Customer Lifetime Value Analysis

By: Mangara Paul Alexander Hutagalung (mangara@sas.upenn.edu)

Customer lifetime value (LTV) is one of the most important metrics to determine the success of a service or a product. It tells us much about our customers' perception of our products and services. Essentially, LTV measures the revenue generated by our customers over their lifetime within our platform. Hypothetically, the higher our LTV, the more loyal and engaged our customers are with our products.

In this exercise, we will go through the process of calculating product A's LTV based on the provided data. We will then answer all the questions in the use case document using analytical methods best suited for each question. However, before going deeper into the analysis, let's look at our available data, clean them, and prepare them for our analysis.

Our database is called `Product A Data Case Study`, and inside, we have two tables, `Revenue Data`, and `Users`. Succinctly, `Revenue Data` stores raw data regarding the activities of our cohorts of customers (labeled as "blue" and "red") in terms of impressions, clicks, and revenue throughout their lifetime. On the other hand, the `Users` table stores raw data regarding how each "color, " similar to marketing campaigns in concept, performed in acquiring new users. The values in the column "color" represent two distinctly different marketing campaigns. One thing to note from this table is a slight anomaly, which the author has limited knowledge of, where activities preceding the install date. For example, suppose a cohort has an installed date of 01/01/23, and yet there are activities for that particular cohort on 12/31/22, which technically makes no sense. Given that we have limited knowledge of this and that these anomaly rows only represent 0.87% of our overall data, we can safely assume to remove these rows of data.

Now, let us combine the two tables and create several new columns that will help us in our expedition to find product A's LTV. We will combine the two tables in Excel with the steps below:

1. SORT the `Revenue Data` table based on "color," then "date_installed," and then "date."
2. Create a "match_identifier" column that merges "date_installed" and "color" into one string that can act as an ID. Do this for both `Revenue Data` and `Users` tables.
3. Left join both tables, using `Revenue Data` as the main table, and move the "user_acquired" column.
4. Remove rows that do not have "user_acquired" values since we cannot measure any metrics per user without this value.
5. Divide the "revenue" by the "user_acquired" to find the "revenue_per_user_acquired."
6. The use case document states that we can assume the acquisition cost per user equals $0.75. We can create an "acquisition_cost" column that states this value across all rows.
7. We then calculate the total cost by multiplying the "user_acquired" with the "acuisition_cost."

8. We then create a cumulative sum column that cumulatively calculates each cohort's revenue, and we can name this column "cumsum_date_color," and then divide this by the number of “user_acquired” to get the “cumsum_per_user.” Remember, the "match_identifier" column represents cohorts.
9. Create a column that labels whether the users in a campaign, represented by the "cumsum_per_user," has reached ROI. The logic is then if cumsum_per_user > 0.75, then "TRUE," else "FALSE," and let us name this column "roi_condition."
10. Now we need to long for a given cohort to reach ROI. This can be done by identifying the first row of our "roi_condition" column as having a " TRUE " value for any campaign. We can find that by identifying the following:
    a. "first_true1":
        i. When "roi_condition" for row n is "TRUE,"
        ii. and “roi_condition” for row n-1 is “FALSE,”
        iii. and "date_installed" for row n is the same as n-1
        iv. and the "color" value is the same for rows n and n-1
    b. "first_true2":
        i. When "roi_condition" for row n is "TRUE,"
        ii. and "date_installed" for row n is not the same as with n-1
11. Lastly, we combine the columns "first_true1" and "first_true2" to find our final list of rows of when a cohort reached ROI for any campaign.

We will use our newly combined table as the base of our analysis. We are now ready to answer each of the use case questions below.

Q1. Graph the Revenue per User over the days from install for the example data set.

To achieve this, we can pivot our table, use the "date_from_install" as the rows, and use "revenue" and "revenue_per_user" as the column values. However, "date_from_install" ranges from 0 to 166, which means we will have 167 X-axis ticks, which will make our graph overcrowded. We can simplify this by grouping our rows into the number of weeks from the install date instead to have a clearer visualization.

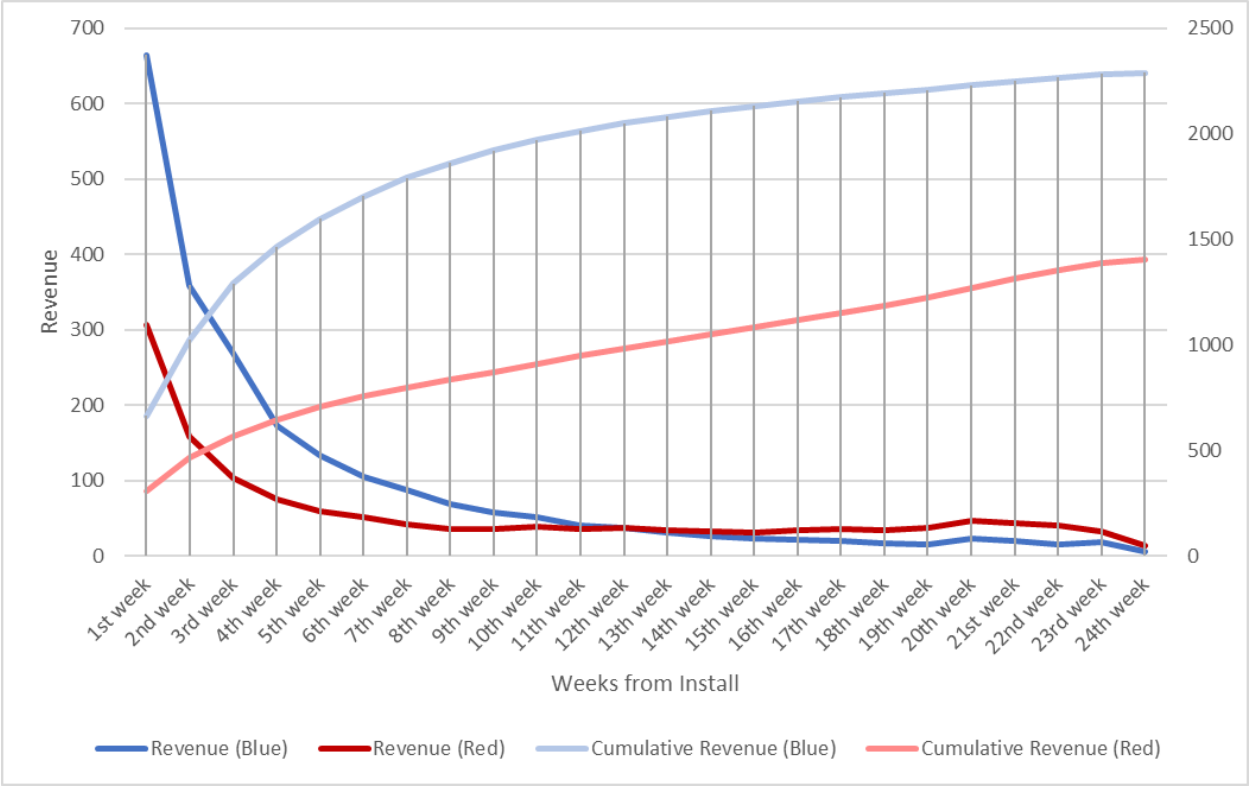


Figure 1. Revenue per cohort over the weeks from install

Our graph above shows that the blue cohort has a significantly higher total revenue per week. The blue cohort generated 1.6x more revenue over its lifetime than the red cohort. With that said, this inference does not consider the number of users per cohort. This leads to our next question, does each cohort have a similar trend if we look at it from the revenue per user point of view?

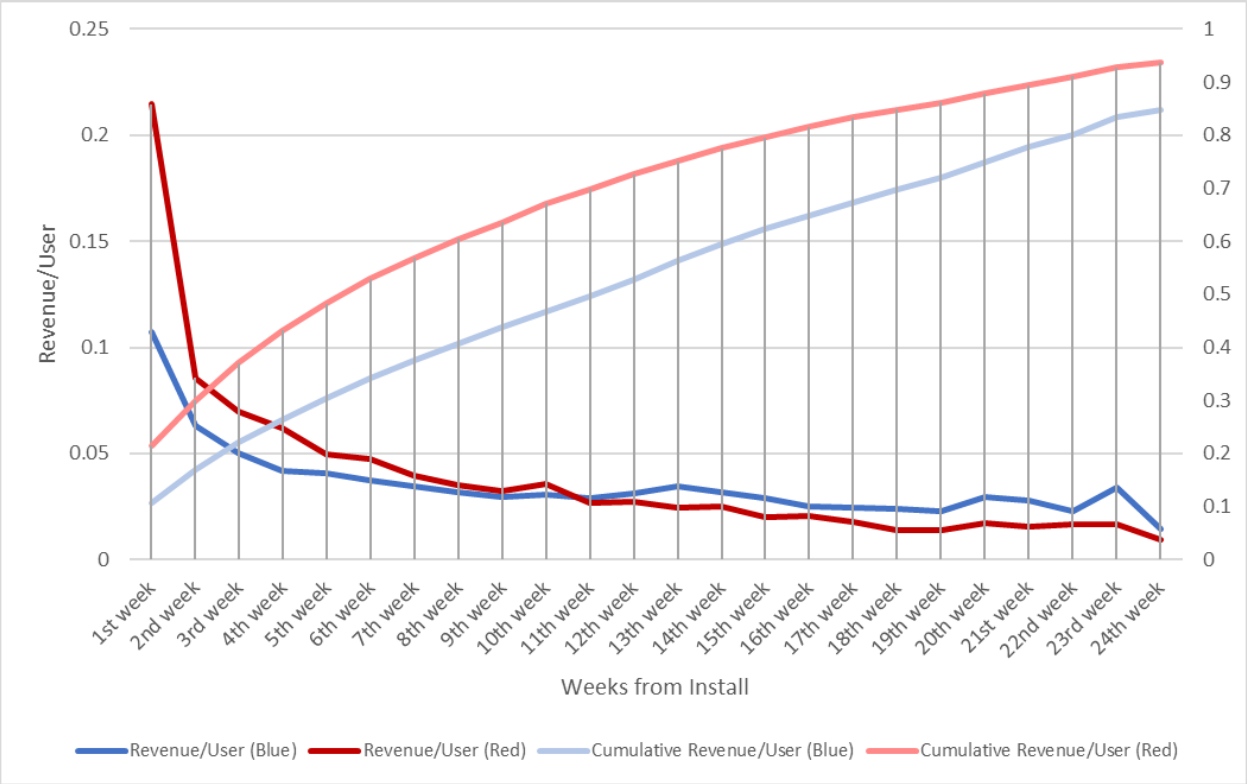


Figure 2. Revenue per user over the weeks from install

If we analyze product A's LTV from the user level instead of the cohort level, the trend flipped, where the red cohort performs significantly better than the blue cohort. Taking both graphs into account, we can infer that:

1. The campaign used to acquire users in the blue cohort seems to perform better, signaled by the total number of users acquired, where the blue cohort acquired 675,032 users while the red cohort acquired 288,522 users (the blue cohort acquired 2.3x more than the red cohort).
2. This, in turn, helps the blue cohort to generate higher revenue (163% higher) over their lifetime relative to the red cohort.
3. However, the campaign used to acquire users in the red cohort seems to attract more loyal users, as signaled by the revenue generated per user.
4. On average, each user in the red cohort generated $0.039 weekly, while the blue cohort generated $0.035 throughout their lifetime.
5. In other words, users from the red cohort generated 10.4% more weekly revenue than those from the blue cohort.
6. Interestingly, the two cohorts have no significant difference if we look at the impressions and clicks per user. The two groups have a significant difference in the total impressions and clicks between the two cohorts, which makes sense given that they have different numbers of acquired users (see Appendix A).

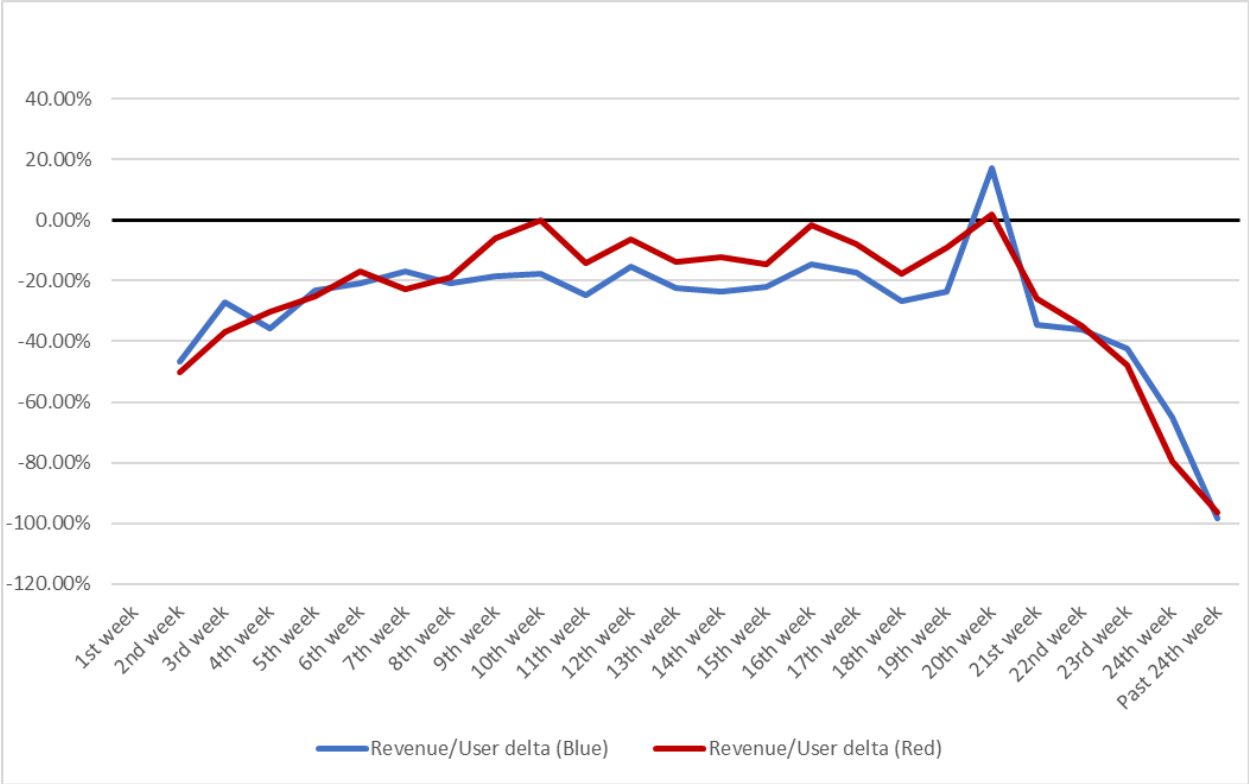


Figure 3. Revenue per user delta over the weeks from install

To further understand our two cohort's activities in product A's application, we can see the graph above showing how each cohort's revenue per user decreases from one week to the next. The closer to the 0% line, the better it is since it means that there is no significant decrease from the previous week. In other words, our red cohort seems to have a more steady trend in weekly revenue per user (on average, 6% decrease per week), while our blue cohort seems to have a more steep decrease in weekly revenue per user (on average, 11% decrease per week).

Q2. Assuming a $0.75 cost per acquisition, on average, how long does it take for the users to pay back their acquisition cost?

Next, let us analyze the quality of each cohort by looking at their return on investment (ROI). ROI is calculated by slicing the data frame by looking at the cohort level, and then the formula is as follows:

$$ROI = \frac{\text{(sum of revenue per cohort - total acquisition cost per cohort)}}{\text{total acquisition cost per cohort}}$$

We also need to slice our data frame by filtering the column "final_first_true" as "TRUE" to determine which cohort has reached ROI and how long it takes to reach ROI. The summary of our overall data regarding ROI can be found in the table below:

Metrics	Total	Blue	Red
Total unique date_installed	328	161	167
Total unique date_installed (ROI = TRUE)	64	13	51
ROI	-30%	-41%	-12%
ROI success rate	38%	20%	56%
Avg ROI period in days (only cohorts that reached ROI)	59	78	52
Avg ROI period in days (all cohorts)	113	141	91

Table 1. Return of investment metrics by cohort

One important thing to note is that not all cohorts have reached ROI. Therefore, separating the performance of successful and unsuccessful cohorts might be beneficial to understand our overall performance better. Therefore, we can summarize the table above: the red cohort is the clear winner out of the two cohorts. Even so, both cohorts resulted in a negative ROI overall. Moreover, only 38% of cohorts have reached the ROI state, which can be considered a low success rate. That means there might be a need to improve the quality of users product A will acquire in the future.

Q3. What is the average value of a user at 30 days? 90 days? 180 days? 365 days? Are any groups of users better than others?

Before fitting our dataset with any predictive model, let us first identify our independent and dependent variables. As the question suggests, we can determine that our dependent variable is the average revenue per user, while our independent variable is the number of days from the install date. Moreover, since we have identified that there are significant differences between the two cohorts activities, we should build our predictive model one for each cohort. Now let us take a look at the relationship between our X and Y variables.

Q3.1. Blue cohort

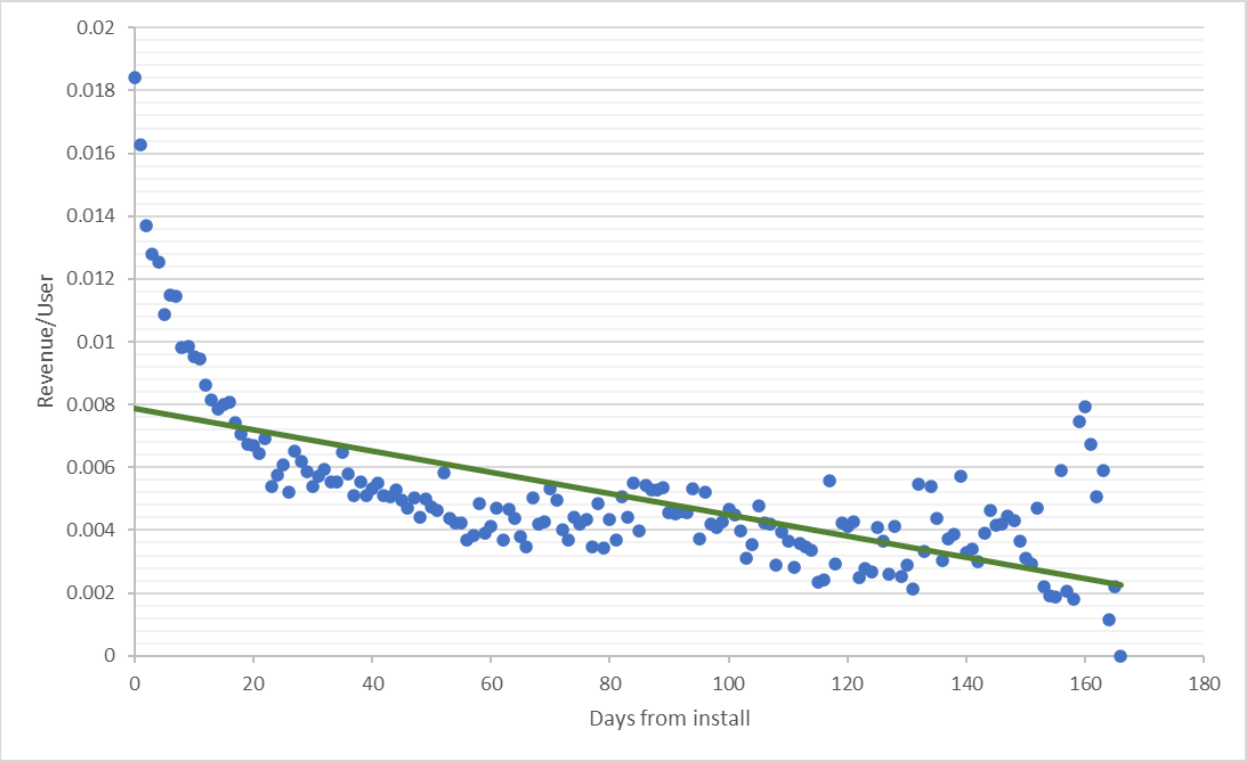


Figure 4. How days from install correlates with revenue/user for the blue cohort

As seen in the graph above, we can see that our variables have a negative correlation. In other words, the further the date from the install date, the lower the revenue per user will be. There are multiple ways to create a predictive model based on our data. However, for this exercise, we can start by fitting our model into a linear regression model and see how accurately we can predict our data. The statistics regarding our first model are as follows:

1. r (correlation coefficient): -0.64971
2. slope (beta): -3.39457E-05
3. intercept: 0.007895309
4. R^2: 0.422118057

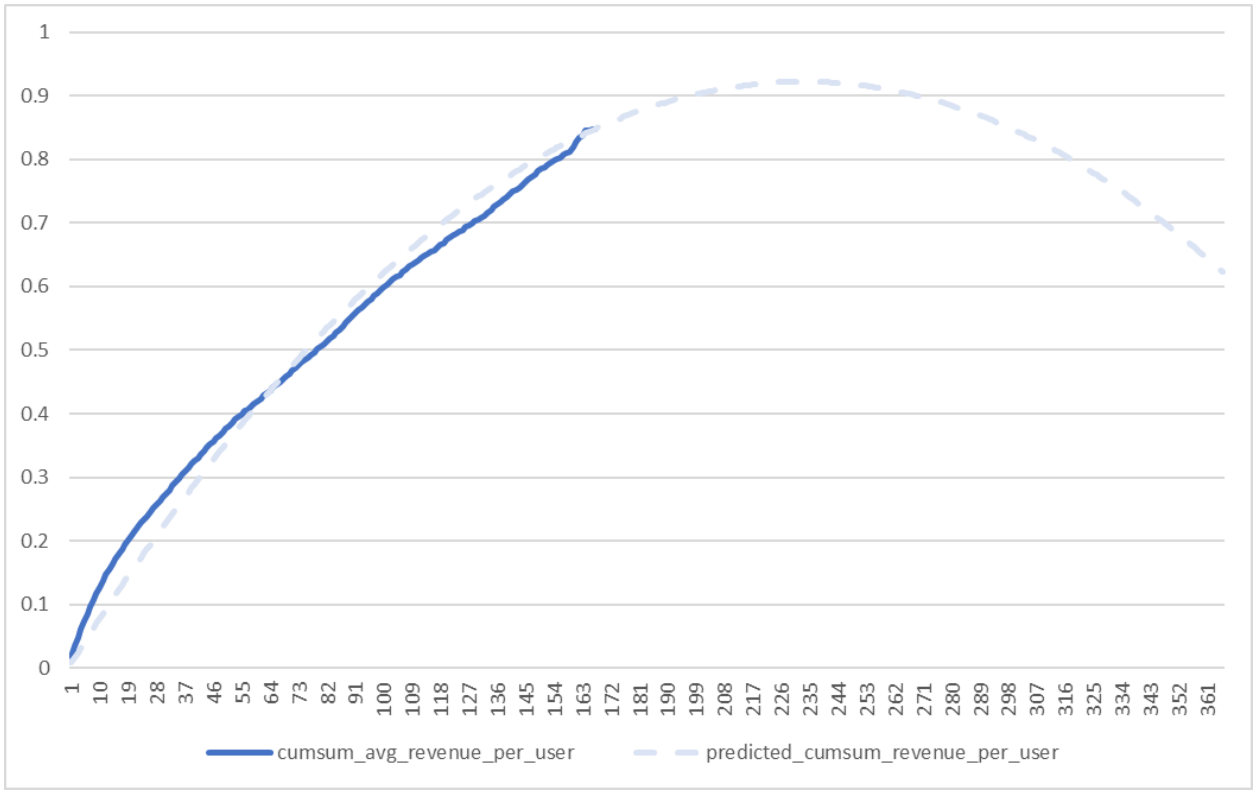


Figure 5. Blue cohort linear regression

The above graph showcases how close or far our predictive model is to the actual data. We can further emphasize the distance between the two by calculating the mean absolute error (MAE). MAE measures how far predicted values are from observed values, and the MAE for our model above is 0.00131694. This means that our linear regression model predicts $0.00131694 revenue per user, more or less on average than the actual value. Not too bad! We can then input our interested number of days after the installation date to this model, and we get the following:

- Cumulative sum for 30 days = 0.221075
- Cumulative sum for 90 days = 0.57157
- Cumulative sum for 180 days = 0.86818
- Cumulative sum for 365 days = 0.614382

Q3.2. Red cohort

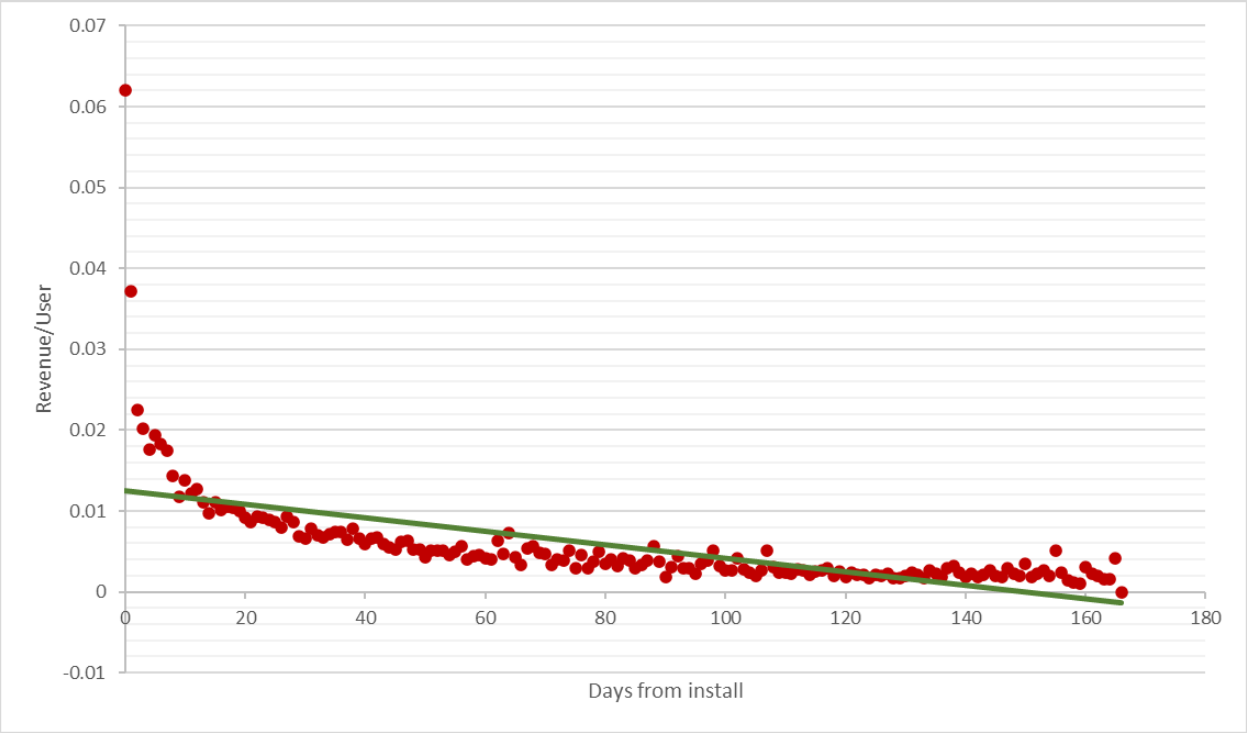


Figure 6. How days from install correlates with revenue/user for the red cohort

The red cohort has a similar relationship between their revenue per user and days from install variables. One main difference from this graph compared to our blue cohort is that the slope seems to be less steep. This can mean that for every additional day after install has passed, the red cohort's revenue per user will decrease, but not as much as the decrease in the blue cohort. The statistics regarding our second model are as follows:

1. r (correlation coefficient): -0.628820816
2. slope (beta): -8.31133E-05
3. intercept: 0.012505103
4. R^2: 0.395415619

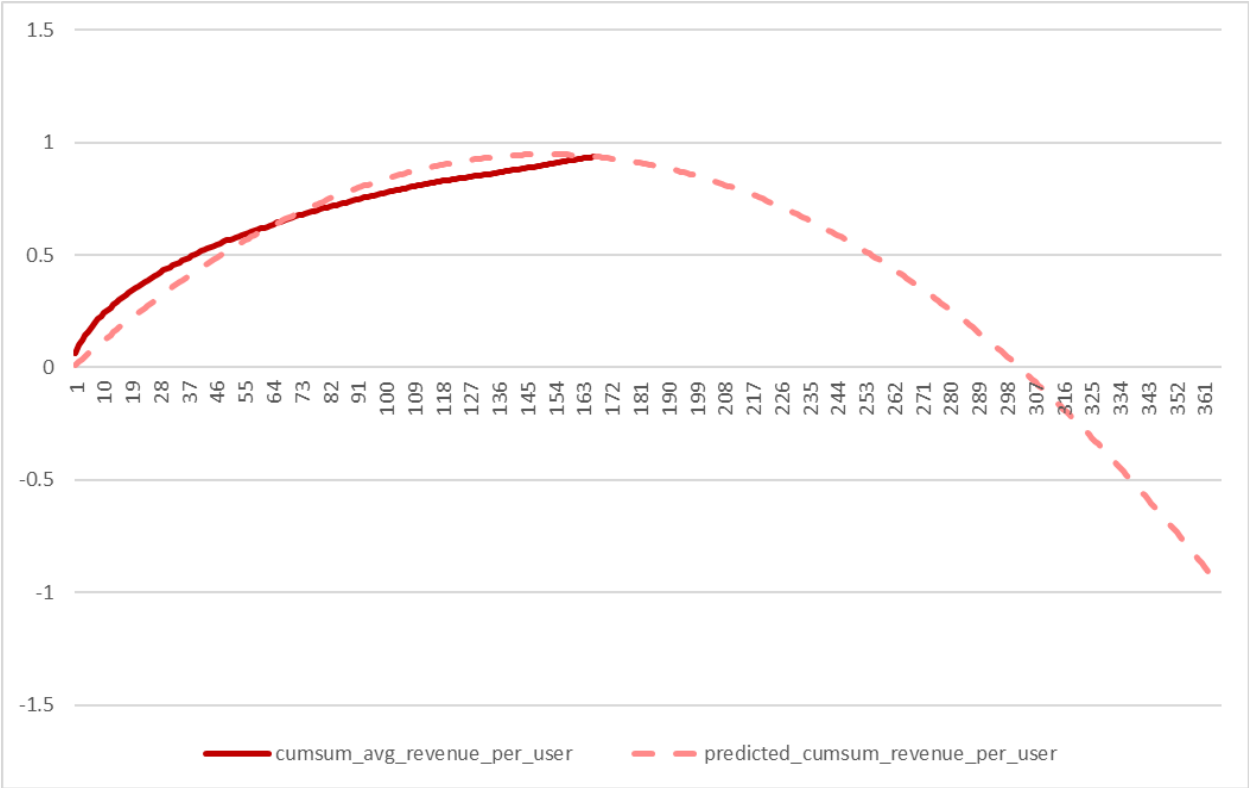


Figure 7. Red cohort linear regression

Similar to our blue cohort predictive model, we can calculate the performance of our linear regression by calculating the MAE. The MAE for our model above is 0.002419981. This means that our linear regression model predicts $0.002419981 revenue per user, more or less on average than the actual value. We can then input our interested number of days after the installation date to this model, and we get the following:

- Cumulative sum for 30 days = 0.336505
- Cumulative sum for 90 days = 0.78511
- Cumulative sum for 180 days = 0.897003
- Cumulative sum for 365 days = -0.98719

Q3.2. Predictive Model Summary

You might be wondering why the curve goes into the negative in both of our predictive models when we stretch it far into the future. This is because we have a negative relationship between our X and Y variables. To counteract this, we can set a floor value for the result of 0. It means that our predictive model will set any negative value to 0, which means that we can identify the number of days our users will have 0 utilities for our product.

There is also another way we can analyze this data and build a different predictive model. Instead of treating this data in a cross-panel format, we can present this data in a time-series or panel format, which opens up various predictive models, such as ARIMA, SARIMAX, and Prophet. However, these

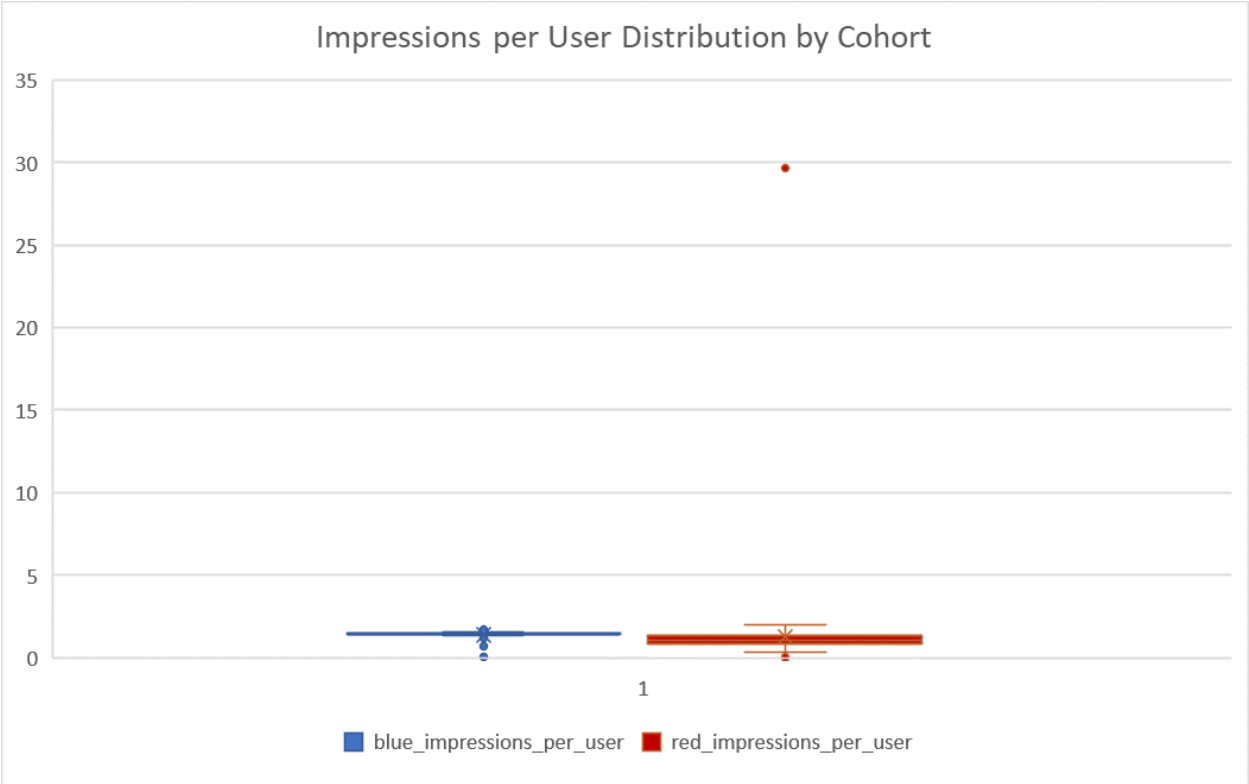
models are much more complex and require many more assumptions to be satisfied. Once we have more variety of data and clarity regarding how the product works, we can then try to build a better model to predict the LTV of product A's users.

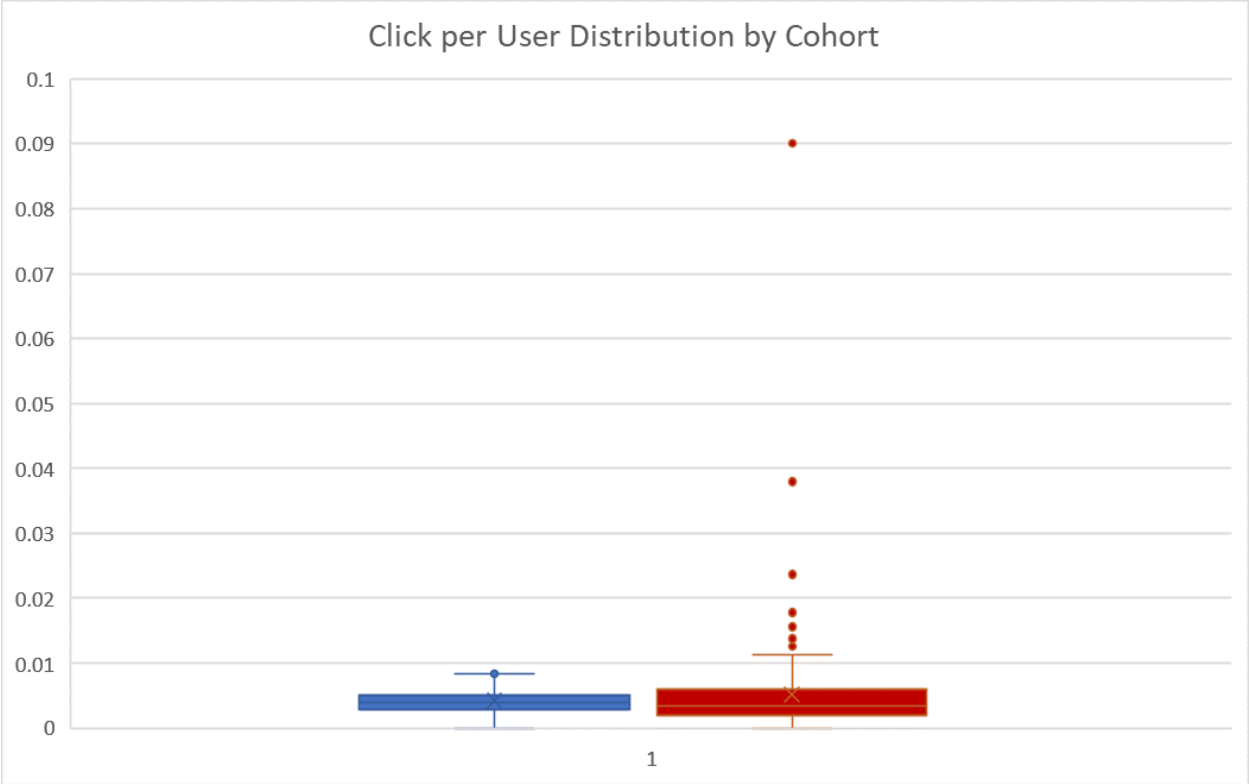
Q4. Now that you've gone through the data, pick the most informative pieces and present a brief summary explaining what you've found.

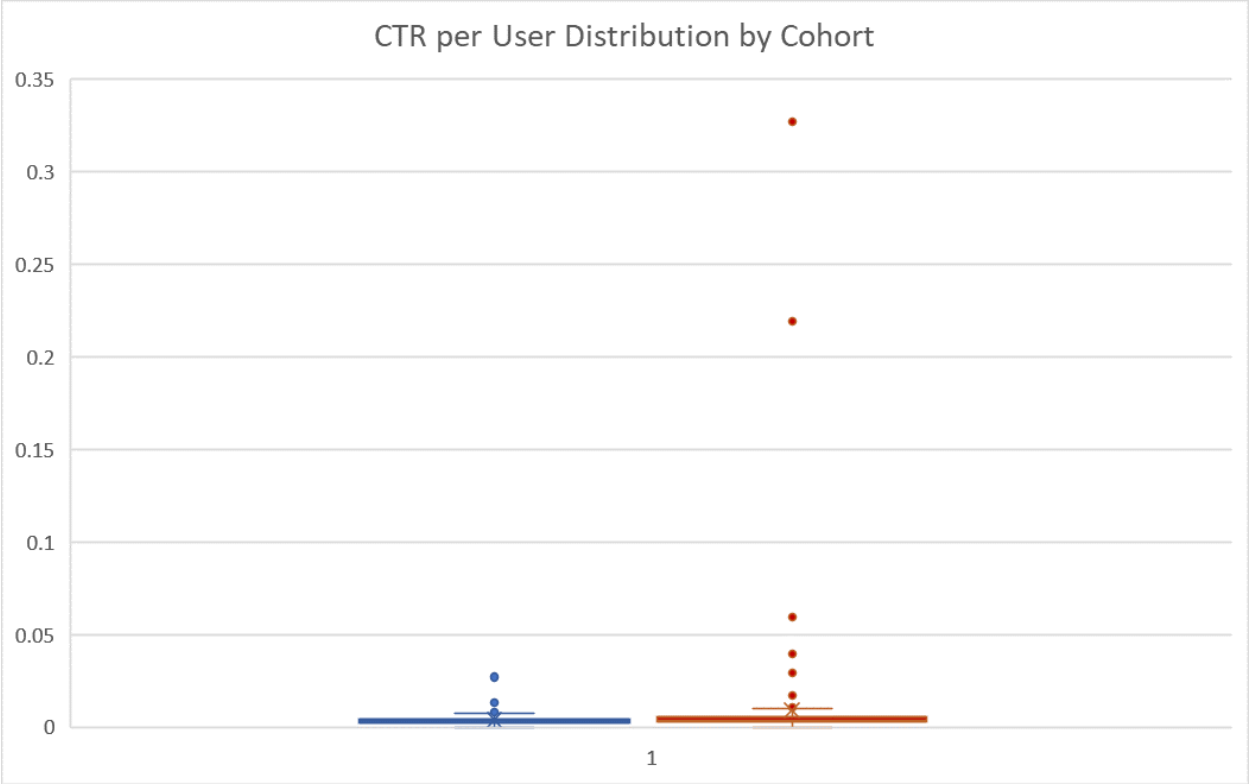
Here are the most important insights from our analysis above:

1. The Blue cohort is the best cohort for acquiring a total number of users, while the red cohort is the best cohort for generating quality and loyal users (in terms of revenue per user). This assumes that cohorts are synonymous with acquisition campaigns, and the two colors represent different ways of acquiring users.
2. The current user quality acquired by both cohorts is still underperforming, with the red cohorts leading with a 56% success rate of reaching the ROI state and the blue cohort with a 20% success rate of reaching the ROI state. Moreover, assuming that this data is the population and not just the sample, this means that product A is operating at a loss.
3. Understanding the prominent features that created the positive quality in each cohort is crucial. We need to find what makes the blue cohort attract more overall users while also understanding what makes the red cohort attract more quality and loyal users. To do this, we can run experiments to identify the most impactful driver after listing all the potential drivers for each cohort.

Appendix A.


Impressions per User Distribution by Cohort
35
30
25
20
15
10
5
0
1
blue_impressions_per_user
red_impressions_per_user


Click per User Distribution by Cohort
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1



Items	Cohort level	Per user level
t-test for impressions	0.0000001*	0.4443047
t-test for clicks	0.0000021*	0.0906585
t-test for click-through-rate	0.0285352	0.0285352

Appendix A.4. T-Test table for cohort and per user level (*p-value < 0.05)