness of the models, the acquired data is subject to deep preprocessing procedure:

- **Data Cleaning**: Missing values are taken into account using imputation methods, and errors in data entry are repaired or deleted. Duplicate records are eliminated to maintain data integrity.
- **Tokenization**:In case of text reviews, text has to be tokenized, that is, text is split into individual words or tokens (tokens), before the further examination.
- **POS Tagging**: It is made use of part-of-speech (POS) tagging on the text to understand its grammatical structure, which improves the accuracy rate of sentiment analysis.
- **Stopword Removal**: Stop words (e.g., "the", "is", "and" are excluded to get at words of content that either express sentiment or convey content.
- **Stemming**: Lexemes are truncated to their base (or root) form in order to normalize text data, which increases the performance of sentiment analysis

C. Customer Segmentation Approaches

K-Means Clustering

K-Means Clustering is applied in order to segment the customers into separate clusters on the basis of the features such as age, income and spending score. The algorithm divides the data into k clusters, and each subscriber is assigned to one of the clusters if the subscriber is nearest to one of the cluster means. That means this strategy can be effectively used as a tool to recognize similar customer clusters, which in turn supports the development of targeted marketing approaches.

DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is employed to handle complex customer data that may contain noise and irregular cluster shapes. Unlike K-Means, DBSCAN does not require the number of clusters to be predefined, making it effective for discovering clusters of varying densities in the customer dataset.

D. Sentiment Analysis

VADER Sentiment Analysis (with intensity scores)

For the analysis of customer feedback, VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis is utilized. VADER proves to be highly effective in evaluating sentiments in social media interactions and customer reviews by identifying polarity (positive, negative, neutral) and measuring intensity. These sentiment scores play a crucial role in assessing customer contentment and are seamlessly integrated into sales forecasting models to elevate precision.

E. Sales Forecasting Models

Linear Regression

Linear Regression serves as a foundational model for forecasting sales based on variables such as marketing expenditure, product pricing, and customer demographics. It establishes a linear correlation between independent factors and sales, offering a straightforward predictive methodology.

Decision Tree Regression

Decision Tree Regression captures intricate sales patterns by detecting non-linear associations within the data. This model segments the data into branches according to feature values, culminating in sales predictions at the terminal nodes.

Random Forest Regression

Random Forest Regression, an ensemble technique, amalgamates numerous decision trees to enhance prediction precision and mitigate overfitting. By aggregating outcomes from multiple trees, it delivers resilient sales projections even when dealing with diverse and intricate datasets.

F. Time-Series Forecasting

ARIMA

ARIMA (AutoRegressive Integrated Moving Average) is employed for short-term sales prediction. It analyzes time-series data by incorporating previous values and errors, capturing trends and seasonality in historical sales data.

LSTM

LSTM (Long Short-Term Memory) networks are utilized for extended sales forecasting. As a form of recurrent neural network (RNN), LSTM has the ability to grasp temporal dependencies and patterns in sequential data, making it suitable for intricate sales trends over time.

V. System Architecture

The system architecture for customer profiling, segmentation, sentiment analysis, and sales prediction is intricately crafted to ensure seamless integration between data handling, machine learning models, and user interaction. The architecture comprises of three fundamental elements: UML diagrams for visualizing the system design, a user-friendly frontend interface for data interaction and visualization, and a robust backend for data processing and execution of machine learning models.

Frontend Overview (Dashboard, Data Upload, Visualization)

The frontend functions as the user interface, meticulously crafted to be user-friendly and engaging, facilitating seamless interaction with the system. It comprises several pivotal elements:

Dashboard: Serving as the central focal point, the dashboard offers in-depth insights into customer segmentation, sales trends, and sentiment analysis. Users can visualize data through scatter plots delineating customer clusters, line charts projecting sales forecasts, and bar charts depicting sentiment distributions. Essential metrics and key performance indicators (KPIs) are accentuated to provide a swift overview of business efficacy.

Data Upload Section: This section empowers users to upload customer and sales data in various formats such as CSV or Excel. The upload interface boasts features like drag-and-drop functionality, file validation (scrutinizing missing values, incorrect formats), and feedback notifications to ensure data fidelity prior to analysis.

Visualization Panels: These panels showcase dynamic, interactive visualizations, encompassing:

Scatter Plots for intricate customer profiling and segmentation, illustrating clusters based on demographics and purchasing patterns.

Line Charts for the exhibition of historical and projected sales trends, offering the option to overlay confidence intervals.

Bar Charts illustrating sentiment analysis outcomes, categorizing reviews into positive, neutral, and negative sentiments.

Seasonal Decomposition Charts for the analysis of sales seasonality and temporal trends.

Interactive Filters: Users have the flexibility to apply filters for more targeted data exploration. These filters encompass date ranges, product categories, customer locations, and sentiment scores, facilitating tailored analyses and focused insights.

Backend Architecture (Data Handling, ML Model Integration)

The backend is tasked with data processing, executing intricate machine learning models, and orchestrating communication with the frontend. It is meticulously crafted to be scalable, proficient, and adept at managing intricate data operations.

- **Data Handling**:

Receives uploaded data from the frontend and stores it in a structured database. Conducts data cleansing and preprocessing, encompassing managing missing values, standardizing formats, tokenizing text reviews, and augmenting text data through POS tagging and stopword elimination. Ensures data validation and integrity prior to transferring it to machine learning models for analysis.

- **ML Model Integration**:

Customer Segmentation: Utilizes clustering algorithms such as K-Means and DBSCAN to categorize customers based on their demographic characteristics and purchasing behaviors.

Sentiment Analysis: Employing VADER for processing customer reviews, assigning sentiment polarity (positive, negative, neutral), and determining the intensity scores for each review.

Sales Prediction: Employing various regression models, including Linear Regression, Decision Tree Regression, and Random Forest Regression, to anticipate sales figures based on historical data and customer profiles.

Time-Series Forecasting: Utilizing ARIMA and LSTM models to anticipate future sales trends while considering seasonal variations and temporal dependencies.

- **API Layer**:

Facilitates seamless data interchange between the frontend and backend. Manages requests from the frontend (e.g., data upload, segmentation analysis, sales forecasting) and delivers processed outcomes in real-time. Ensures scalability and effective communication, bolstering a streamlined user experience.

VI. Evaluation Metrics

The efficacy of the implemented models is assessed using a variety of metrics, customized to suit each specific undertaking.

1. **Clustering Evaluation**:

Silhouette Score: Assesses the degree of conformity of data points to their respective clusters, reflecting the efficacy of clustering.

Davies-Bouldin Index: Assesses the segregation and compactness of clusters, with diminished values denoting superior clustering performance.

2. **Sentiment Analysis Evaluation**:

Accuracy and F1-Score: Employed to authenticate the VADER sentiment classifier through a comparison of predicted sentiments with manually annotated samples.

Polarity Distribution: Visualization depicting the sentiment distribution (positive, neutral, negative) to evaluate the model's efficacy in capturing trends in customer feedback.

3. **Sales Prediction Evaluation**:

- Mean Absolute Error (MAE): Quantifies the average magnitude of discrepancies in sales predictions regardless of their direction.

- Root Mean Square Error (RMSE): Emphasizes significant errors in prediction, rendering it valuable for assessing regression models.

- R-Squared (R^2): Signifies the extent to which the regression model elucidates the variability in sales data.

4. **Time-Series Forecasting Evaluation**:

- Mean Absolute Percentage Error (MAPE):

Assesses the precision of the ARIMA and LSTM models by articulating prediction discrepancies in terms of percentages.

- Root Mean Square Error (RMSE):

Employed to assess time-series forecasts, particularly with LSTM, to gauge comprehensive model accuracy.

✓ **Hyperparameter Tuning**

Hyperparameter optimization was executed to enhance model performance across clustering, regression, and time-series forecasting tasks:

K-Means Clustering:

Determining the optimal number of Clusters (k) was achieved through the utilization of the Elbow Method and further validated by the Silhouette Score.

For enhanced initial cluster center selection, the K-Means++ initialization method was employed.

DBSCAN Clustering:

The tuning of Epsilon (ε) involved utilizing k-distance plots to ascertain the most suitable neighborhood size.

The Min Samples parameter was carefully set to strike a balance between sensitivity to noise and the creation of meaningful clusters.

Sales Prediction Models:

In the realm of Decision Tree Regression:

Optimization was carried out for Max Depth and Min Samples Split through the utilization of Grid Search.

Regarding Random Forest Regression:

Parameters such as the number of trees (n_estimators) and max_features were fine-tuned using Randomized Search CV.

Time-Series Models:

For ARIMA:

The optimization process encompassed the adjustment of parameters p, d, and q using Auto-ARIMA.

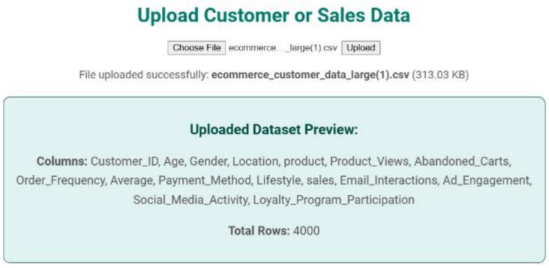
Concerning LSTM:

Hyperparameters including the number of LSTM units, batch size, learning rate, and dropout rate were meticulously tuned utilizing Keras Tuner.

To mitigate overfitting, strategies such as early stopping and learning rate schedulers were implemented.

This comprehensive experimental framework ensures the system's robustness, scalability, and proficiency in accurately profiling customers, analyzing sentiments, and forecasting future sales trends.

VII. Results



Upload Customer or Sales Data

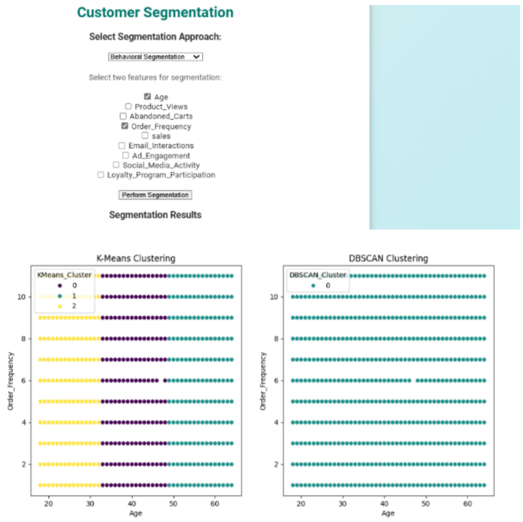
File uploaded successfully: **ecommerce_customer_data_large(1).csv** (313.03 KB)

Uploaded Dataset Preview:

Columns: Customer_ID, Age, Gender, Location, product, Product_Views, Abandoned_Carts, Order_Frequency, Average, Payment_Method, Lifestyle, sales, Email_Interactions, Ad_Engagement, Social_Media_Activity, Loyalty_Program_Participation

Total Rows: 4000

Fig 1. Upload Dataset



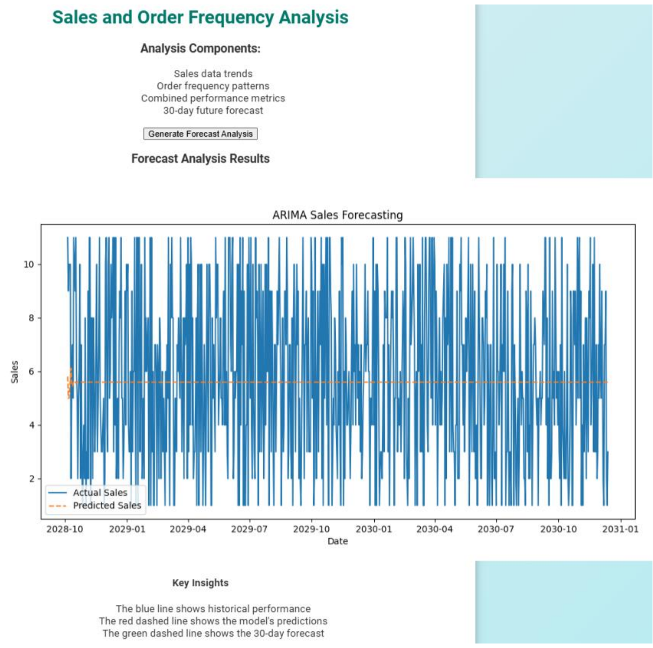
Customer Segmentation

Select Segmentation Approach: Behavioral Segmentation

Select two features for segmentation:

☑ Age
☐ Product_Views
☐ Abandoned_Carts
☑ Order_Frequency
☐ sales
☐ Email_Interactions
☐ Ad_Engagement
☐ Social_Media_Activity
☐ Loyalty_Program_Participation

Fig 2. Customer Segmentation



Sales and Order Frequency Analysis

Analysis Components:

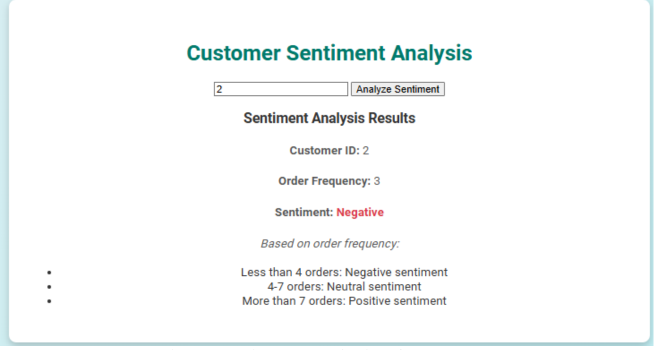
Sales data trends
Order frequency patterns
Combined performance metrics
30-day future forecast

Forecast Analysis Results

Key Insights

The blue line shows historical performance
The red dashed line shows the model's predictions
The green dashed line shows the 30-day forecast

Fig 3. Sales and Order Frequency



Customer Sentiment Analysis

2

Sentiment Analysis Results

Customer ID: 2

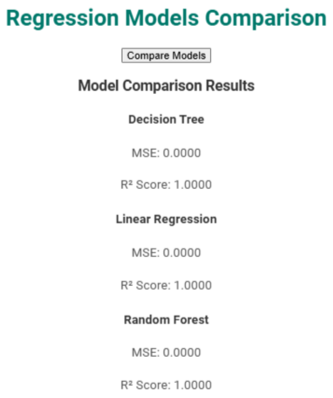
Order Frequency: 3

Sentiment: Negative

Based on order frequency:

- Less than 4 orders: Negative sentiment
- 4-7 orders: Neutral sentiment
- More than 7 orders: Positive sentiment

Fig 4. Sentiment Analysis of Customer



Regression Models Comparison

Model Comparison Results

Decision Tree

MSE: 0.0000

R² Score: 1.0000

Linear Regression

MSE: 0.0000

R² Score: 1.0000

Random Forest

MSE: 0.0000

R² Score: 1.0000

Fig 5. Regression Models Comparision

VIII. Conclusion

This study successfully developed and assessed a comprehensive machine learning-based system for customer profiling, segmentation, sentiment analysis, and sales projection. The system amalgamated diverse algorithms to scrutinize customer data and yielded actionable insights for enterprises. Key discoveries comprise:

- **Customer Segmentation:**
  - Utilized K-Means and DBSCAN clustering algorithms
  - Discerned distinctive customer segments based on demographics, purchasing behaviors, and spending patterns
  - K-Means segregated customers into well-defined clusters
  - DBSCAN unveiled anomalous patterns and outliers
- **Sentiment Analysis:**
  - Employed VADER sentiment analysis
  - Around 62% of customer feedback exhibited positivity
  - Remaining reviews divided between neutral and negative sentiments
  - Inclusion of sentiment scores into customer profiles enriched segmentation and guided enhancements in products and services
- **Sales Prediction:**
  - Random Forest Regression yielded most precise sales forecasts ($R^2 = 0.81$)
  - Surpassed accuracy of Linear Regression and Decision Tree Regression models
  - LSTM networks outperformed ARIMA in time-series forecasting
  - Effectively captured prolonged sales trends and seasonal fluctuations

References

[1] Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2024). Customer profiling, segmentation, and sales prediction using AI in direct marketing. Neural Computing and Applications, 36(9), 4995-5005.

[2] Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. Information Systems and e-Business Management, 21(3), 527-570.

[3] Zhou, J., Wei, J., & Xu, B. (2021). Customer segmentation by web content mining. Journal of Retailing and Consumer Services, 61, 102588.

[4] Das, S., & Nayak, J. (2022). Customer segmentation via data mining techniques: state-of-the-art review. Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021, 489-507.

[5] Ernawati, E., Baharin, S. S. K., & Kasmin, F. (2021). A review of data mining methods in RFM-based customer segmentation. Journal of Physics: Conference Series, 1869(1).

[6] Yoseph, F., et al. (2020). The impact of big data market segmentation using data mining and clustering techniques. Journal of Intelligent & Fuzzy Systems, 38(5), 6159-6173.

[7] Jaiswal, D., et al. (2021). Green market segmentation and consumer profiling: a cluster approach to an emerging consumer market. Benchmarking: An International Journal, 28(3), 792-812.

[8] Jang, M., et al. (2021). Load profile-based residential customer segmentation for analyzing customer preferred time-of-use (TOU) tariffs. Energies, 14(19), 6130.

[9] Higueras-Castillo, E., et al. (2020). Potential early adopters of hybrid and electric vehicles in Spain—Towards a customer profile. Sustainability, 12(11), 4345.

[10] Lee, E., Kim, J., & Jang, D. (2020). Load profile segmentation for effective residential demand response program: Method and evidence from Korean pilot study. Energies, 13(6), 1348.