IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

A Study On Customer Lifetime Value Prediction Using Machine Learning

Harini J and Dr. Suganthi P
School of Management
SASTRA Deemed University
SASTRA, Thanjavur-613 401

Abstract

The study examines demographic profiles and factors affecting customer relationships among Spotify premium subscribers using machine learning methods called BG/NBD Model and Gamma-Gamma Model by predicting Customer lifetime value in music platform named Spotify. It reveals that smaller consumer bases contribute more due to dissatisfaction with restricted services, while larger bases are less generous. Students are a significant customer base, prioritizing entertainment expenditure. The study suggests adjusting marketing and pricing strategies to attract and retain students, monitoring user ratios, and offering incentives for upgrades. It emphasizes the importance of demographic targeting, post-restriction analysis, and proactive issue resolution for optimizing customer relationships and revenue generation.

Keyword: Machine Learning, BG/NBD Model, Gamma-Gamma Model, Customer relationships, Customer lifetime value.

I. INTRODUCTION

Music platforms, or streaming services, have revolutionized music listening by providing instant access to vast collections of songs, albums, and playlists. Spotify, a popular global platform, offers both free and premium subscription options, including ad-free listening, offline downloads, and higher audio quality, making it a popular choice for music lovers.

Customer Lifetime Value (CLV) is a company's net profit from a customer's relationship, considering purchases, repeat purchases, referrals, and costs. It helps businesses understand marketing and retention strategies, allocate resources efficiently, focus on high-value customers, and prioritize customer retention efforts.

II. LITERATURE REVIEW

This paper explores customer lifetime value measurement literature, presents a contextual framework, reviews prominent models, discusses strengths and weaknesses, and identifies key issues for further literature advancement**.** Singh, S. S., & Jain, D. C. (2013).

E-commerce, the online buying and selling of products, has created numerous employment opportunities and enabled retailers to analyze customer relationships. This work focuses on predicting customer lifetime value (CLV) using the Beta-Geometric/Negative Binomial Distribution Model (BG\NBD) and Gamma-Gamma. Yashaswini, H. S., & Prabhudeva, S. (2022).

The Pareto/NBD model has gained popularity for its accurate prediction of customer purchase frequency. This paper analyzes its disadvantages, proposes an improved model that incorporates customer personalized information, and tests its superiority using real sales data from a Dalian Mall. It also introduces a segmentation method based on customer lifetime value and management strategy. Guo, Y., et al. (2013).

The study proposes an alternative to the Pareto/NBD model for valuing customers, focusing on the dependence between transaction number and profitability. This approach increases the accuracy of predicting customer lifetime value (CLV), as demonstrated in a new empirical case from the retail banking sector. This approach helps predict future profits and customer transactions. Glady, et al. (2009).

This paper explores the use of AI algorithms in predicting and optimizing Customer Life Cycle (CLV), enabling better resource allocation and personalized retention strategies. By analyzing customer data, AI can predict CLV accurately, enabling businesses to tailor marketing and retention efforts, thereby maximizing ROI and long-term profitability.

Blessing et al. (2023).

The civil aviation industry's competitiveness is largely determined by high-quality customer resources. Research on customer lifetime value helps identify high-value, medium-value, and low-value travellers, enabling rational resource allocation. However, models for calculating customer life value remain controversial. The author proposes an optimized China Eastern Airlines passenger network value assessment model, combining it with the TravelSky value score to identify significant customers. Chen, S. (2018).

A study using machine learning has improved Customer Relationship Management (CRM) by enhancing CLV predictions. The research, conducted using advanced machine learning tools, found that machine learning significantly outperforms traditional CLV estimation methods, leading to better accuracy and improved resource allocation. This highlights the importance of CLV in crafting personalized customer journeys and refining CRM strategies. Kumar, A., et al. (2023).

Customer Lifetime Value (CLTV) is a crucial metric for businesses to monitor, as it helps make informed decisions about investing in acquiring and retaining customers, and helps build marketing strategies with a positive ROI. Aigerim Shopenova. (2021).

This paper focuses on customer lifetime value measurement and segmentation using customer lifecycle value theory. It conducts feature engineering, machine learning algorithms, and relationship management analysis models to construct a customer value segmentation identification model. The paper also conducts empirical analysis with real customer transaction data from an online shopping platform to verify the validity and applicability of the proposed methods. Sun, Y., et al. (2023).

III. METHODOLOGY

A. Problem Definition

The study uses machine learning algorithms with Lifetime package in Python to predict the lifetime value of Spotify users, focusing on their value, overall customer segment contribution, customer relationships, and future predictions and their overall perception towards Spotify using sentimental analysis.

B. Scope

Customer Lifetime Value (CLV) is a metric that evaluates the total value a customer brings to a business over their entire relationship, aiding in optimizing customer acquisition, retention, and loyalty strategies, thereby boosting profitability and sustainable growth.

C. Proposed Methodology

1. Data Collection

This study is based on real-time data which is collected as tentative sample size of 385(Cochran (1977), Mark (2005) and Singh and Chaudhury (1985)) of moving population using non-probability purposive sampling. The aim of the calculation is to determine an adequate sample size which can estimate results for the whole population with a good precision.

Formula: $((1.96)^2 (0.5) (0.5)) / (0.05)^2 = 385$

The dataset consists of 385 observations with 6 attributes. The study focuses on Spotify premium user behavior, and this it also includes factors such as demographics, transaction data, and their overall satisfaction data.

2. Tools used

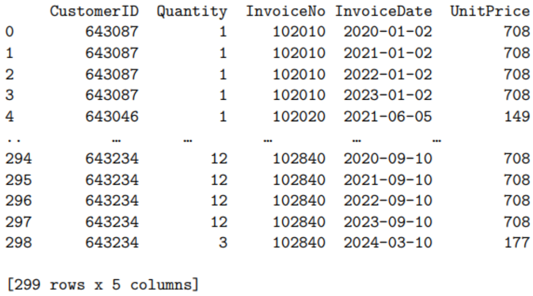
Python-Lifetime packages:

A "lifetime package" is a one-time payment for access to a product or service, such as software licenses, subscription-based services, or memberships, providing ongoing access without the need for recurring payments, ensuring the product or service's existence or lifetime of the purchaser.

D. Proposed Algorithm

1)Customer Lifetime Value using Lifetime package

Table. 1 Dataset of premium users



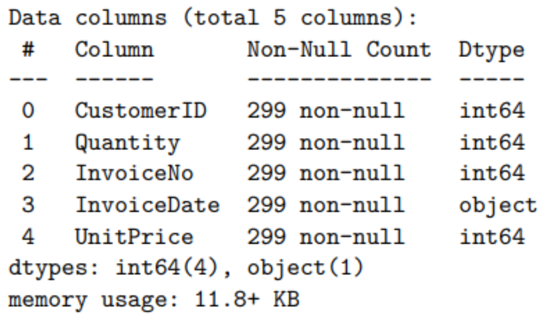
```
     CustomerID  Quantity  InvoiceNo InvoiceDate  UnitPrice
0        643087         1     102010  2020-01-02        708
1        643087         1     102010  2021-01-02        708
2        643087         1     102010  2022-01-02        708
3        643087         1     102010  2023-01-02        708
4        643046         1     102020  2021-06-05        149
..          ...       ...        ...         ...        ...
294      643234        12     102840  2020-09-10        708
295      643234        12     102840  2021-09-10        708
296      643234        12     102840  2022-09-10        708
297      643234        12     102840  2023-09-10        708
298      643234         3     102840  2024-03-10        177

[299 rows x 5 columns]
```

(Source: Primary data)

The dataset contains details of purchase data of Spotify users with Column name Customer ID- identifies the unique customer, Quantity- number of items purchased by the user, Invoice No- a code for the unique products, Invoice Date- the date where the customers purchase the product(subscribed), Unit Price- the amount of each product.

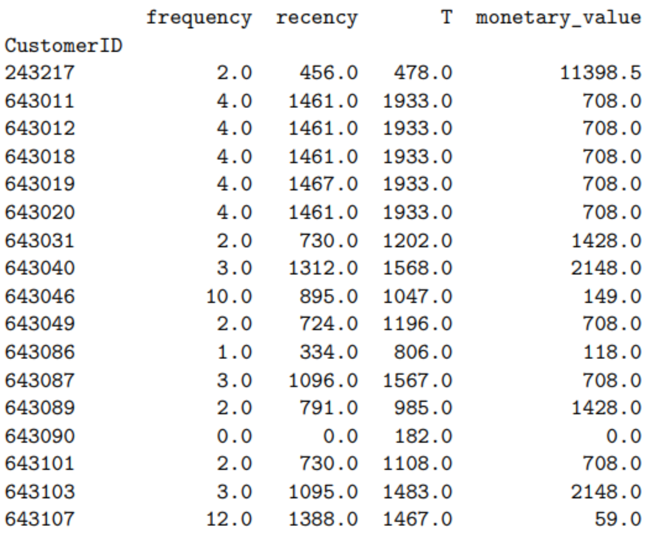
Table. 2 Identification of datatypes



```
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   CustomerID   299 non-null    int64 
 1   Quantity     299 non-null    int64 
 2   InvoiceNo    299 non-null    int64 
 3   InvoiceDate  299 non-null    object
 4   UnitPrice    299 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 11.8+ KB
```

(Source: Primary data)

Table. 3 Computation of summary data using lifetime method



```
            frequency  recency       T  monetary_value
CustomerID
243217            2.0    456.0   478.0         11398.5
643011            4.0   1461.0  1933.0           708.0
643012            4.0   1461.0  1933.0           708.0
643018            4.0   1461.0  1933.0           708.0
643019            4.0   1467.0  1933.0           708.0
643020            4.0   1461.0  1933.0           708.0
643031            2.0    730.0  1202.0          1428.0
643040            3.0   1312.0  1568.0          2148.0
643046           10.0    895.0  1047.0           149.0
643049            2.0    724.0  1196.0           708.0
643086            1.0    334.0   806.0           118.0
643087            3.0   1096.0  1567.0           708.0
643089            2.0    791.0   985.0          1428.0
643090            0.0      0.0   182.0             0.0
643101            2.0    730.0  1108.0           708.0
643103            3.0   1095.0  1483.0          2148.0
643107           12.0   1388.0  1467.0            59.0
```

(Source: Primary data)

The recency variable is a measure of the time elapsed since the customer’s last purchase.

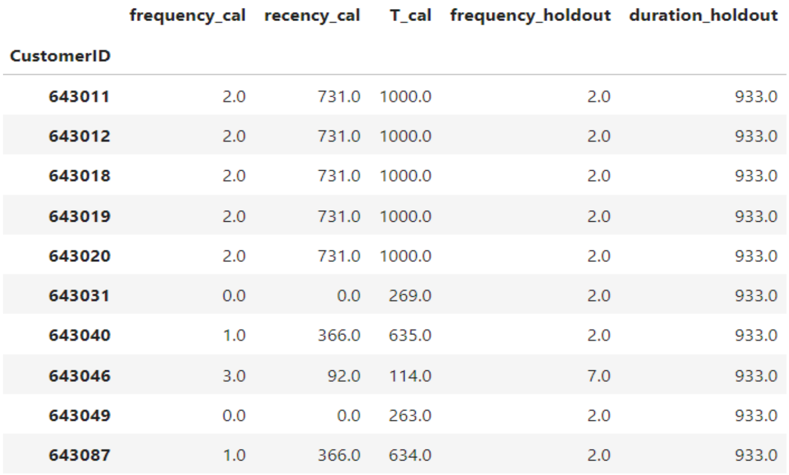
The frequency corresponds to the number of orders placed by a customer.

The monetary metric shows the total revenue generated by each customer.

Tenure, or T, represents how long a given customer has been with the company.

Fitting and testing the BG/NBD model.

Table.4 Split data into frequency_holdout and duration_holdout

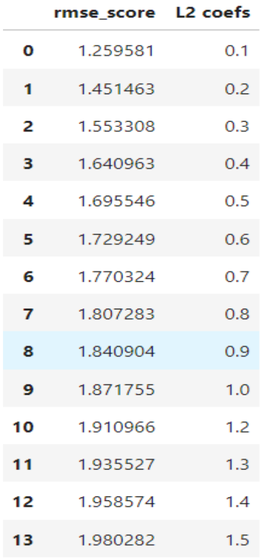


	frequency_cal	recency_cal	T_cal	frequency_holdout	duration_holdout
CustomerID					
643011	2.0	731.0	1000.0	2.0	933.0
643012	2.0	731.0	1000.0	2.0	933.0
643018	2.0	731.0	1000.0	2.0	933.0
643019	2.0	731.0	1000.0	2.0	933.0
643020	2.0	731.0	1000.0	2.0	933.0
643031	0.0	0.0	269.0	2.0	933.0
643040	1.0	366.0	635.0	2.0	933.0
643046	3.0	92.0	114.0	7.0	933.0
643049	0.0	0.0	263.0	2.0	933.0
643087	1.0	366.0	634.0	2.0	933.0

(Source: Primary data)

The dataset, with a maximum of 1933 days, has been split using the Lifetimes method for calibration and holdout data, with 1000 data for model fit and 933 data for testing.
The BG/NBD model employs L2 regularization, and the optimal L2 coefficient is determined through the "Grid Search."

Table.5 L2 regularization



	rmse_score	L2 coefs
0	1.259581	0.1
1	1.451463	0.2
2	1.553308	0.3
3	1.640963	0.4
4	1.695546	0.5
5	1.729249	0.6
6	1.770324	0.7
7	1.807283	0.8
8	1.840904	0.9
9	1.871755	1.0
10	1.910966	1.2
11	1.935527	1.3
12	1.958574	1.4
13	1.980282	1.5

(Source: Primary data)

The model was fitted with the optimal L2 coefficient of 0.10.

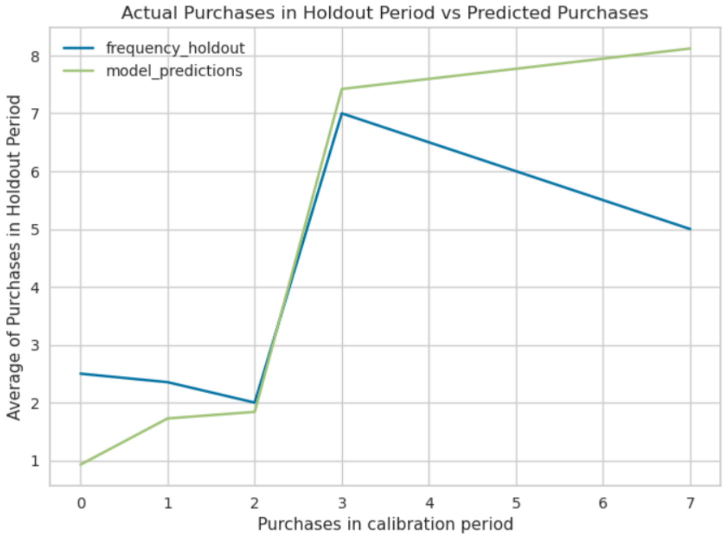
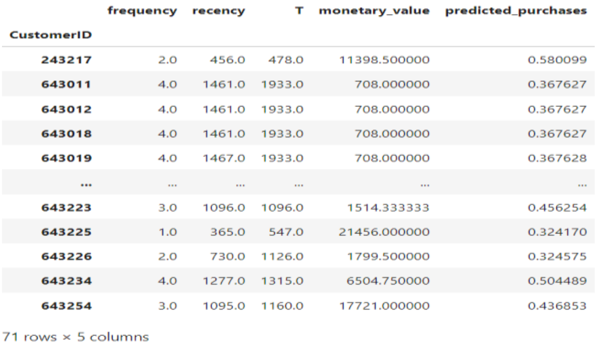


Figure. 1 Actual Purchases in Holdout period vs Predicted Purchases

The plot indicates that the model effectively predicts frequencies in the holdout dataset.

(Source: Primary data)

Table. 6 Predicting the next six months CLV for each customer.



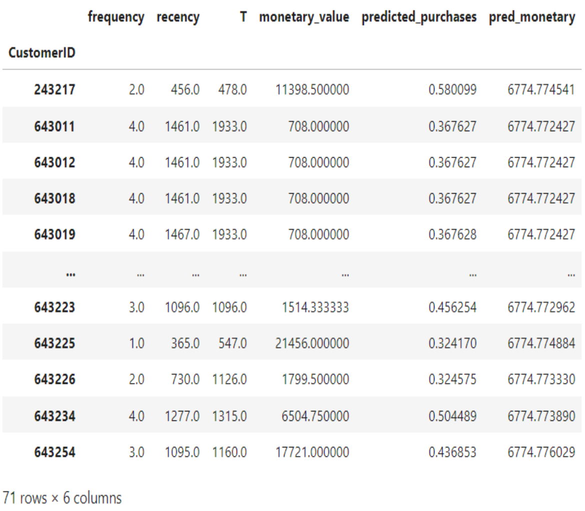
	frequency	recency	T	monetary_value	predicted_purchases
CustomerID					
243217	2.0	456.0	478.0	11398.500000	0.580099
643011	4.0	1461.0	1933.0	708.000000	0.367627
643012	4.0	1461.0	1933.0	708.000000	0.367627
643018	4.0	1461.0	1933.0	708.000000	0.367627
643019	4.0	1467.0	1933.0	708.000000	0.367628
...	...	...	...	...	...
643223	3.0	1096.0	1096.0	1514.333333	0.456254
643225	1.0	365.0	547.0	21456.000000	0.324170
643226	2.0	730.0	1126.0	1799.500000	0.324575
643234	4.0	1277.0	1315.0	6504.750000	0.504489
643254	3.0	1095.0	1160.0	17721.000000	0.436853

71 rows × 5 columns

(Source: Primary data)

The model was fitted and the number of customer purchases predicted for the next six months was determined using the "conditional_expected_number_of_purchases_up_to_time" method.

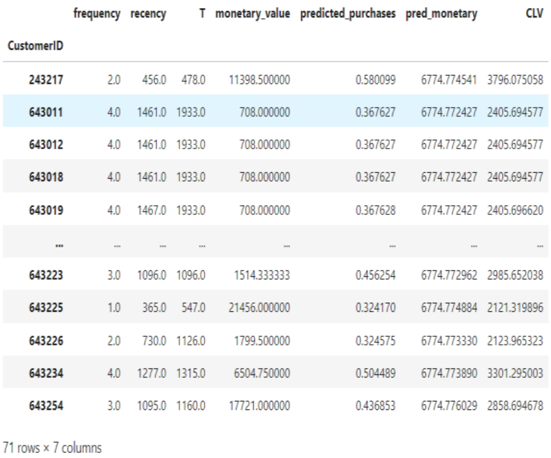
Table. 7 Fitting the Gamma-Gamma Model



	frequency	recency	T	monetary_value	predicted_purchases	pred_monetary
CustomerID						
243217	2.0	456.0	478.0	11398.500000	0.580099	6774.774541
643011	4.0	1461.0	1933.0	708.000000	0.367627	6774.772427
643012	4.0	1461.0	1933.0	708.000000	0.367627	6774.772427
643018	4.0	1461.0	1933.0	708.000000	0.367627	6774.772427
643019	4.0	1467.0	1933.0	708.000000	0.367628	6774.772427
...	...	...	...	...	...	...
643223	3.0	1096.0	1096.0	1514.333333	0.456254	6774.772962
643225	1.0	365.0	547.0	21456.000000	0.324170	6774.774884
643226	2.0	730.0	1126.0	1799.500000	0.324575	6774.773330
643234	4.0	1277.0	1315.0	6504.750000	0.504489	6774.773890
643254	3.0	1095.0	1160.0	17721.000000	0.436853	6774.776029

71 rows × 6 columns

Table. 8 Computation of CLV for next six months



	frequency	recency	T	monetary_value	predicted_purchases	pred_monetary	CLV
CustomerID							
243217	2.0	456.0	478.0	11398.500000	0.580099	6774.774541	3796.075058
643011	4.0	1461.0	1933.0	708.000000	0.367627	6774.772427	2405.694577
643012	4.0	1461.0	1933.0	708.000000	0.367627	6774.772427	2405.694577
643018	4.0	1461.0	1933.0	708.000000	0.367627	6774.772427	2405.694577
643019	4.0	1467.0	1933.0	708.000000	0.367628	6774.772427	2405.696620
...	...	...	...	...	...	...	...
643223	3.0	1096.0	1096.0	1514.333333	0.456254	6774.772962	2985.652038
643225	1.0	365.0	547.0	21456.000000	0.324170	6774.774884	2121.319896
643226	2.0	730.0	1126.0	1799.500000	0.324575	6774.773330	2123.965323
643234	4.0	1277.0	1315.0	6504.750000	0.504489	6774.773890	3301.295003
643254	3.0	1095.0	1160.0	17721.000000	0.436853	6774.776029	2858.694678

71 rows × 7 columns

(Source: Primary data)

The Customer Lifetime Value (CLV) for each customer is calculated for the next six months using the method "customer_lifetime_value".

2) Customer segmentation using the KMeans algorithm.

The customer clustering algorithm KMeans is used to determine the optimal number of clusters for the dataset, which is achieved using the "Elbow" method, implemented using the "KElbowVisualizer" from the Yellowbrick library.

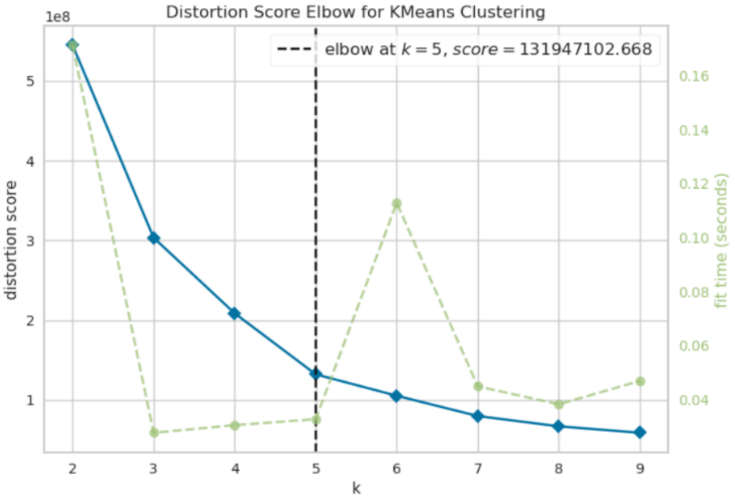
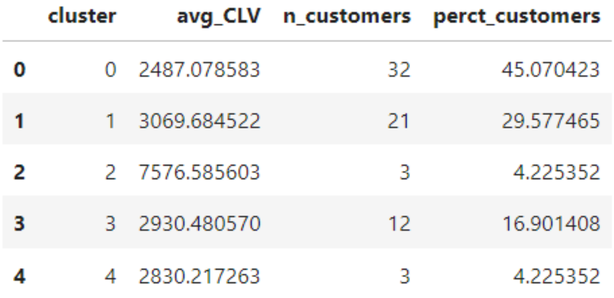


Figure 2. Distortion Score Elbow for KMeans Clustering

(Source: Primary data)

The optimal number of clusters, or elbow value, is 5. The distortion score is determined by the sum of squared errors, where the error is the distance between a point and its assigned centroid. The KMeans model can now be fitted with this optimal number of clusters.

Table. 9 Average cluster values and customer base



	cluster	avg_CLV	n_customers	perct_customers
0	0	2487.078583	32	45.070423
1	1	3069.684522	21	29.577465
2	2	7576.585603	3	4.225352
3	3	2930.480570	12	16.901408
4	4	2830.217263	3	4.225352

(Source: Primary data)

The table shows that cluster 2, comprising 4.225% of the customer base, has the highest average CLV of ₹7576.58. Cluster 1 follows with an average CLV of ₹3069.68. Cluster 3 follows with an average CLV of ₹2930.48, followed by cluster 4 with an average CLV of ₹2830.21, and finally cluster 0 with an average CLV of ₹2487.07 (45% of the customer base). The clusters should be renamed "Diamond", "Platinum", "Gold", "Silver", and "Bronze" respectively.

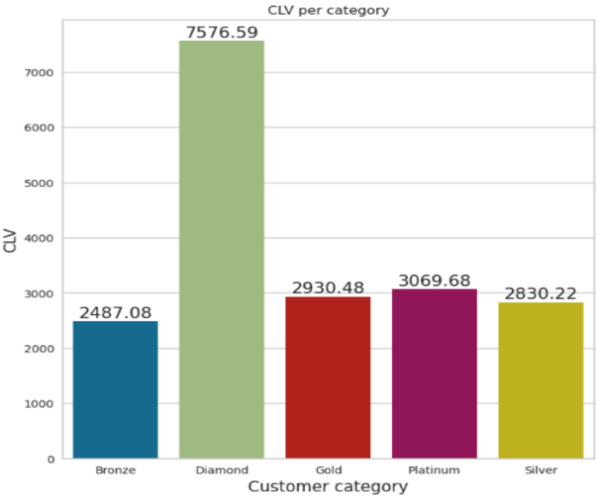


Figure.3 CLV per category

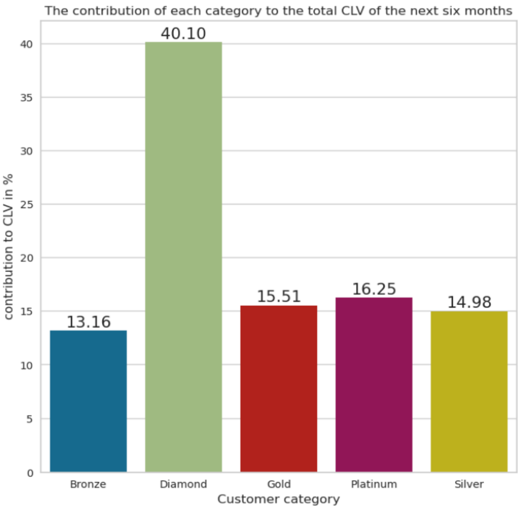


Figure. 4 The contribution of each category to the total CLV of the next six months

(Source: Primary data)

The bar plot reveals that "Diamond" and "Platinum" customers contribute over 56% of the total CLV for the next six months, with a 33% customer base, while "Bronze" customers contribute only 13%, despite representing 45% of the customer base.

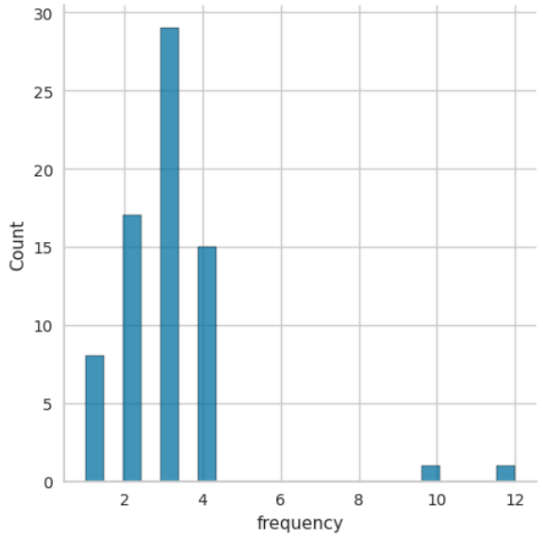


Figure. 5 Analyzing Frequency Metric

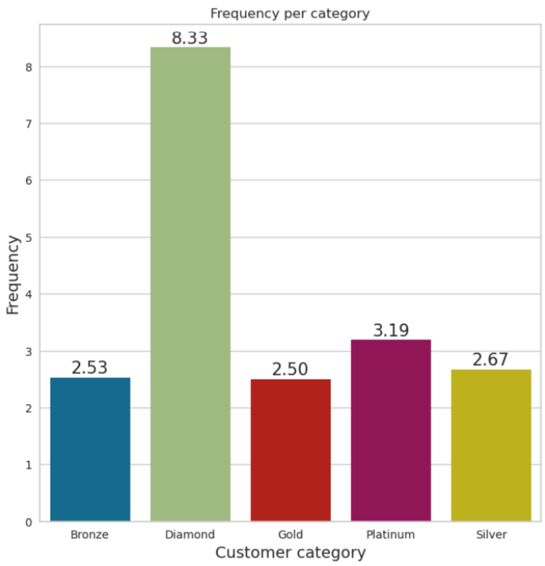


Figure. 6 Frequency per category

(Source: Primary data)

The data indicates that "Diamond" customers exhibit the highest mean frequency.

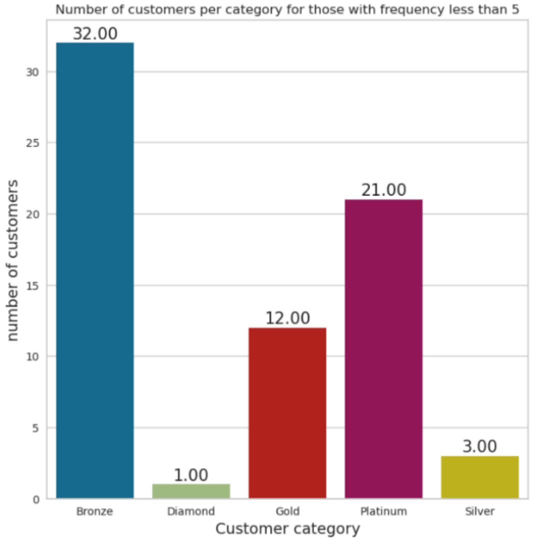


Figure. 7 Frequency less than 5

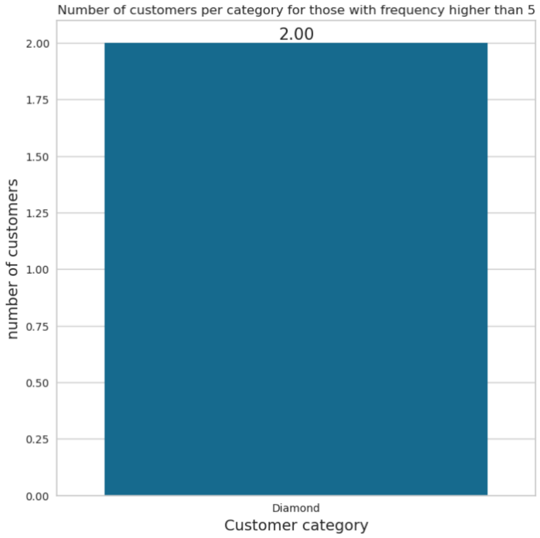
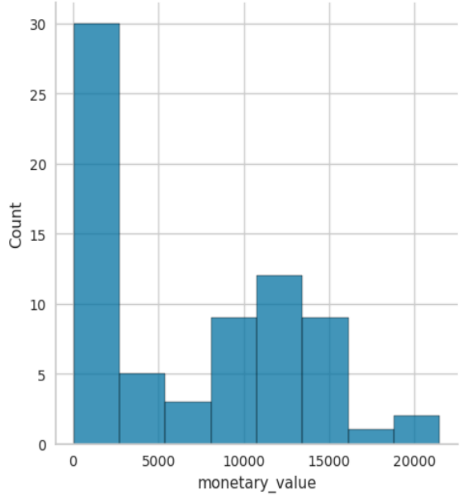


Figure. 8 Frequency higher than 5

(Source: Primary data)

The frequency variable distribution is skewed towards the left, with 90% of customers having a frequency less than 5. No "Diamond" customers have a frequency less than 5, indicating that all "Diamond" customers made more than 5 purchases during the given period, with an average of 8 purchases. (Fig. 6 Frequency per category)



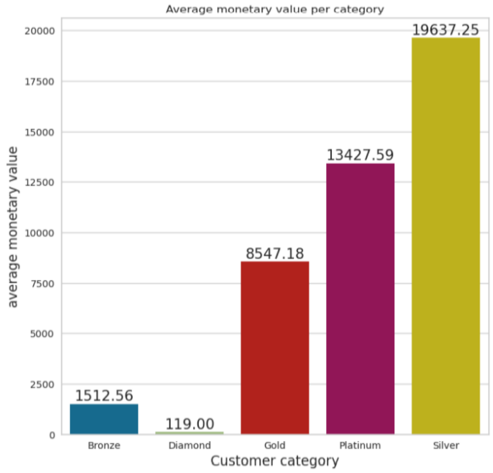
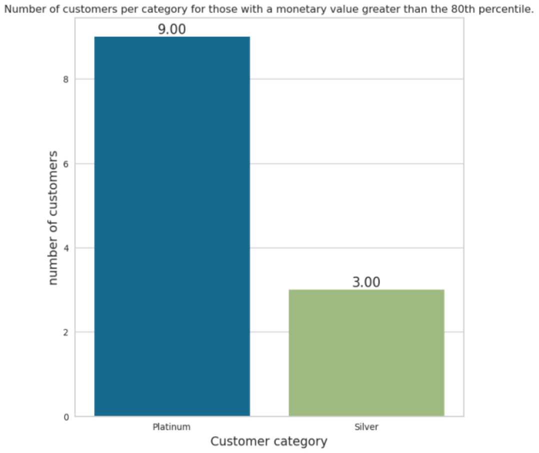


Figure. 9 Analyzing monetary value **Figure. 10 Average monetary value per category**

(Source: Primary data)



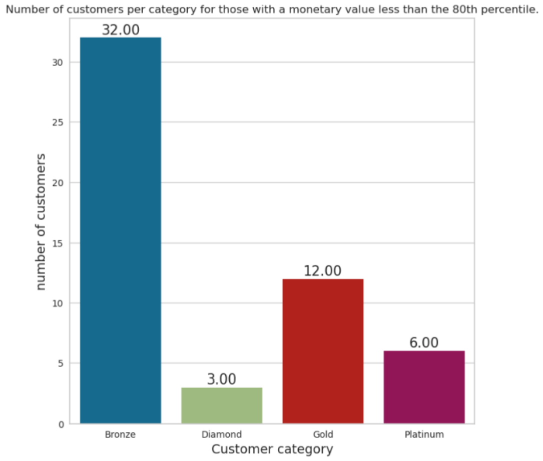


Figure. 11 Value greater than 80th **Figure. 12 Monetary value less than percentile**

80th percentile

(Source: Primary data)

The distribution plot shows a distribution highly skewed towards the left, meaning that a large number of customers spent small amounts of money and few customers spent relatively high amounts of money during the given period.

3) Sentimental Analysis

Sentiment analysis is a method that uses natural language processing and machine learning to understand the emotional tone of text, such as reviews or customer feedback. It is used in various applications like market research, customer service, and social media monitoring to gauge public opinion, customer satisfaction, and brand perception.

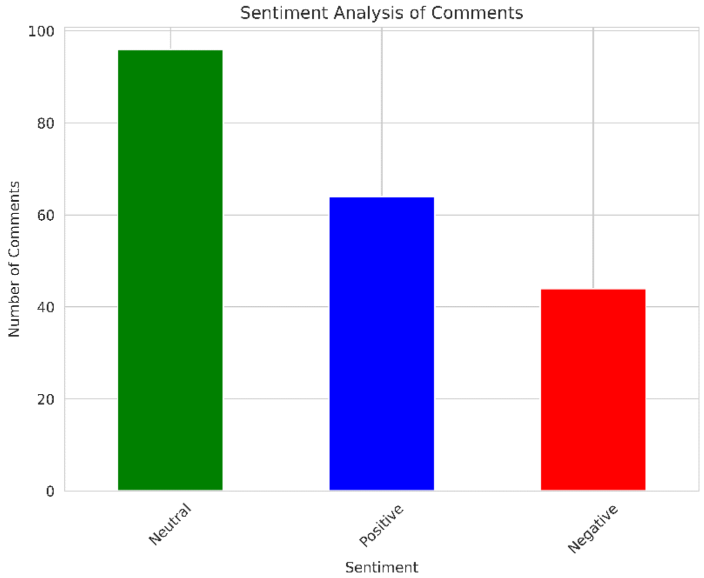


Figure. 13 Sentimental Analysis of comments

(Source: Primary data)

The visualization shows the distribution of comments, categorizing them into positive, neutral, and negative sentiments based on their polarity.

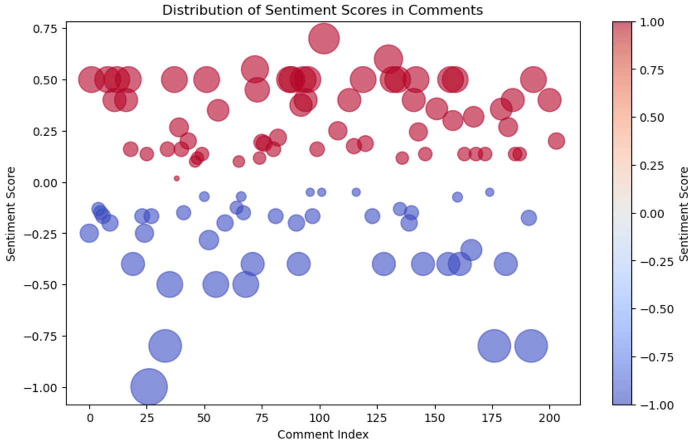


Figure. 14 Distribution of Sentimental Scores in comments

(Source: Primary data)

The sentiment score is represented by bubbles, with larger bubbles indicating stronger sentiments. The color indicates sentiment polarity, with warmer red indicating positive sentiments and cooler blue indicating negative ones. The chart shows how sentiments are spread across comments, with the x-axis representing the index and the y-axis showing the sentiment score.

IV. FINDINGS

Based on data research of Spotify premium subscribers, it can be observed that consumers with a smaller base make larger contributions to the company. This may be attributed to their discontent with the restricted services provided and their irritation with interruptions and advertisements. The majority of these customers work and earn between ₹100,000 and ₹300,000 annually. Larger customer bases tend to be less generous with their contributions since they believe the services are overpriced and unreliable. Frequently, they choose freemium services or go to alternative providers. These customers have stopped using Spotify because they are irritated by its limitations. Students make up the bulk of Spotify customers; they prioritise budgeting and entertainment expenditures.

V. SUGGESTION

Spotify needs to adapt its marketing and pricing strategies to attract and retain students, as they make up a significant portion of its customer base. Offering discounts or packages can help retain this demographic. Conducting research on customer satisfaction and loyalty can help devise strategies to address issues. Monitoring the ratio of free and paid users can help gauge the effectiveness of marketing campaigns and pricing structures. If there's a decline in satisfaction or loyalty, offering incentives or promotions can encourage users to upgrade to paid subscriptions. Analysing feedback and sentiment data can help identify pain points and address them proactively.

VI. CONCLUSION

The study reveals that demographic profiles significantly impact consumer contributions, with smaller consumer bases contributing more due to dissatisfaction with restricted services and interruptions. These consumers, earning between ₹100,000 and ₹300,000 annually, are more willing to invest in premium services. Factors like service restrictions and reliability also influence customer relationships. Consumers, particularly those with larger bases, are more likely to opt for freemium services or switch to alternative providers when they perceive services as overpriced or unreliable. Understanding the proportion of free and paid users is crucial for devising strategies to attract and retain paying customers. Post-restriction analysis of Average Revenue per User (ARPU), customer satisfaction, and loyalty can guide pricing strategies and service offerings. Targeting marketing efforts towards specific demographic groups, such as students, can capitalize on their entertainment expenditure priorities.

References

[1]Singh, S. S., & Jain, D. C. (2013). Measuring customer lifetime Value: Models and analysis. Social Science Research Network. https://doi.org/10.2139/ssrn.2214860

[2] Yashaswini, H. S., & Prabhudeva, S. (2022). Customer lifetime value prediction. International Journal of Advanced Research in Science, Communication and Technology, 805–810. https://doi.org/10.48175/ijarsct-5162

[3] Guo, Y., Wang, H., & Liu, W. (2013). Improved Pareto/NBD Model and Its Applications in Customer Segmentation based on Personal Information Combination. International Journal of Database Theory and Application, 6(5), 175–186. https://doi.org/10.14257/ijdta.2013.6.5.16

[4] Glady, N., Baesens, B., & Croux, C. (2009). A modified Pareto/NBD approach for predicting customer lifetime value. Expert Systems with Applications, 36(2), 2062–2071. https://doi.org/10.1016/j.eswa.2007.12.049

[5] Blessing, Elisha & Klaus, Hubert. (2023). Discuss how AI algorithms predict and optimize the customer lifetime value, enabling better resource allocation and personalized retention strategies. 3329. 12.

[6] Chen, S. (2018). Estimating Customer Lifetime Value Using Machine Learning Techniques. In Estimating Customer Lifetime Value Using Machine Learning Techniques (pp. 1–19). https://doi.org/10.5772/intechopen.76990

[7] Kumar, A., Singh, K. U., Kumar, G., Choudhury, T., & Kotecha, K. (2023). Customer Lifetime Value Prediction: Using Machine Learning to Forecast CLV and Enhance Customer Relationship Management. IEEE Xplore. https://doi.org/10.1109/ismsit58785.2023.10304958

[8] Aigerim Shopenova. (2021). Predict Customer Lifetime Value with Machine Learning. The Startup.

[9] Sun, Y., Liu, H., & Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. Heliyon, 9(2), e13384. https://doi.org/10.1016/j.heliyon.2023.e13384