Customer Lifetime Value

Tarun Kumar Arya
IIITD
New Delhi
tarun21295@iiitd.ac.in

Rohit Raj
IIITD
New Delhi
rohit21279@iiitd.ac.in

Harshit Pal
IIITD
New Delhi
harshit21255@iiitd.ac.in

Aryan Dhawan
IIITD
New Delhi
aryan21023@iiitd.ac.in

Abstract

In today's business world, understanding customer behavior becomes very important. Our Project's aim is to analyze the customer's dataset which contains the crucial information like the customer's basic information(age, gender etc), brand loyalty, purchase history etc. By understanding the importance of these factors and analyzing them in depth, we will get valuable insights about our customer. With the help of our ML model, businesses will able to optimize their strategies in order to maximize their profits. Ultimately, they will be able to to segment customers effectively, which helps them to target their customers more precisely and in reducing churn. ***GitHub Link:*** *Click Here*

1. Introduction

1.1. Background and Importance of CLV

Customer Lifetime Value (CLV) is a critical metric for predicting the total revenue a customer generates throughout their relationship with a business. Understanding CLV helps companies make data-driven decisions, optimize marketing strategies, allocate resources, and enhance customer retention. In competitive markets, focusing on CLV enables sustainable growth by identifying and retaining high-value customers.

1.2. Objective and Segmentation

This report segments customers into three categories based on CLV:

- **Gold**: High-value customers requiring personalized attention.
- **Silver**: Moderate-value customers with growth potential.
- **Bronze**: Low-value customers at risk of churn but with potential for re-engagement.

This segmentation helps prioritize resources, enhance profits, and reduce churn.

1.3. Business Implications

CLV-based segmentation allows businesses to provide tailored services for Gold customers, nurture Silver customers into higher-value segments, and re-engage Bronze customers to improve retention and profitability.

2. Literature Survey

2.1. Paper 1: Retail Data Predictive Analysis Using Machine Learning Models

Research Significance:
Highlights the importance of predictive analytics in the retail sector for optimizing sales strategies and improving profitability.

Results:
The findings from this study emphasize the value of machine learning in transforming retail operations. By utilizing XGBoost for forecasting, K-Means for segmentation, and time series analysis for trend identification, businesses can optimize their supply chains, enhance customer engagement, and make strategic decisions backed by data. These insights pave the way for future exploration with ensemble models and price optimization strategies to further improve retail performance.

2.2. Paper 2: Machine Learning for Revenue Forecasting: A Case Study in Retail Business

Research Significance:
This study demonstrates the importance of machine learn-

ing in revenue forecasting for retail businesses by leveraging data-driven insights to enhance decision-making and optimize performance. The key outcomes highlight the practical applications and impact of forecasting techniques on business operations.

Results:

The findings from this case study underscore the value of using machine learning models for accurate revenue forecasting in retail. By applying XGBoost and time series forecasting, businesses gain a competitive advantage through precise predictions and operational efficiency. The study suggests future exploration of hybrid models and automated forecasting solutions to further improve accuracy and decision-making in dynamic retail environments.

3. Dataset and Preprocessing

3.1. Datset Description

The dataset contains various customer-related fields that are used to analyze and predict customer lifetime value. The key columns include:

- **Transaction_ID**: Unique identifier for each transaction (287,005 unique values).
- **Customer_ID**: Unique identifier for each customer (86,485 unique values).
- **Name, Email, Phone, Address, City, State, Zipcode, Country, Age, Gender, Income, Customer Segment**: Personal details and demographic information about customers. Date, Year, Month, Time: Transaction timestamps, useful for temporal analysis.
- **Total_Purchases**: Number of items purchased per transaction.
- **Total_Amount**: Total monetary value spent in a transaction.
- **Product_Category, Product_Brand, Product_Type**: Details about the products purchased.
- **Feedback, Shipping_Method, Payment_Method, Order_Status, Ratings**: Transactional feedback, logistics, and status indicators.

This dataset contains 30 columns with 300k+ unique data enteries.

3.2. Exploratory Data Analysis (EDA)

To understand the dataset's characteristics, we visualized the distribution of transactions across key attributes such as Country, Gender, Income, Product Category and Total Income. This provides an initial understanding of the dataset that guided feature selection and further analysis.

A correlation heatmap was created to analyze the relationships between different numerical variables to visualise correlation between the features of the dataset.

3.3. Data Preprocessing and Feature Engineering

Handling Missing Values:

- Checked for null values in each column.
- Imputed missing values using mode for categorical and mean for numerical columns.
- Removed rows with missing values based on business context if applicable.

Duplicate and Redundant Data Removal: Removed duplicate records and unnecessary columns to ensure data integrity.

Feature Engineering and Data Transformation: Generating RFM features:

- **Recency**: Days since last purchase.
- **Frequency**: Total number of purchases.
- **Monetary**: Total amount spent.
- Encoded categorical features (*Gender*, Income etc.) using label and one-hot encoding.
- Scaled features using *StandardScaler*.
- Normalized RFM values to the range [0, 1].

Feature Selection:

- Conducted correlation analysis to remove redundant features.
- Focused on RFM metrics for final clustering and segmentation.

Target Variable Creation: Segmented CLV into **Gold**, **Silver**, and **Bronze** categories based on *Total_Amount* for supervised learning.

Dataset Splitting: Divided data into training, testing, and validation sets for model building.

4. Methodology

Our methodology visually depicted in Figure 1.

4.1. Logistic Regression (With or without scaling)

Logistic regression is a statistical model used to predict the probability of an outcome based on one or more input features. In this model, we use the log loss function and set the learning rate to 0.01 for 100 epochs. Model training is performed using cross-validation, with or without scaling or normalization of the data. Loss: 0.67 Acc: 57.56

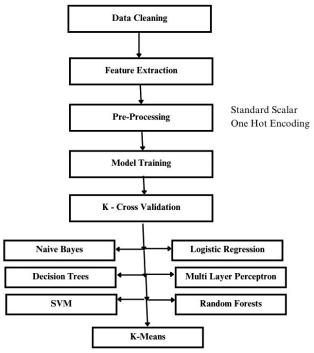

Figure 1. Flowchart

4.2. Gaussian Naive Bayes Classifier

A Gaussian Naive Bayes model was used to classify customer lifetime value (CLV) into High, Mid, and Low categories. Initial training without hyperparameter tuning indicated potential overfitting, prompting the use of GridSearchCV to optimize the var smoothing parameter. Although retraining improved performance, Naive Bayes may struggle with complex datasets, suggesting a need for more advanced algorithms or feature engineering. Loss: 0.73 Acc: 46.73

4.3. Decision Tree Classifier

A Decision Tree is a supervised learning algorithm used for both classification and regression tasks that splits data into branches based on feature values, creating a tree-like model to make decisions. We use GridSearchCV to find optimal max depth, max leaves, min sample split.

We use `GridSearchCV` to find the optimal hyperparameters for the Decision Tree, such as:

- `max_depth`: The maximum depth of the tree, which controls how deep the tree can grow.
- `max_leaf_nodes`: The maximum number of leaf nodes.
- `min_samples_split`: The minimum number of samples required to split an internal node.

Acc: 61.75 Loss: 0.66

4.4. Random Forest Classifier

The Random Forest model is an ensemble learning method(Bagging) that constructs multiple decision trees and merges their results to improve predictive accuracy and control overfitting. We use `GridSearchCV` to fine-tune hyperparameters by exploring the following ranges:

- `n_estimators` (number of trees): [50, 100, 200]
- `max_depth` (maximum depth of trees): [None, 10, 20, 30]
- `min_samples_split` (minimum samples required to split a node): [2, 5, 10]
- `min_samples_leaf` (minimum samples required to be at a leaf node): [1, 2, 4]

Loss: 0.19 Acc: 75.7

4.5. Multilayer Perceptron (MLP)

A Multilayer Perceptron (MLP) is a type of feedforward artificial neural network consisting of at least three layers: an input layer, hidden layers, and an output layer. We include three hidden layers with 64, 32, and 16 neurons, each using the ReLU activation function, which introduces non-linearity. The output layer uses softmax activation to handle multiclass classification. The model is trained with the Adam optimizer and uses categorical cross-entropy loss for multiclass classification. Weights are initialized using Glorot Uniform initialization.

Loss: 0.60 Acc: 69.45

4.6. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised learning model suitable for classification tasks. For the CLV prediction, an SVM with a radial basis function (RBF) kernel was employed to manage non-linear decision boundaries. The model was implemented using the Scikit-learn library, with hyperparameters (regularization and kernel) tuned through grid search. It was trained to optimize classification performance, yielding:

Loss: 0.63 Accuracy: 66.89

4.7. K-Means Clustering

K-means clustering was applied for customer segmentation based on standardized RFM (Recency, Frequency, Monetary) features. The optimal cluster count was determined using the Elbow Method and Silhouette Score, settling on three clusters:
Gold: High frequency, high monetary value, low recency.
Silver: Moderate values across metrics.
Bronze: Low frequency, low monetary value, high recency.

This segmentation informed targeted strategies for customer engagement.

Model	**Acc. (%)**	**Loss**
Random Forest Classifier	76.57	0.19
MLP	68.70	0.60
Decision Tree Classifier	62.16	0.66
Logistic Regression with Scaling	57.56	0.67
SVM	66.89	0.63
Naive Bayes Classifier	46.61	0.73

Table 1. Comparison of Model Performance

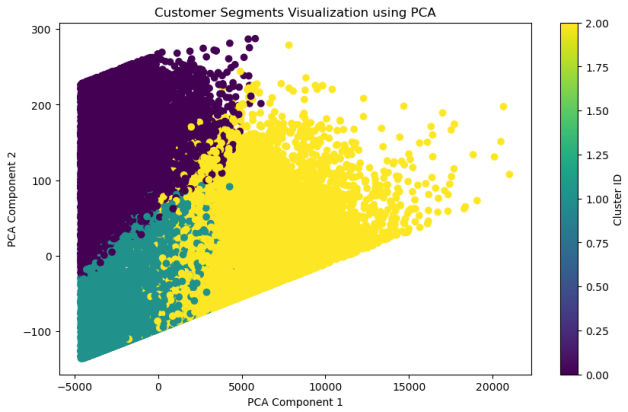

Figure 2. Clustering using K-Means Algo.

5. Results And Analysis

5.1. Results

The best-performing supervised model used in this study was the Random Forest (RF) model, achieving an accuracy of 76.57% and a loss of 0.19. The accuracies and losses for all other supervised models are summarized in Table 1, highlighting the comparative performance of the algorithms.

For the unsupervised analysis, we employed the K-means algorithm with RFM features (Recency, Frequency, and Monetary value) to determine customer lifetime value. The clustering results were evaluated using the Davies-Bouldin Score (0.94) and the Calinski-Harabasz Score (69412.07), which confirmed the effectiveness of the K-means approach for differentiating customer behaviors and identifying valuable customer clusters. These metrics demonstrate that K-means clustering was the most suitable method for segmenting customers based on their lifetime value.

The results for clustering customers can be seen in Figure 2.

5.2. Analysis

Analyzing our results based on metrics chosen (RFM):

Monetary:
Regular, Premium, and New customers show similar spending distributions, with no significant differences in median or interquartile range. This indicates that the monetary metric does not strongly differentiate between these segments.

Recency:
All three segments show similar distributions of recency, with slight variations in spread but comparable medians. No segment stands out in terms of purchase frequency over time.

Frequency:
Regular, Premium, and New customers again exhibit comparable frequency distributions. This suggests that customer segmentation (Regular, Premium, New) is not significantly influencing frequency behavior.

5.3. Key insights

The analysis reveals that Cluster_Id (Gold, Silver, Bronze) is more effective than Customer_Segment (Regular, Premium, New) in distinguishing customer behaviors. Clusters demonstrate clear differences across key metrics such as monetary value, recency, and frequency, with Gold customers consistently outperforming other clusters, followed by Silver and then Bronze.
Bronze Cluster customers show lower spending, higher recency (indicating less frequent purchases), and lower frequency, suggesting lower engagement levels. In contrast, Gold Cluster customers are the most valuable, characterized by high spending, frequent purchases, and low recency, making them a priority for retention strategies. On the other hand, Customer_Segment classifications fail to reveal distinct behavioral patterns, indicating potential overlap or insufficient differentiation criteria.

6. Conclusion

In this project, we gained hands-on experience with large datasets, mastering feature extraction, preprocessing, scaling, and normalization techniques. We implemented various ML algorithms, including logistic regression, Naive Bayes, decision trees, random forests, K-Means, and MLPs, each offering unique strengths for customer segmentation and CLV prediction. While MLPs and random forests excelled in accuracy, simpler models like logistic regression provided interpretability. Key challenges included managing data quality, optimizing complex models, and balancing interpretability with performance. This work sharpened our understanding of practical machine learning workflows and their application to real-world problems.

Contributions:

Rohit: Feature extraction, Implemented decision trees and kmeans, Worked on scaling and normalization.

Tarun: Preprocessing large datasets, Implemented Naive Bayes and logistic regression, Contributed to feature extraction.

Harshit: Scaling and normalization, Implemented logistic regression and SVM, Worked on random forests.

Aryan: Preprocessing large datasets, Implemented MLP, Contributed to random forest tuning.

7. References

1. Güner, M. (2021). *Retail Data Predictive Analysis Using Machine Learning Models*.

2. Pundir, A. K., Ganapathy, L., Maheshwari, P., & Kumar, M. N. Machine Learning for Revenue Forecasting: A Case Study in Retail Business.

3. Turkmen, B. Customer Segmentation With Machine Learning for Online Retail Industry.