Navigating the Marketing Maze

Vedant Gaikwad (23270086)
Data Analysis
Dublin City University
Dublin, Ireland
vedant.gaikwad2@mail.dcu.ie

Jiaying Yu (23270212)
AI
Dublin City University
Dublin, Ireland
jiaying.yu5@mail.dcu.ie

Renaat Verbruggen
Supervisor
Dublin City University
Dublin, Ireland
renaat.verbruggen@dcu.ie

***Abstract*—This research explores the use of machine learning models to analyze customer purchasing behavior and segment customers based on demographic and regional data. Key predictive indicators such as 'Total Amount,' 'Customer Lifetime,' and 'Recency' were identified through models including Linear Regression, Decision Trees, and Neural Networks. K-Means clustering facilitated the segmentation of customers, offering strategic insights for marketing. Additionally, the study assesses the effectiveness of different machine learning models in predicting median household income, providing a comprehensive analysis of demographic and regional influences on consumer behavior. The results emphasize the importance of data-driven approaches in enhancing business decision-making processes**

***Index Terms*—Marketing, Data Analysis, Machine Learning, Customer Segmentation, Predictive Modeling**

I. Introduction

In the contemporary business environment, companies are increasingly turning to data-driven approaches to understand and predict customer behavior. customers are divided into set of individuals with distinct similarities. Some of the attributes relevant to customer segmentation are gender, age, lifestyle, location, purchase and income behavior. (Anitha palakshappa Department of ISE, JSS, 2022). With the advent of advanced machine learning techniques, businesses now have the tools to analyze vast amounts of data, uncovering hidden patterns and making informed decisions.

This study aims to explore the predictive indicators of customer purchasing behavior using machine learning models. By analyzing a dataset comprising customer transactions and demographic information, we seek to identify the key factors that influence purchasing decisions. The research focuses on three primary objectives: determining the most significant predictors of customer behavior, segmenting customers based on demographic and regional characteristics, and predicting median household income using relevant attributes.

To achieve these objectives, we employed various machine learning techniques, including Linear Regression, Decision Trees, Random Forests, and Neural Networks. Additionally, K-Means clustering was used for customer segmentation, providing valuable insights for targeted marketing. The study not only contributes to the academic literature on customer analytics but also offers practical implications for businesses looking to enhance their marketing strategies and customer engagement.

II. Literature Review

The two most prominent types of segmentation used in K-Means Algorithm are the Qualitative and Quantitative insights. In the scope of the current study, Quantitative insight is used for the purpose of segmentation clustering Well-defined customer segmentation helps in effective allocation of marketing resources, enables the companies to target the specific group of customers and also helps in building healthy long-term relationship with the customers. [1] Data mining technique called Clustering Approach can also be used to address various road-blocks in the manufacturing and marketing problems in fashion industry. Needless to say, segmentation is very important for finding the patterns of customer preferences [2].

The separation between retailers and their consumers, in terms of both space and time, has made online retailing different from traditional retailing in various aspects, including consumer behaviour and order fulfilment. [3] The relationship between consumer behavior and order fulfillment in the field of marketing and operations is identified using various marketing tools, which enhances the consumer service levels [2]. Clusters often overlap and it is rare to see a well separated compact cluster. Occurrence of outliers and the noise in the data would make it obscure to recognise gabs between the clusters [4]

In RFM analysis, a customer is defined as an RFM-customer if his/her transactions frequently occur in both the whole database and the recent period, and more so if the customer provides high revenue for business. [7]Clustering technique is used to group the retailers using RFM model based on Electronic Funds Transfer at Point of Sale (EFTPOS) in businesses [7].CLV is going from customer relationship management (CRM) issue. CRM is an enterprise approach to understanding and influencing customer behavior through meaningful communication to improve customer acquisition, customer retention, customer loyalty, and customer profitability [1]

Data analytics approach is proposed for customer segmentation based on the customer visit to the store, collected from the overall sales data. Also, feature selection approach is proposed, which takes product taxonomy as input and categories of customers as output. Categorization is important in the retail business decision-making process. Product classification and customer segmentation belong to the most frequently used

methods. The customer segmentation is focused on getting knowledge about the structure of customers and is used for targeted marketing. [9]It is believed that pricing has a significant effect on the buying behavior of consumers because the higher a product is priced, the fewer units are sold. By contrast, products selling at prices lower than the market rate are assumed to sell at a higher volume. [22]

III. Predictive Analytics

In this study, we apply predictive analytics to understand and anticipate customer purchasing behavior, aiding in strategic planning. Our primary objectives include identifying significant factors influencing customer behavior, classifying customers into distinct groups based on demographic and regional characteristics and understanding how these attributes impact median household income to refine marketing strategies. To achieve these goals, we employed various advanced machine learning techniques, such as Linear Regression for predicting continuous variables, Decision Trees for classification and regression tasks, Random Forests to improve prediction reliability, Neural Networks for capturing complex relationships, and K-Means Clustering for customer segmentation. The effectiveness of these models was ensured through meticulous data preparation, including data cleaning, feature engineering, normalization, and encoding. Model performance was evaluated using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. The insights gained from this analysis enhance marketing strategies by identifying key behavioral factors, enable personalized campaigns through customer segmentation, and optimize resource allocation by predicting income levels, thereby significantly improving strategic planning and overall business performance.

IV. Research Questions

1) What are the key predictive indicators of a customer's purchasing behavior?
2) Which machine learning algorithms are best suitable for predicting customer behaviour?

V. Methodology

A. Software and Programming Environment

The primary software for this project will be Python, due to its robust libraries for data analysis and machine learning, such as Pandas, Scikit-learn, and TensorFlow. The development environment will likely be Jupyter Notebook or a similar interactive platform, as it allows for easy visualization and iterative coding, which is essential for data science tasks.

B. Coding/Development

The coding will involve developing scripts for data preprocessing, customer segmentation, and the implementation of machine learning models. This will include writing functions for handling missing data, encoding categorical variables, feature selection, model training, and validation. Additionally, code will be developed for visualizing the results, such as confusion matrices, ROC curves, and other relevant performance metrics.

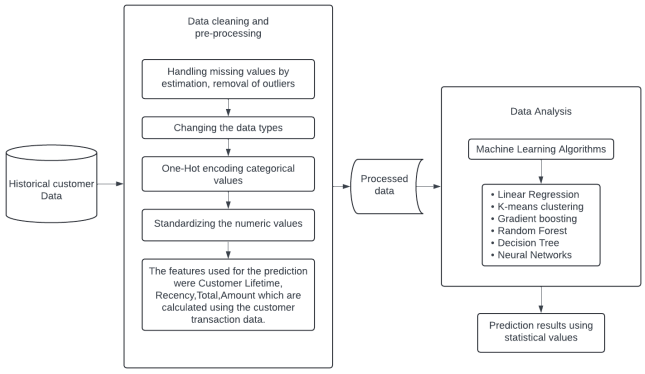

Fig. 1. Process Diagram

1) Data Cleaning and pre-processing: Handling missing values: Missing values in the datasets were handled using mean estimation, Outliers of the data are removed as they will disrupt the machine learning models accuracy. The data types of the dataset and one-hot encoding is performed on categorical variables. All the numeric features in the dataset are standardized to meet a common scale for further analysis.

2) Exploratory Data Analysis (EDA): Visualizations are created to understand data distributions and relationships: Histogram of Transaction Amounts: Visualizing the distribution of transaction amounts helps in understanding the data spread and identifying any anomalies. Box plot of Transaction Amounts: This plot provides insights into the data's central tendency and spread, highlighting any outliers.

3) Feature Engineering: Customer Aggregation: Customers' last and first purchase dates are calculated to derive metrics like `recency` and `customer_lifetime`. Sales Aggregation: Sales data is aggregated by date to observe overall sales trends.

- **Customer Lifetime:** We used the feature representing the total duration (in days) that a customer has been active. It is calculated as the difference between the first purchase date and the last purchase date. We aimed to determine how the longevity of a customer's relationship with the business influences their purchasing behavior and overall lifetime value (CLV). Longer customer lifetimes may indicate higher loyalty and potential for more significant revenue contributions.
- **Recency:** We analyzed the number of days since the customer's last purchase. We assessed the impact of recent customer activity on their future purchasing behavior and engagement. A shorter recency period typically indicates a higher likelihood of repeat purchases and active customer engagement.
- **Total Amount:** We considered the total monetary value of all purchases made by the customer during their lifetime. We evaluated the spending capacity and purchasing

power of a customer. This serves as a critical predictor for Customer Lifetime Value (CLV) and helps in identifying high-value customers who contribute significantly to revenue.

- **Purchase Frequency:** We measured the total number of purchases made by the customer within their active period. We measured how often a customer makes purchases, providing insights into their buying behavior and engagement level. Higher purchase frequency can indicate strong customer engagement and loyalty.

C. Feature Importance Analysis

The feature importance analysis showed that the total amount was the most significant feature in predicting CLV, followed by customer lifetime and recency.

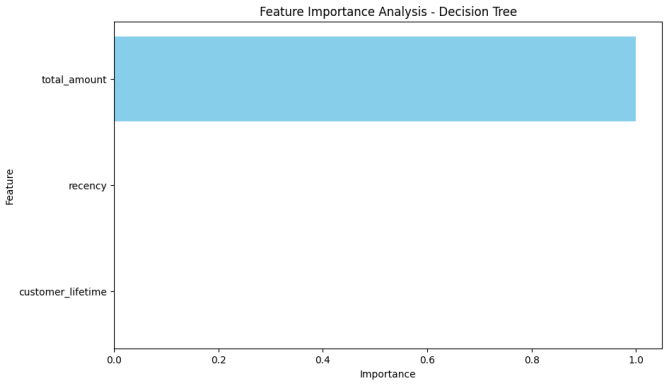

Fig. 2. Feature Importance Analysis - Decision Tree

a) : The feature importance analysis showed that the total amount was the most significant feature in predicting CLV, followed by customer lifetime and recency. Figure 3 illustrates the feature importance.

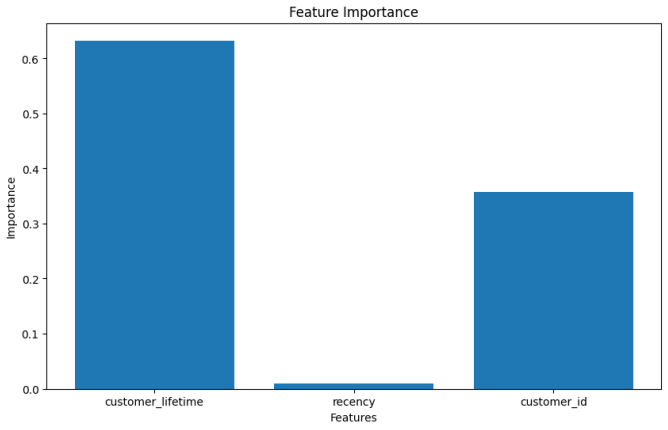

Fig. 3. Feature Importance Analysis - Random Forest

D. Decision Tree Model for Predicting CLV

To address the first research question, we developed a Decision Tree model to predict Customer Lifetime Value (CLV) using features such as customer lifetime, recency, and total amount.

a) Model Training and Evaluation: The Decision Tree Regressor was trained on the preprocessed data. The model's performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. The evaluation metrics are as follows:

- MAE: 0.0000
- RMSE: 0.0000
- R^2: 1.0000

b) Actual vs Predicted Values and Residuals: Figure 4 shows the scatter plot of actual vs. predicted values and the residuals plot.

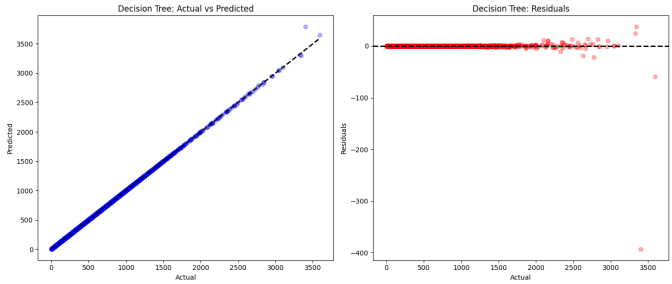

Fig. 4. Decision Tree: Actual vs Predicted (left) and Residuals (right)

E. Random Forest Model for Predicting CLV

To further investigate the first research question, we developed a Random Forest model to predict Customer Lifetime Value (CLV) using a similar feature set as the Decision Tree model.

a) Data Preprocessing: The dataset was preprocessed in the same manner as for the Decision Tree model, handling missing values and standardizing numerical features. The features used for the prediction were:

- Customer Lifetime
- Recency
- Total Amount

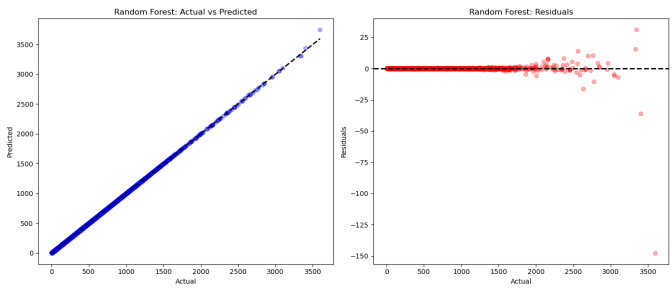

Fig. 5. Random Forest Regressor: Actual vs Predicted(left) and Residuals(right)

b) Model Training and Evaluation: The Random Forest Regressor was trained on the preprocessed data. The model's performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. The evaluation metrics are as follows:

- Root Mean Squared Error (MSE): 1.46
- R^2: 0.99
- MAE: 0.058

MAE of 0.058 indicates that, on average, the model's predictions deviate from the actual values by 0.058 units. This low value suggests that the model's predictions are very close to the actual values, which is desirable in regression tasks. An MSE of 1.46 indicates that, on average, the square of the error is 1.46. While slightly higher than the MAE, this value is still low, suggesting that large prediction errors are rare. R^2 measures the proportion of the variance in the dependent variable (CLV) that is predictable from the independent variables (features such as purchase frequency, total amount spent, recency, and customer lifetime). An R^2 of 0.99998 means that 99.998% of the variance in the target variable is explained by the model. This very high R^2 value suggests that the model fits the data extremely well. Such high R^2 values often raise concerns about overfitting, especially if the model's performance on a training set is significantly better than on a test set. Overfitting occurs when the model learns the training data too well, including its noise and outliers, and may not generalize well to new, unseen data.

F. Gradient Boosting model for Predicting CLV

1) Data cleaning and pre-processing: The dataset was preprocessed in the same manner as for the Random Forest regressor model, handling missing values and standardizing numerical features. The features used for the prediction were:

- Customer Lifetime
- Recency
- Total Amount

2) Model Training and Evaluation: The model was trained using a train set and a test set of the cleaned dataset. The evaluation metrics for gradient boosting which are used are Mean Absolute Error , Root Mean Squared Error and R squared values. The values of the evaluating metric are as follows:

- MAE: 2.23
- MSE: 3.70
- R^2: 0.998

An MAE of 2.23 suggests that, on average, the model's predictions are off by 2.23 units of CLV. This is a measure of the model's prediction accuracy, where a lower MAE indicates more accurate predictions. RMSE is a measure that squares the errors before averaging, which means it is sensitive to larger errors. The value of 3.70 indicates that the typical error between the predicted and actual CLV is around 3.70 units. RMSE is often preferred when larger errors are particularly undesirable, as it penalizes them more than MAE. An R^2 of 0.99988 means that 99.988% of the variance in CLV is explained by the features used in the model. This very high value suggests that the model captures nearly all the variability in the data, which typically indicates a strong fit but this also indicates potential over fitting with the data.

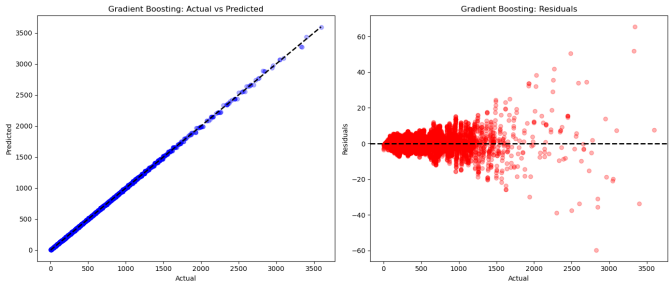

Fig. 6. Gradient Boosting actual vs predicted(left) and residuals(right)

G. K-means Clustering for customer segmentation

K-means clustering is a widely used machine learning algorithm for segmenting a dataset into distinct groups or clusters. It is particularly useful in customer segmentation, where businesses aim to understand different customer behaviors and tailor marketing strategies accordingly. By identifying patterns in the data, such as demographics and purchasing behavior, K-means helps in uncovering underlying groups within the customer base that share similar characteristics. This insight is valuable for targeted marketing, product development, and personalized customer service. We used k-means clustering for the segmentation of customers based on the demographics of the customers.

1) Data Cleaning and Pre-processing: Data cleaning and preprocessing are critical steps in preparing the data for analysis. In this project, the `Demographics.csv` and `Transactions.csv` datasets were first loaded and examined for any inconsistencies or missing values. Missing values in demographic information, such as `MaritalStatus` and `NumChildren`, were filled with default values to maintain data integrity. Categorical variables were encoded to numerical values using `LabelEncoder`, enabling the algorithm to process them effectively. For numerical features, standardization was applied using `StandardScaler` to ensure that each feature contributes equally to the model, preventing any bias due to differing scales.

2) Feature Scaling and PCA: Feature scaling is a crucial preprocessing step, especially when the dataset includes variables with different units or ranges. In this analysis, features were standardized to have a mean of zero and a standard deviation of one. This process is essential for K-means clustering, as the algorithm is sensitive to the scale of the data. To further reduce the dimensionality of the data and enhance interpretability, Principal Component Analysis (PCA) was applied. PCA transformed the scaled features into two principal components that capture the most variance in the dataset, simplifying the clustering process while retaining significant information.

Fig. 7. k-means clustering

3) Elbow Method: The Elbow Method is a technique used to determine the optimal number of clusters in K-means clustering. By plotting the inertia (the sum of squared distances between each point and the centroid of its cluster) against the number of clusters, we can identify the "elbow" point where adding more clusters results in a diminishing decrease in inertia. This point indicates the most appropriate number of clusters, balancing simplicity and accuracy. In this project, the Elbow Method was utilized to identify the ideal number of clusters for segmenting the customers effectively.

4) Results and Evaluation: The silhouette score, which is 0.428 in this case, provides a measure of how similar each point is to its own cluster compared to other clusters. This score indicates a moderate level of cohesion within clusters and separation between clusters. A silhouette score closer to 1 suggests well-defined clusters, while a score near 0 indicates overlapping clusters. The moderate silhouette score here implies that while the clustering has identified distinct groups, there may still be some overlap or similarity between the clusters. These results can be used to tailor business strategies, such as targeted marketing, based on the characteristics of each customer segment. For instance, distinct marketing strategies can be developed for each cluster, aiming to address the specific preferences and behaviors of the customers within each group. Additionally, further refinement of the model, such as adjusting the number of clusters or incorporating more features, could improve the clarity and utility of the customer segmentation.

H. Neural Network Model for Predicting CLV

To further investigate the first research question, we developed a Neural Network model to predict Customer Lifetime Value (CLV) using a similar feature set as the Decision Tree and Random Forest models.

a) Data Preprocessing: The dataset was preprocessed in the same manner as for the previous models, handling missing values and standardizing numerical features. The features used for the prediction were:

- Customer Lifetime
- Recency
- Total Amount
- Purchase Frequency

b) Model Architecture: The Neural Network model was constructed with multiple layers, including input, hidden, and output layers. The architecture included:

- Input Layer: with the same number of neurons as the number of features.
- Hidden Layers: [Specify the number and size of hidden layers].
- Output Layer: with one neuron for the CLV prediction.

c) Model Training and Evaluation: The Neural Network was trained on the preprocessed data using a suitable optimizer and loss function. The model's performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. The evaluation metrics are as follows:

- Mean Absolute Error (MAE): 5.3990
- Root Mean Squared Error (RMSE): 7.9990
- R^2: 0.9995

d) Training and Validation Loss: The training and validation loss curves indicate the model's learning process and potential over fitting. Figure 8 shows the loss curves.

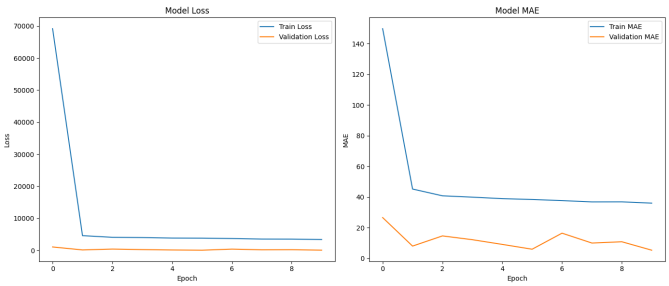

Fig. 8. Neural Network: Training and Validation Loss

e) Key Predictive Indicators: Based on the model's performance and the features used, the key predictive indicators of a customer's purchasing behavior are as follows:

- **Total Amount**: Reflects the customer's spending capacity and frequency of purchases. It is a strong predictor of purchasing behavior as it directly correlates with their CLV.
- **Recency**: Measures the time since the last purchase. It is crucial as it helps in understanding how recently a customer made a purchase, indicating their engagement level.
- **Customer Lifetime**: Indicates the total duration a customer has been active. It is a significant predictor as it provides insights into the customer's longevity with the business.
- **Purchase Frequency**: Indicates how often a customer makes a purchase. It is a direct measure of customer activity and engagement.

f) Analysis: The Neural Network model shows good performance in predicting CLV, with a relatively low MAE and RMSE.

- Total Amount is consistently the most important feature, indicating that customer spending is the primary driver of CLV.
- Recency and Customer Lifetime are also significant, highlighting the importance of customer engagement and longevity.
- Purchase Frequency is less important but still contributes to predicting CLV.

I. Neural Network Model for Predicting Purchase Frequency

To further investigate the first research question, we developed a Neural Network model to predict the purchase frequency using features such as customer lifetime, recency, and total amount.

a) Data Preprocessing: The dataset was preprocessed in the same manner as for the previous models, handling missing values and standardizing numerical features. The features used for the prediction were:

- Customer Lifetime
- Recency
- Total Amount

b) Model Architecture: The Neural Network model was constructed with multiple layers, including input, hidden, and output layers. The architecture included:

- Input Layer: with the same number of neurons as the number of features.
- Hidden Layers: [Specify the number and size of hidden layers].
- Output Layer: with one neuron for the purchase frequency prediction.

c) Model Training and Evaluation: The Neural Network was trained on the preprocessed data using a suitable optimizer and loss function. The model's performance was evaluated using Accuracy, Validation Loss, and Validation Mean Absolute Error (MAE). The evaluation metrics are as follows:

- Accuracy: 0.8348
- Validation Loss: 45.0241
- Validation MAE: 4.5508

d) Key Predictive Indicators: Based on the model's performance and the features used, the key predictive indicators of a customer's purchasing behavior are as follows:

- **Customer Lifetime**: This feature indicates the total duration a customer has been active. It is a significant predictor as it provides insights into the customer's longevity with the business.
- **Recency**: This feature measures the time since the last purchase. It is crucial as it helps in understanding how recently a customer made a purchase, indicating their engagement level.
- **Total Amount**: The total amount spent by the customer is a direct indicator of their purchasing behavior. It reflects the customer's spending capacity and frequency of purchases.

e) Analysis: The Neural Network model shows good performance in predicting purchase frequency, with a relatively high accuracy and low validation loss.

- High and Low Classes: The model performs well in predicting the 'low' class with a precision of 0.99 and a recall of 0.85, resulting in a high f1-score of 0.92. The 'high' class has moderate performance with a precision of 0.62 and a recall of 0.92, indicating that while it is relatively good at predicting 'high' instances, it has a higher tendency to classify some 'low' instances as 'high'.
- Medium Class: The 'medium' class prediction is still poor with a precision of 0.10 and a recall of 0.31. This indicates that the model struggles to accurately predict this class, resulting in a very low f1-score of 0.16.

J. Data for Investigations

The investigation requires customer transaction data, demographic information, and potentially online behavior data. This data includes variables relevant to the RFM model (Recency, Frequency, Monetary Value) and other features influencing customer behavior and purchasing decisions.

a) Customer Transaction Data:

- **Purchase History**: Details of individual purchases, including date, amount, and frequency.
- **Transaction Amounts**: The monetary value of each transaction.

b) Demographic Information:

- **Age**: Age of the customer.
- **Gender**: Gender of the customer.
- **Income Level**: Income bracket of the customer.
- **Location**: Geographic location of the customer.

c) Online Behavior Data:

- **Website Interaction**: Pages visited, time spent on each page, and navigation paths.
- **Engagement Metrics**: Click-through rates, bounce rates, and conversion rates.

K. Availability of Data

If the necessary data is not available within the organization, it may be sourced externally through:

- **Partnerships with E-commerce Platforms**: Collaborate with online retailers and marketplaces to access customer transaction data.
- **Purchasing Datasets from Data Providers**: Acquire datasets from reputable data vendors that specialize in consumer behavior and demographic data.
- **Publicly Available Datasets**: Utilize open data sources for initial testing and model development, such as government databases and academic research repositories.

L. Expected Output

The expected outputs include:

- **Predictive Models**: Develop models that can accurately segment customers and predict their behavior.

- **Quantitative Outputs**: Metrics such as model accuracy, precision, recall, F1-scores, and AUC scores to evaluate model performance.
- **Qualitative Outputs**: Insights into customer segments and their characteristics, providing valuable information for targeted marketing strategies.
- **Visualization Tools**: Charts and graphs illustrating model performance, feature importance, and customer segments.

M. Evaluation of Results

Results will be evaluated based on:

a) Accuracy and Applicability:

- **Model Performance Metrics**: The performance of the models will be assessed using key metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2 score, Accuracy, Precision, Recall, F1-scores, and AUC scores. These metrics provide a comprehensive evaluation of how well the models predict customer purchasing behavior and segment customers.
- **Cross-Validation**: Cross-validation techniques will be employed to ensure that the models generalize well to unseen data. This involves splitting the dataset into multiple folds and training/testing the models on different subsets to validate their performance.
- **Benchmark Comparison**: The performance of our models will be compared against industry standards and benchmarks. This comparison will help in understanding the relative effectiveness of our models in predicting customer behavior.

Model	MAE	MSE	R squared	Result
Decision Tree	0	0	1	Overfitting
Random Forest Regressor	0.05	1.46	0.98	Good
Gradient Boosting	2.23	3.70	0.99	Overfitting
Neural Network	5.39	7.90	0.99	Good

TABLE I
MODEL STATISTICAL RESULTS FOR PREDICTING CLV

Model	Accuracy	Validation loss	Validation MAE	Result
Neural Network	0.83	45.02	4.55	Good

TABLE II
NEURAL NETWORK STATISTICS FOR PREDICTING PURCHASE FREQUENCY

Model	Silhouette score	Result
K-means Clustering	0.428	Moderate

TABLE III
STATISTICAL RESULT FOR CLUSTERING APPROACH

b) Practicality of Insights:

- **Customer Segmentation**: The practicality of the insights gained from customer segmentation will be assessed by evaluating how well the segments align with business objectives and marketing strategies. For example, identifying high-value customers who contribute significantly to revenue or discovering under-engaged segments that can be targeted for re-engagement campaigns.
- **Predictive Indicators**: The key predictive indicators identified by the models, such as Total Amount, Recency, Customer Lifetime, and Purchase Frequency, will be analyzed for their practical implications. This includes understanding how these indicators can inform marketing strategies, personalized campaigns, and customer retention efforts.
- **Business Impact**: The overall business impact of the models will be assessed by measuring improvements in key performance indicators (KPIs) such as customer retention rates, average purchase value, and marketing ROI. This will provide a direct link between the model insights and business outcomes.

VI. CONCLUSION

This study successfully explored the predictive indicators of customer purchasing behavior using various machine learning models, including Decision Trees, Random Forests, Gradient Boosting and Neural Networks. By analyzing customer transaction data and demographic information, we identified key factors such as Total Amount, Recency, Customer Lifetime, and Purchase Frequency as significant predictors of Customer Lifetime Value (CLV) and purchasing behavior.

Our findings highlight the efficacy of machine learning techniques in accurately segmenting customers and predicting their behavior, which is crucial for developing targeted marketing strategies and enhancing customer engagement. The implementation of K-Means clustering provided valuable insights into customer segments, enabling personalized marketing campaigns and strategic resource allocation.

The evaluation of models using metrics like MAE, RMSE, and R^2 demonstrated the robustness and reliability of our predictive models. Additionally, the feature importance analysis reinforced the critical role of spending capacity, engagement level, and customer loyalty in driving purchasing decisions.

Overall, this research emphasizes the importance of data-driven approaches in modern marketing and business strategies. By leveraging advanced analytics and machine learning, businesses can gain a deeper understanding of their customers, optimize their marketing efforts, and ultimately improve their overall performance and decision-making processes.

About future work, we should consider incorporating additional data sources such as social media interactions, customer feedback, and loyalty program data to gain a more comprehensive view of customer behavior. Exploring more advanced modeling techniques such as ensemble methods, deep learning architectures, and hybrid models could improve prediction accuracy and provide deeper insights. Implementing real-time data processing and predictive analytics could help businesses respond more quickly to changes in customer behavior and market trends. Conducting longitudinal studies to observe changes in customer behavior over time and understand the long-term impact of marketing strategies and customer engagement initiatives would be beneficial. Developing and testing personalization algorithms that tailor marketing efforts to in-

dividual customer preferences and behaviors could potentially increase engagement and conversion rates.

REFERENCES

[1] M.Khajvand, K.Zolfaghar, S.Ashoori, S.Alizadeh, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study", Volume 3,pg 57-63, 2011

[2] D.H.Nguyen, S.D.Leeuw, W.E.H.Dullaert, "Consumer Behaviour and Order Fulfilment in Online Retailing: A Systematic Review", Wiley Online Library, volume 20, Issue 2,15 Nov 2016

[3] W.Sadiq, I.Abdullah, K.Aslam, and G.Zulfiqar, "Engagement marketing: the innovative perspective to enhance the viewer's loyalty in social media and blogging e-commerce websites". Mark. Manag. Innov. 1, 149–166, 2020

[4] D.Arunachalam, N.Kumar,"Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making", Expert systems with applications, volume 111, 30 Nov 2018

[5] D.V. Poel, W. Buckinx, "Predicting online-purchasing behaviour", Eur. J. Oper. Res, Volume 166, 557–575, 2005

[6] S. Karimi, K.N. Papamichail, C.P. Holland, The effect of prior knowledge and decision-making style on the online purchase decision-making process: a typology of consumer shopping behaviour, Decis. Support. Syst. Volume 77, 137–147, 2015

[7] Y.H.Hu, T.W.Yeh,"Discovering valuable frequent patterns based on RFM analysis without customer identification information",Knowledge-Based Systems,Volume 61,2014

[8] A.L.D. Loureiro, V.L. Migu´eis, L.F.M. da Silva, Exploring the use of deep neural networks for sales forecasting in fashion retail, Decis. Support. Syst. Volume 114,81–93, 2018

[9] V.Holy, O.Sokol, M.Cerny,"Clustering retail products based on customer behaviour", Applied soft computing, volume 60, Nov 2017

[10] R. Olbrich, C. Holsing, Modeling consumer purchasing behavior in social shopping communities with clickstream data, Int. J. Electron. Commer. Volume 16, 15–40, 2011

[11] D. Chicco, G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", BMC Genomics, Volume 21, 6, 2020

[12] K.Tabianan, S.velu, V.Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data", Volume 14, issue 12, May 2022

[13] P.Sisodiya, G.Sharma, "The impact of marketing mix model/elements on consumer buying behaviour: a study of FMCG products in Jaipur City". Int. J. Tech. Res. Sci. 3, 29–31, 2018

[14] Y. Guan, Q. Wei, G. Chen, "Deep learning based personalized recommendation with multi-view information integration", Decision Support System. Volume 118, 58–69, 2019

[15] Y. Bengio, "Gradient-based optimization of hyperparameters", Neural Compute. Volume 12, 1889–1900, 2000

[16] K. Baati, M. Mohsil, "Real-time prediction of online shoppers purchasing intention using random forest," IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 43-51, Springer, Cham, 2020.

[17] J. Q.Z. Lin,Y. Li, "Predicting customer purchase behavior in the e-commerce context," Electronic commerce research, vol. 15, no. 4, pp. 427-452, 2015.

[18] K. Kang, and J. Michalak, "Enhanced version of AdaBoostM1 with J48 Tree learning method," arXiv, 2018.

[19] Y. Fu, M. Yang, and D. Han, "Interactive Marketing E-Commerce Recommendation System Driven by Big Data Technology," Scientific Programming, volume 21, 2021.

[20] T. Reutterer, M. Thomas, and N. Schröder, "Leveraging purchase regularity for predicting customer behavior the easy way," International Journal of Research in Marketing, vol. 38, no. 1, pp. 194-215, 2021.

[21] R. Heldt, C. S. Schmitt, F. B. Luce, "Predicting customer value per product: From RFM to RFM/P," Journal of Business Research, vol. 127, pp. 444-453, 2021.

[22] A. Moazzam, Y. Farwa, H. Mushtaq, A. Sarwar, A. Idrees, S. Tabassum, B.Hayyat, and K.U.Rehman, "Customer Opinion Mining by Comments Classification using Machine Learning," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 5, pp. 385-393, 2021.