95-851 /451 Making Products Count: Data Science for Product Managers

HW1: Customer Lifetime Value – Fall 2023

Due September 13, 2023

Background: Customer Lifetime Value

In marketing, customer lifetime value (CLV or often CLTV), lifetime customer value (LCV), or life- time value (LTV) is a prediction of the net profit attributed to the entire future relationship with a customer. Customer lifetime value can also be defined as the dollar value of a customer relationship, based on the present value of the projected future cash flows from the customer relationship. Customer lifetime value is an important concept in that it encourages firms to shift their focus from quarterly profits to the long-term health of their customer relationships. Customer lifetime value is an important number because it represents an upper limit on spending to acquire new customers. For this reason, it is an important element in calculating payback of advertising spent in marketing mix modeling.[1] In this assignment, you will have the opportunity to practice estimating CLV using Python.

Dataset

The dataset is provided in CSV format in the "customer_sales.csv" and contains roughly 8000 transactions records representing sales records over a couple of years. Each row in the dataset represent a single transaction. There are 17 columns in the dataset:

Order ID	A unique ID given to each order placed. An order can have multiple transactions.
Order Date	The date at which the order was placed.
Customer Name	Name of the customer placing the order.
Country	The country to which the customer belongs to.
State	The state to which the customer belongs from the country.
City	Detail about the city to which the customer resides in.
Region	Contains the region details.
Segment	The ordered product belongs to what segment.
Ship Mode	The mode of shipping of the order to the customer location.
Category	Contains the details about what category the product belongs to.
Sub-Category	Contains the details about what sub - category the product belongs to.

Product Name	The name of the product ordered by the customer.
Discount	The discount applicable on a product.
Sales	The actual sales happened for a particular order.
Profit	Profit earned on an order.
Quantity	The total quantity of the product ordered in a single order.
Feedback	The feedback given by the customer on the complete shopping experience. If feedback provided, then TRUE. If no feedback provided, then FALSE.

Tasks

We will work on finding the most profitable region-category combo and calculating the CLV.

To estimate the CLV, you should complete following steps using Python. We recommend you use a ipython notebook. If you have not used ipython notebook before, the easiest tool you can start with is the Jupyter Notebook, which can be downloaded here: https://ipython.org/notebook.html. You can put your answers to questions in a comments block in your notebook or neatly print them to standard output. The number of points out of 20 total for the assignment is indicated in parentheses after the description of the step.

You are not allowed to use ChatGPT or similar generative AI tools for this assignment. In submitting a notebook for this assignment, you certify that you did not use generative AI in preparation of the notebook.

We recommend reading the blog by Ben Gorman here:

https://usermanual.wiki/Pdf/gormanalysiscomPractical20Guide20to20Calculating20Customer20Lifetim e20Value20CLV.201499060 (also attached to Canvas)

Step 1: Understand the dataset (5 points)

The first thing you should do is to load the CSV file and understand the dataset. It is important to understand the dataset and examine if there is any missing values or outliers before doing analysis. In this step, you should write Python scripts to answer the following questions.

1. Are there any missing values in the dataset? (0.5)
2. What is the range of dates in the dataset? (1.5)
3. How many unique customers are there in the dataset? (0.5)
4. How many unique orders are there in the dataset? (0.5)
5. Profile the data to give the standard descriptive statistics for the Sales field. What are the min, max, variance, and standard deviations? (0.5)
6. Do transaction amounts (sales) in general increase over time (perhaps due to inflation)? (1.5)

Step 2: Explore the dataset (4)

Next, explore the dataset to check if there are any outlier or if there are values that don't make sense. You can use statistical tests to check for outliers. Or, you can simply plot the histogram of the Sales and see if there is any value that appears to be abnormal. (Hint, is there any value that appears to be abnormally large or small? Could it be caused by bad entries (e.g. forgetting decimal separator?). In this step, you should write python scripts to help you answer the following questions.

1. Are there any outliers? (2)
2. If so, how would you treat them? (2)

Step 2.5: Finding the best region-category combo (5)

1. What are the top and bottom three cities, and states for sales and profits, overall, and for each year? (.5)
2. Does giving discounts increase sales and/or profits? Make a graph to show that. Comment on what happens at 80% discount level, and what's the optimum level for maximizing total profit. Does the optimum change over years? (1)
3. How do the different categories and sub-categories perform over time by region? Comment on whether a category-subcategory combo is cyclical, increasing, decreasing for each region. (.5)
4. What factors affect the sales the most? You can run a simple regression to understand the impact. (.5)

And now for calculating CLV:

Step 3: Determine origin year of customers (1)

Some of the underlying customers are brand new and others have been customers for almost five years. The newer customers will have (generally) spent less on average than the old ones. You need to separate the customers into groups based on how long ago they were acquired (e.g. customers acquired in 2010, vs customers acquired in 2011, …). First, assign customers into different groups based on the date of acquisition (origin year). For instance, the earliest transaction date that can be found for customer 2 in the dataset is 5/15/12. Then we assign customer 2 into group 2012 as this customer was acquired in the year 2012. Write Python scripts to assign the appropriate origin year to all customers,

Step 4: Calculate cumulative transaction amounts (1.5)

Now, calculate the cumulative transaction amounts (sales) for customers in each group of origin year. Your output should be something similar to the table below. In the table, each row represents the cumulative amount for customers of each origin year. Each column represents the cumulative amounts at age 12, 24, 36, 48, 60, and 72 months. Note: age represents the time elapsed since the start of each customer group.

Amount.cmltv

Origin	12	24	36	48	60	72
2010-01-01 – 2010-12-31	2255.07	3613.85	5271.87	6627.43	7922.95	8956.55
2011-01-01 – 2011-12-31	2238.46	3758.03	5465.12	6702.14	7861.77	
2012-01-01 – 2012-12-31	2182.92	3878.26	5230.43	6505.42		
2013-01-01 – 2013-12-31	2181.85	3611.81	5230.75			
2014-01-01 – 2014-12-31	1833.85	3263.05				
2015-01-01 – 2015-12-31	1912.37					

To help you complete this step, we provide an example for calculating the first value 2255.07 in the table. First, let's understand the meaning of this value. Since it is in the first row, it represents the transaction amounts made by customers acquired in the year 2010 (2010-01-01 to 2010-12-31). Since it is in the first row, it represents the cumulative amounts up to age 12 months. Therefore, to obtain this value:

- First, find the transaction records for customers with the origin year 2010 (use your result from step 3)
- Then, sum up the amounts made by these customers up to 12 months later than the origin year. Therefore, you should check the TransactionDate to determine if you should include a particular transaction into the sum (e.g. you sum the amount only if the TransactionDate is no greater than 12 months since the start of 2010). Now, do the same calculations for the rest of the entries in the table and print your table to standard output (or document it in comments block). Note: you will not get the same numbers as those in the example table above because the data set is different.

Step 5: Calculate the number of new customers (1)

- Again using Python, calculate the number of new customers by origin year in each year. Your output should be similar to the table below (again, you will have slightly different numbers because the input data is different for the assignment):

NewCustomers.cmltv

Origin	12	24	36	48	60	72
2010-01-01 – 2010-12-31	172	172	172	172	172	172
2011-01-01 – 2011-12-31	170	170	170	170	170	
2012-01-01 – 2012-12-31	163	163	163	163		
2013-01-01 – 2013-12-31	180	180	180			
2014-01-01 – 2014-12-31	155	155				
2015-01-01 – 2015-12-31	160					

- In this table, each entry represents the number of new customers acquired during each origin year. Since this number is irrelevant to age, values for each column in the same row should be the same. Print your table to standard output (or document it in a comments block).

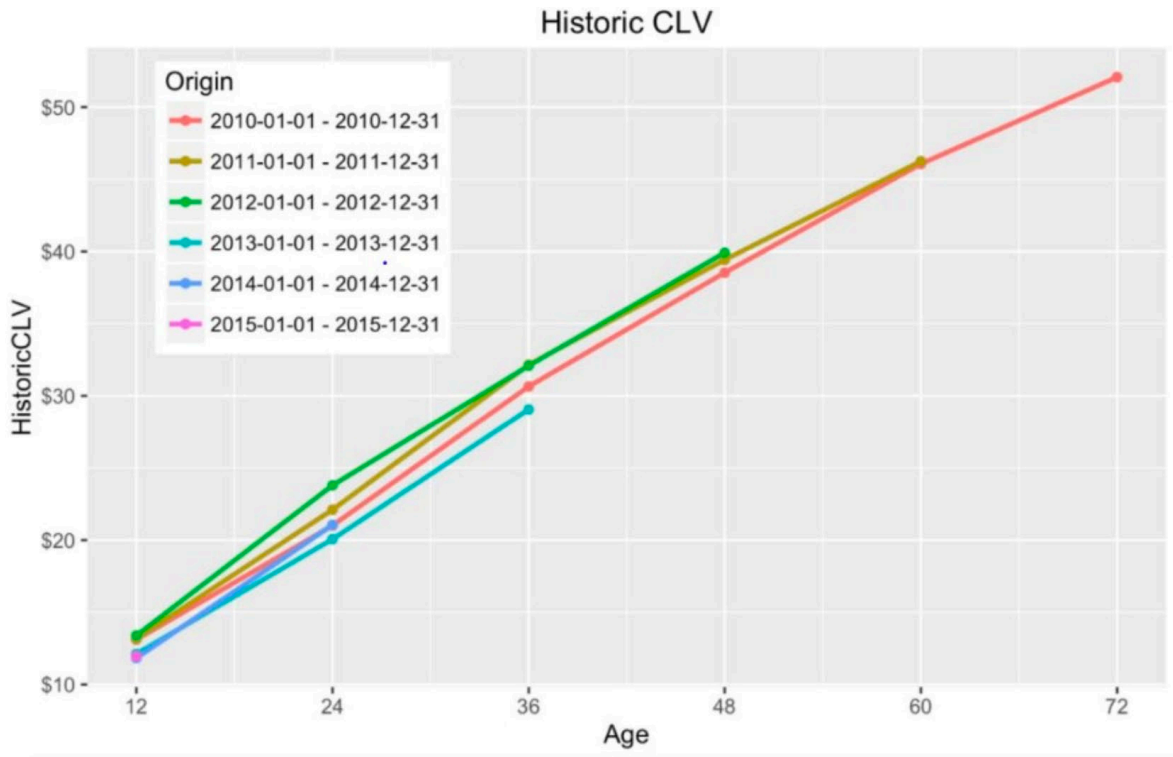
Step 6: Historic CLV (1.5)

- Finally, you are ready to calculate the Historic CLV. Dividing the Amount.cmltv triangle by the NewCustomers.cmltv triangle will give us annual measurements of the cumulative amount spent per customer in each group of annually acquired customers. This is also known as Historic CLV. Print your table to standard output (or document it in comments block). Your output should be similar to the table below

HistoricCLV=Amount.cmltv/NewCustomers.cmltv

Origin	12	24	36	48	60	72
2010-01-01 – 2010-12-31	13.11	21.01	30.65	38.53	46.06	52.07
2011-01-01 – 2011-12-31	13.17	22.11	32.15	39.42	46.25	
2012-01-01 – 2012-12-31	13.39	23.79	32.09	39.91		
2013-01-01 – 2013-12-31	12.12	20.07	29.06			
2014-01-01 – 2014-12-31	11.83	21.05				
2015-01-01 – 2015-12-31	11.95					

For better visualization: plot the results for each year on the same graph. A plot of the historic CLV for each cohort looks like this:



At this point, we'd like to combine all of our data to create a single curve of Historic CLV. A simple, but effective approach to doing this is to take a volume-weighted average of the Historic CLV for each group at each Age, weighted by the number of customers in each group. For example, we'd get:

Age	HistoricCLV
12	12.6
24	21.58
36	30.95
48	39.28
60	46.15
72	52.07

Step 7: Interpreting your results (1)

Interpret the historic CLV and briefly answer the following questions: -

1. How much have customers acquired in 2011 spent to date?
2. Does each group of customers exhibit similar or different patterns of spending? What's the implication for the business?

What to hand in

You should neatly document or print your summary statistics, answers to questions, tables and plots in the ipython Notebook. Make sure to comment your code whenever necessary. Use the file name convention DSPM_HW<HW#3>_<AndrewID>.ipynb, e.g. DSPM_HW1_steier.ipynb.

Useful Python Tutorials

Jupyter Notebook Tutorial

http://nbviewer.jupyter.org/github/jvns/pandas-cookbook/blob/v0.1/cookbook/A%20quick%20tour%20of%20IPython%20Notebook.ipynb

Reading and plotting data using Pandas

http://nbviewer.jupyter.org/github/jvns/pandas- cookbook/blob/v0.1/cookbook/Chapter%201%20-%20Reading%20from%20a%20CSV.ipynb

GroupBy function in Pandas

http://nbviewer.jupyter.org/urls/bitbucket.org/hrojas/learn- pandas/raw/master/lessons/06%20-%20Lesson.ipynb

External Sources

[1] https://en.wikipedia.org/wiki/Customer_lifetime_value

[2]https://usermanual.wiki/Pdf/gormanalysiscomPractical20Guide20to20Calculating20Customer20Lifetim e20Value20CLV.2014990060