THE CHINESE UNIVERSITY OF HONG KONG
STAT3011 Workshop on Data Analysis and Statistical Computing
Project II Proposal of the additional statistical or computation methods

Our team

Group 1

Topic

Chapter 40: Determining Your Most Valuable Customer (customer lifetime value)

Purpose of the original solution

Determining the Most Valuable Customer by predicting the future 3 month of CLV of customer. We would keep this purpose in our extension solution.

Definition of customer lifetime value (CLV)

In the original solution, CLV is defined as the total sales from a customer over a specific future period, based on their Monetary Value and Purchase Frequency. In our extended solution, we redefine CLV to also consider Recency.

Motivation / Weakness of the original solution

Linear regression in original solution may not work well to predict the 3 month of CLV, the reason are as follows:

1. Multicollinearity: Sales average (e.g. sales_avg_M_3) is directly related to sales sum and sales count. Specifically, sales_avg is calculated as sales sum / sales count.

	feature	coef
0	sales_avg_M_2	-0.074322
1	sales_avg_M_3	-0.210439
2	sales_avg_M_4	1.068602
3	sales_avg_M_5	-0.883733
4	sales_count_M_2	50.762158
5	sales_count_M_3	35.165081
6	sales_count_M_4	33.589453
7	sales_count_M_5	-56.795327
8	sales_sum_M_2	0.544971
9	sales_sum_M_3	0.160771
10	sales_sum_M_4	-0.404613
11	sales_sum_M_5	1.043029

2. Recency is Not Explicitly Accounted: Linear regression does not directly consider recency , which is important for predicting future CLV.

3. The rolling window method may used to predict the Customer Lifetime Value (CLV) for the next 3 months. However, this method assumes that the data is stationary, In practice, this assumption may not always hold.

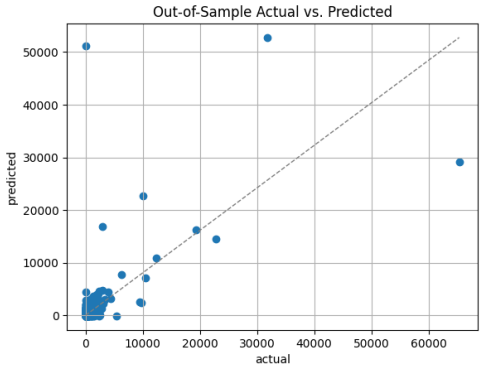
Also While the original solution successfully trains and evaluates the linear regression model, it does not use the trained model to predict the Customer Lifetime Value (CLV) for the upcoming 3 months, also have not determined the Most Valuable Customer.

Introduction of the Extended Solution

In this section, we propose an extension to the original solution by comparing the Linear Regression model with a more advanced BG/NBD model with Gamma-Gamma. We will examine their performance by plotting the actual vs. predicted values and analyzing the differences. Additionally, we introduce the use of K-Means Clustering to identify the most valuable customer segments based on their predicted Customer Lifetime Value (CLV).

1. Gamma-Gamma with BG/NBD Model vs. Linear Regression in the Original Solution:
Original Solution (Linear Regression):

- In the original solution, 30% of the data is used to generate the actual vs. predicted plot. This data helps evaluate how well the linear regression model predicts the (CLV).



Gamma-Gamma with BG/NBD Model:

- For the Gamma-Gamma with BG/NBD model, we will use the same 30% of the data that was used in the original solution to ensure a fair comparison.
- The goal is to predict the CLV for the most recent 3-month period, as the actual data for this period is already known. This allows us to compare the predicted CLV with the actual CLV.

Comparison:

- We will compare the actual vs. predicted plots for both models
- Since the recent 3-month data is known, we can directly compare how well each model predicted the CLV for this period.
- The Gamma-Gamma with BG/NBD model is expected to perform better because it considers Recency, Frequency, and Monetary Value.

2. K-Means Clustering:
In this section, we use K-Means Clustering to segment customers based on their **Conditional expected average profit** and **Predicted purchases**. The expected result may classify as follows:

- Group 1: High Frequency, High Value (Most Valuable Customers)
- Group 2: Low Frequency, High Value (Needs Increased Repurchase Rate)
- Group 3: High Frequency, Low Value (Needs to Increase Average Order Value)
- Group 4: Low Frequency, Low Value (Potential for Churn)

Group 1 will be identified as our **Most Valuable Customers** group.

Summary of three model

	Model	Features	Response	CLV prediction
1.	**Linear regression in original solution**	- sales avg Mi - sales count Mi - sales sum Mi i= 2,3,4,5	Sales sum Mi	Sales sum M0 (future 3 month)
2.	**Gamma-Gamma**	-frequency -monetary value	Conditional expected average profit	Conditional expected average profit × Predicted purchases
3.	**BG/NBD model**	-frequency -recency -T	Predicted purchases	

Concept explains

Mi	Represent different time periods (e.g., the 2nd, 3rd, 4th, and 5th 3-month periods). M1: most recent 3-month period**.** M0: future 3 month
Sales sum	The total sales within a specific time period.
Sales avg	The average transaction value within a specific time period.
Sales count	The number of transactions within a specific time period.
Frequency	The number of repeat purchases made by a customer during the observation period.
Monetary value	The average transaction value of a customer, calculated as the total amount spent divided by the number of transactions
Recency	The time since the customer's last purchase, measured from their first purchase to their most recent purchase. *Example*: If a customer's first purchase was on January 1, 2023, and their most recent purchase was on June 1, 2023, then recency = 5 months.
T	The customer's age, measured from their first purchase to the end of the observation period. *Example*: If a customer's first purchase was on January 1, 2023, and the observation period ended on December 31, 2023, then T = 12 months.
Conditional expected average profit	The conditional expected average transaction value of a customer.
Predicted purchases	The expected number of future transactions for a customer within a given time period.

Conclusion

the proposed extended solution addresses the limitations of the original linear regression model by introducing more advanced statistical methods. The BG/NBD and Gamma-Gamma models improve the prediction of Customer Lifetime Value (CLV). Additionally, K-Means Clustering helps identify the most valuable customer segments based on their predicted CLV. These methods not only enhance the accuracy of CLV predictions but also provide actionable insights for customer segmentation and targeted marketing strategies.

Archieve

4. Method

For doing the prediction, we separate all the transaction data into 6 groups (M1-M6) by their transaction date. 3 months as one interval and M1 will contain the closest 3-month transaction records and M6 will contain the oldest 3-month transaction records.
In order to find out which models give the best results, we are going to use RSS to check if the model gives the best results.

4.1 linear regression model

For linear regression model, we are going to train the model by using transaction record from M2 to M6 for predicting the average transaction amount of M1. Then using the true average transaction amount of M1 and the one we predict to find the RSS.

Since the customer lifetime value is based on the average transaction amount and the transaction count, we also need to predict the future transaction frequency of each customer. For frequency prediction, we are going to use BG/NBD model to do the prediction.

5. Conclusion

At the end, by using the data from predicting the average transaction amounts which gives the best results and the frequency prediction by using BG/NBD model, we can find out the customer with the highest customer lifetime value.

Not submitted

Allocation of work for report and presentation

Introduction (Choi Chit, Chong Chun Hin)
- Problem statement and background
- Importance of Identifying Valuable Customers
- Description of Dataset

Original Analysis (Tang Yu Ching, Ng Yuk)
- Data Preprocessing Steps
- Code Blocks for Original Analysis
- Statistical and Computational Methods Applied
- Strengths and Limitations of the Analysis

Extended Analysis 1 (Chong Chun Hin, Choi Chit, Tang Yu Ching, Ng Yuk)
- Additional Statistical and Computational Methods Applied
- Justification for Chosen Methods

- Comparison with Original Analysis
- Strengths and Limitations of the Analysis

Extended Analysis 2 (Wu Tong, Zhang Yuchen, Jiang Xinwen)
- Additional Statistical and Computational Methods Applied
- Justification for Chosen Methods
- Comparison with Original Analysis
- Strengths and Limitations of the Analysis

Results and Visualization (All members of our group)
- Key Findings
- Visualizations of Results
- Usage for real world

Conclusion (All members of our group)
- Summary of Findings
- Recommendations for Future Analysis