Penn
Master of Behavioral
and Decision Sciences

Case Study - Analyzing Customer Lifetime Value

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

Overview

Background:

- A large part of success in any given product comes from measuring your bets, and using the data as your guide to bet big when the indicators are right.
- One of the primary ways that we measure our initiatives is by looking at the customer lifetime value of an acquired user.
- In the data that is provided to us, we have indicators to find the quality of the users we have acquired that are clustered into two groups of colors, red and blue.
- We can think of these colors as a categorical variable that represents our user acquisition campaigns.
- Therefore, given the data that we received:

Can we determine which group of users would generate the most revenue for a product based on their lifetime value?

Source: ReachMobi Data Case Study (2023)

Research Questions

Research question 1:

Graph the Revenue per User over the days from install for the example data set. What conclusions/information can we ascertain from this graph?

Research question 2:

Assuming a $0.75 cost per acquisition, on average, how long does it take for the users to pay back their acquisition cost? Are any groups better than others?

Research question 3:

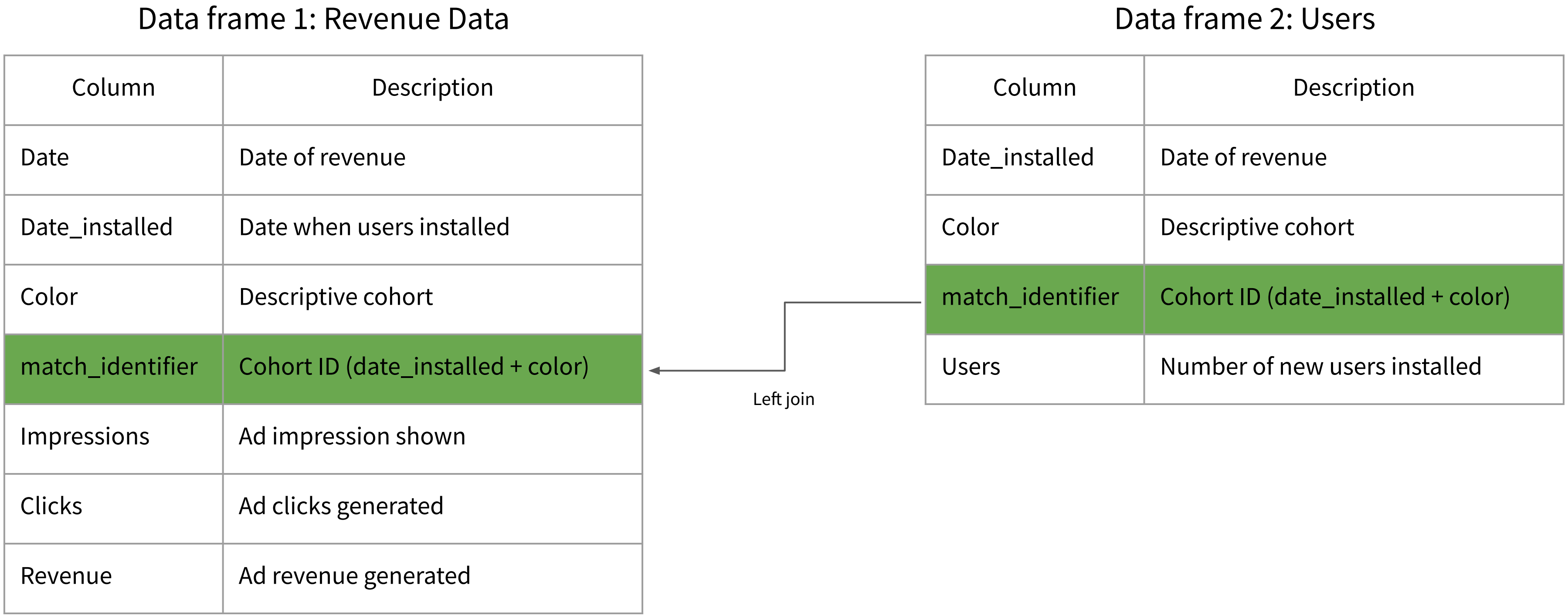
What is the average value of a user at 30 days? 90 days? 180 days? 365 days? Are any groups of users better than others?

Research question 4:

Pick the most informative pieces and present a brief summary explaining what you've found.

Source: ReachMobi Data Case Study (2023)

Data Preparation



Data frame 1: Revenue Data

Column	Description
Date	Date of revenue
Date_installed	Date when users installed
Color	Descriptive cohort
match_identifier	Cohort ID (date_installed + color)
Impressions	Ad impression shown
Clicks	Ad clicks generated
Revenue	Ad revenue generated

Left join

Data frame 2: Users

Column	Description
Date_installed	Date of revenue
Color	Descriptive cohort
match_identifier	Cohort ID (date_installed + color)
Users	Number of new users installed

Data Preparation

Data manipulation:

- Remove rows that do not have "user_acquired" values since we cannot measure any metrics per user without this value.
- Divide the "revenue" by the "user_acquired" to find the "revenue_per_user_acquired."
- 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.
- We then calculate the total cost by multiplying the "user_acquired" with the "acuisition_cost."
- 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.
- 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."
- 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:
  - "first_true1":
    - When "roi_condition" for row n is "TRUE,"
    - and “roi_condition” for row n-1 is “FALSE,”
    - and "date_installed" for row n is the same as n-1
    - and the "color" value is the same for rows n and n-1
  - "first_true2":
    - When "roi_condition" for row n is "TRUE,"
    - and "date_installed" for row n is not the same as with n-1
- 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.

Data Preparation

Final table:

Column*	Description
Date	Date of revenue
Date_installed	Date when users installed
date_from_installed	date - date_installed
revenue_per_user_acquired	revenue / number of users acquired
total_acquisition_cost	$0.75 * number of users acquired
cumsum_per_user	Cumulative sum of any given cohort ID (match_identifier) from MIN(Date) to current Date
final_first_true	A boolean variable (TRUE/FALSE) to identify the first date of a given cohort ID reached ROI
roi_group	A boolean variable (TRUE/FALSE) to identify whether a given cohort ID had reached ROI, regardless of when it reached said state

*The list of columns only represents those deemed important and predictive of our research question and not the overall columns available in our newly merged data frame.

Research Question 1

Graph the Revenue per User over the days from install for the example data set. What conclusions/information can we ascertain from this graph?

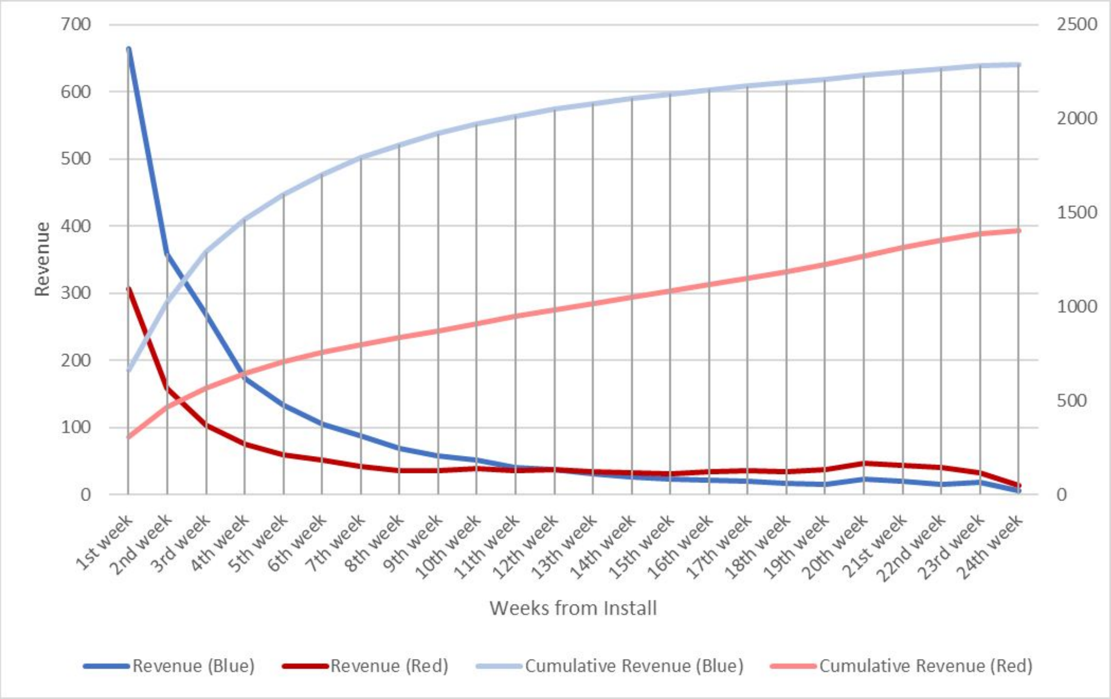


Figure 1. Revenue per cohort over the weeks from install

$95.4

weekly revenue generated by the **blue** cohort

$58.4

weekly revenue generated by the **red** cohort

1.6x

more revenue generated by **blue** over its lifetime relative to red

Research Question 1

Graph the Revenue per User over the days from install for the example data set. What conclusions/information can we ascertain from this graph?

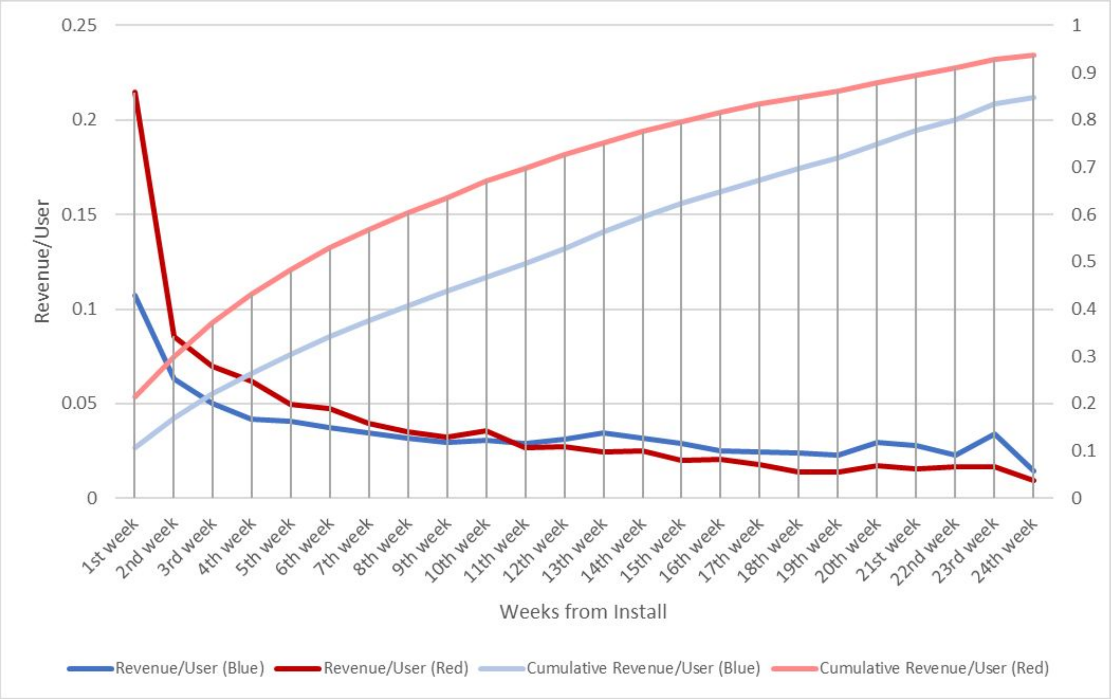


Figure 2. Revenue per user over the weeks from install

$0.035

weekly revenue generated by the **blue** cohort

$0.039

weekly revenue generated by the **red** cohort

1.1x

more revenue generated by **red** over its lifetime relative to red

Research Question 1

Taking both graphs into account, we can infer that:

The campaign used to acquire users in the blue cohort seems to perform better (the blue cohort acquired **2.3x** more than the red cohort)

Overall, blue cohort to generate higher revenue (**163%** higher) over their lifetime relative to the red cohort

From revenue-per-user perspective, users from the red cohort generated **10.4%** more weekly revenue than those from the blue cohort

Research Question 1 - Auxiliary Analysis

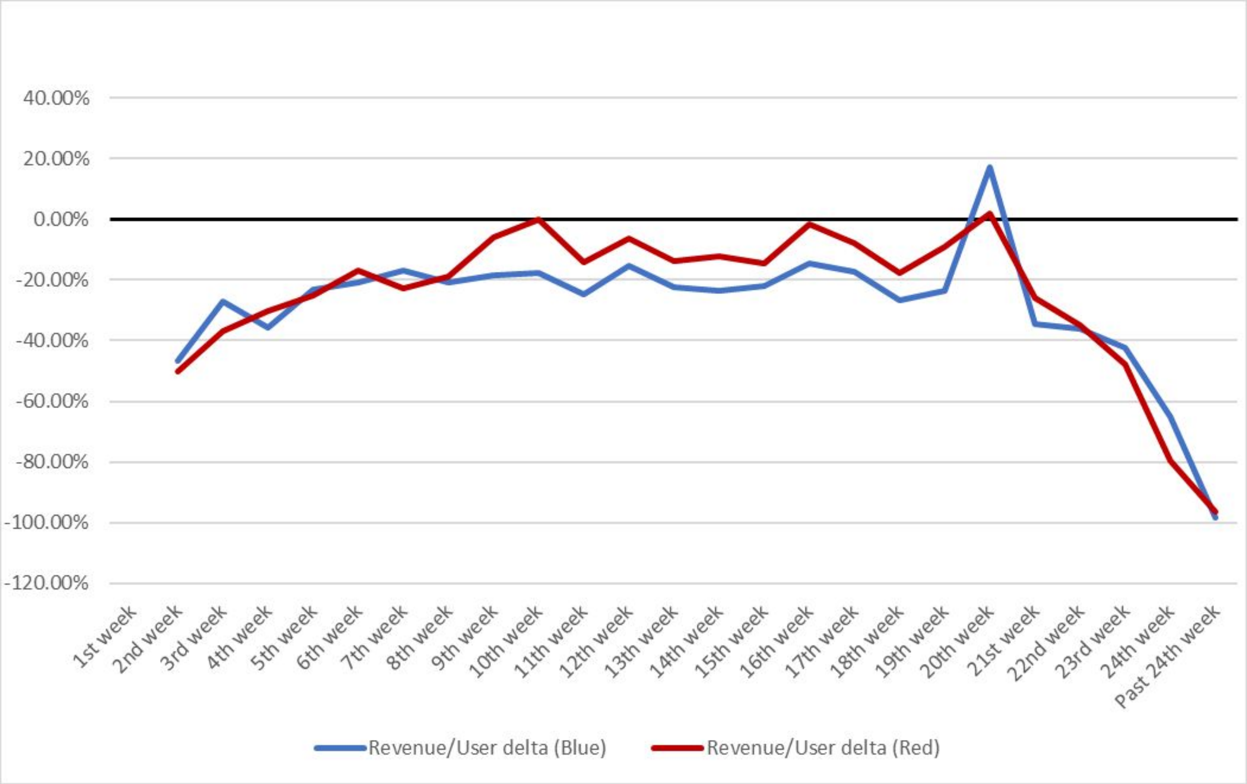


Figure 3. Revenue per user delta over the weeks from install

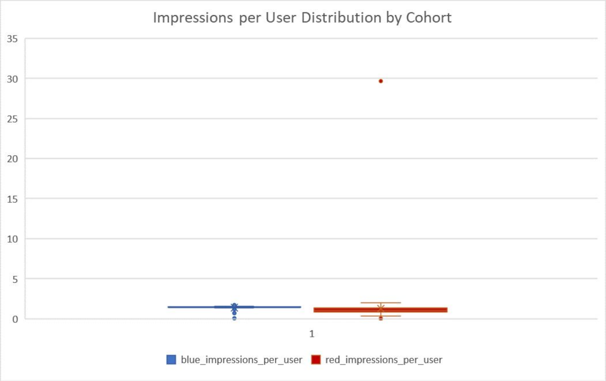
-11%

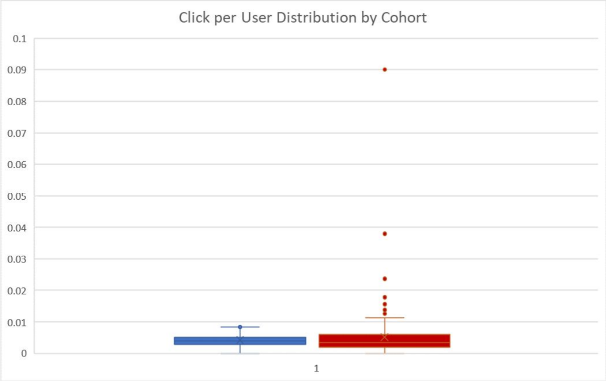
average revenue decrease from one week to the next for **blue** cohort

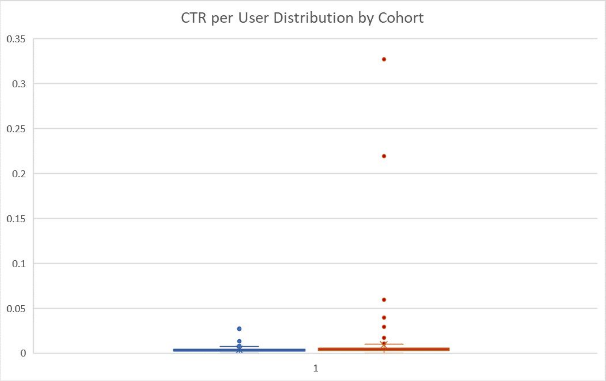
-6%

average revenue decrease from one week to the next for **red** cohort

Research Question 1 - Auxiliary Analysis


Impressions per User Distribution by Cohort


Click per User Distribution by Cohort


CTR per User Distribution by Cohort

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)

Research Question 2

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

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

Research Question 3

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?

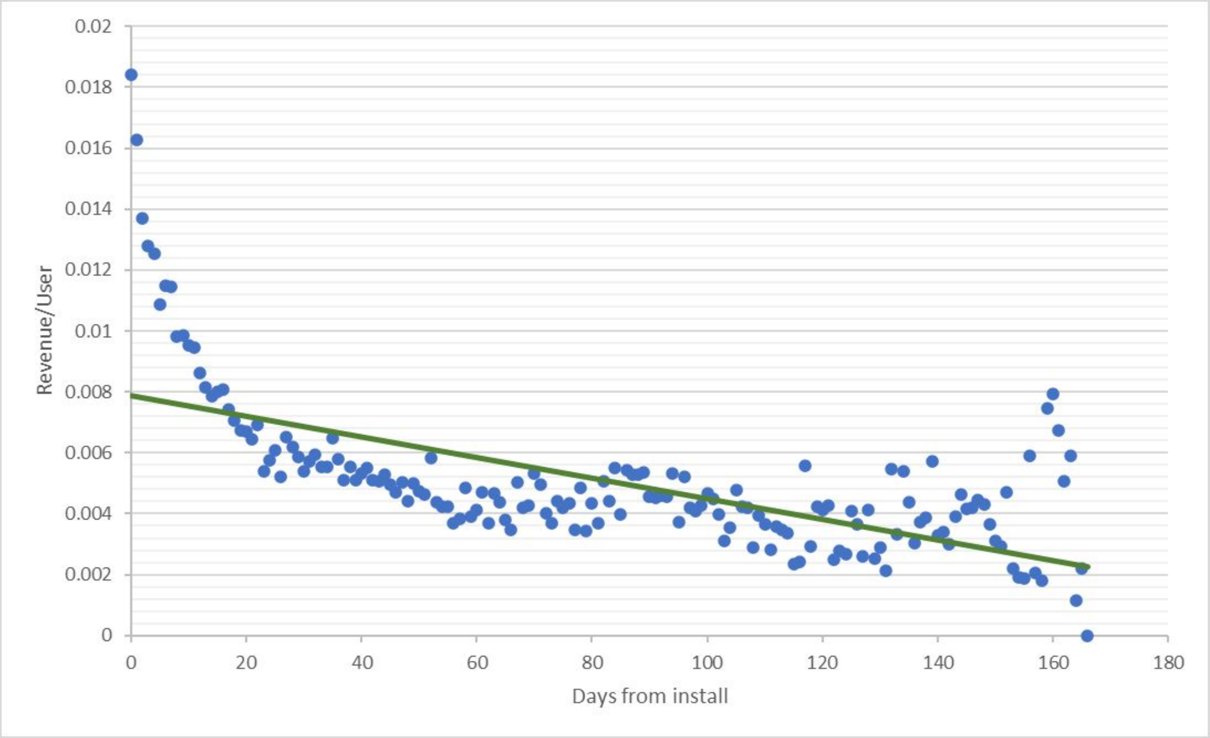


Figure 4. How days from install correlates with revenue/user for the **blue cohort**

Statistics:

r = -0.64971
slope (beta) = -3.39457E-05
intercept = 0.007895309
R^2 = 0.422118057

Research Question 3

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?

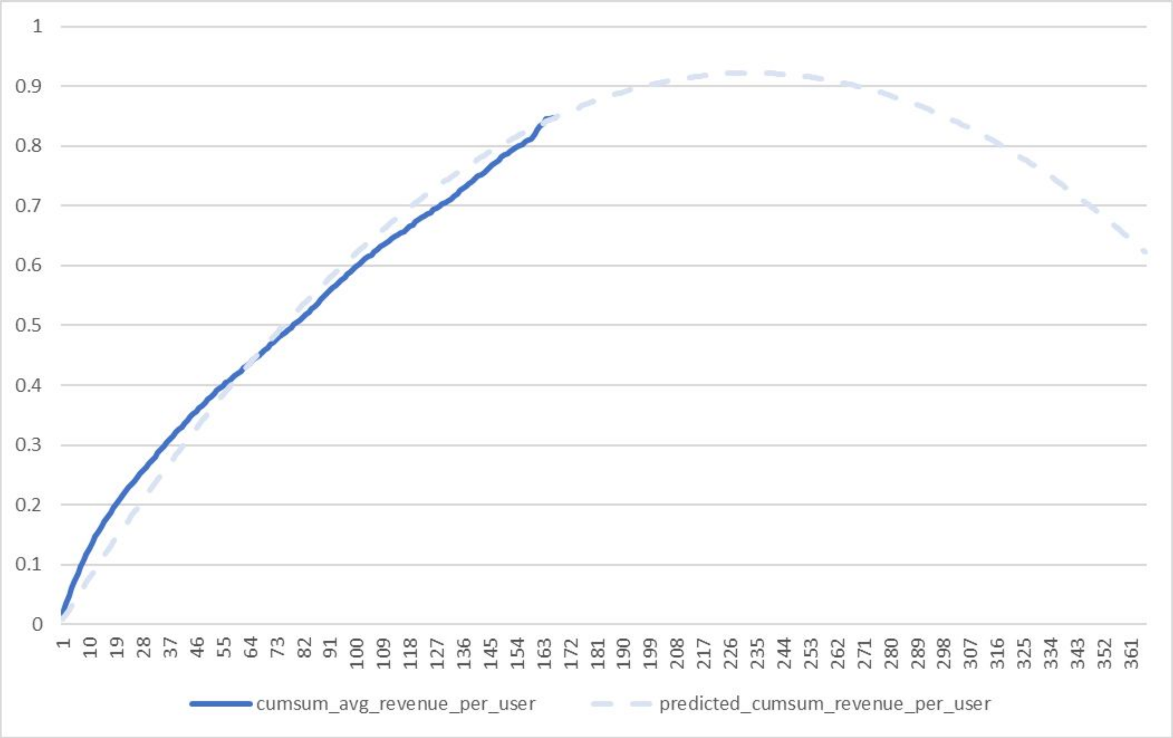


Figure 5. **Blue cohort** linear regression

MAE* = 0.00131694

Therefore:

- 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**

*This model predicts $0.00131694 revenue per user, more or less on average than the actual value

Research Question 3

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?

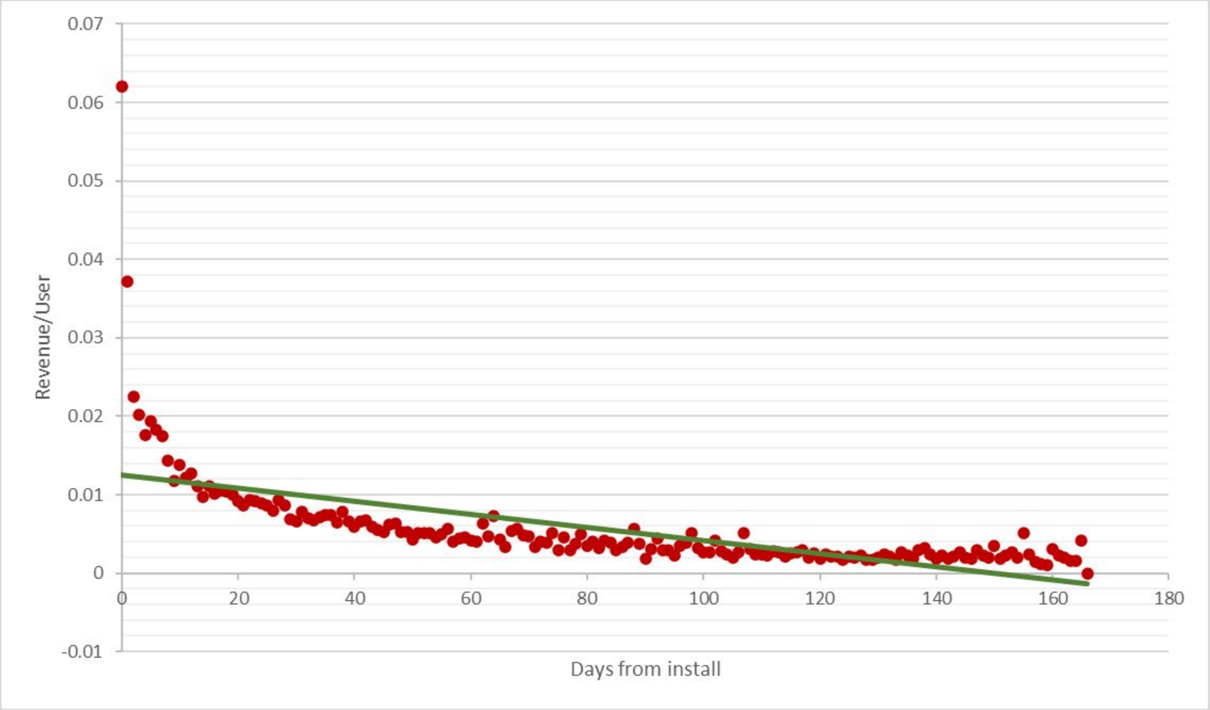


Figure 6. How days from install correlates with revenue/user for the **red cohort**

Statistics:

r = -0.628820816
slope (beta) = -8.31133E-05
intercept = 0.012505103
R^2 = 0.395415619

Research Question 3

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?

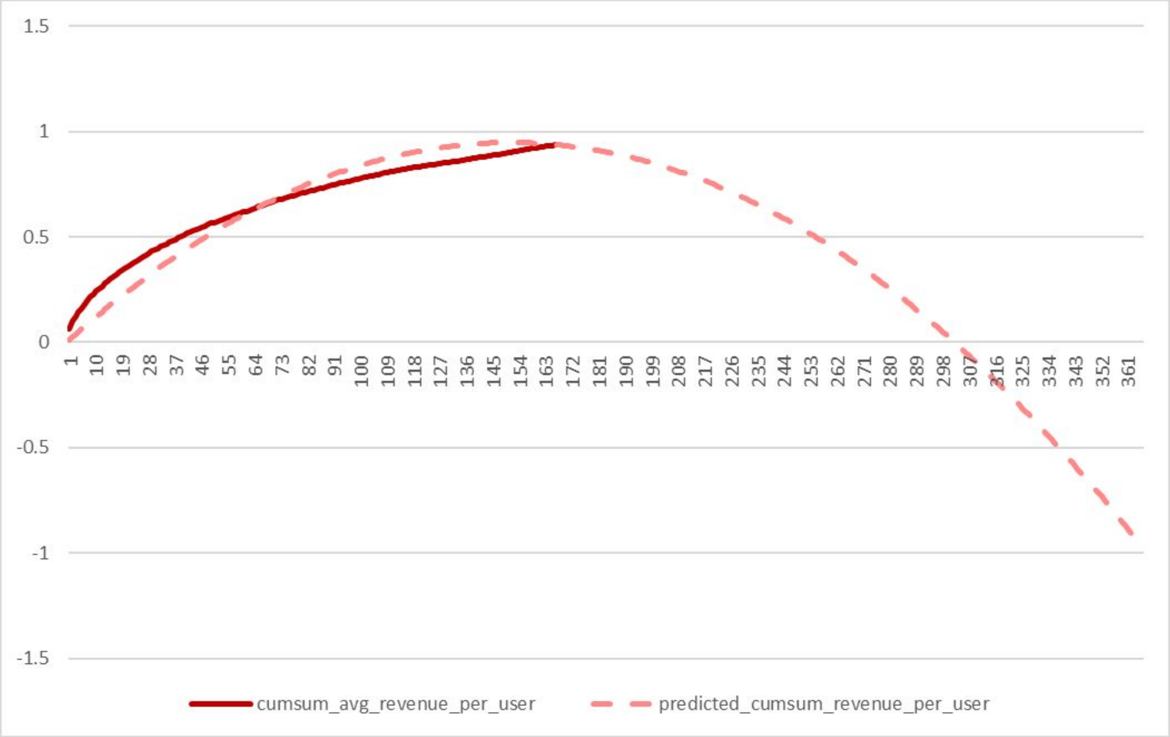


Figure 7. **Red cohort** linear regression

MAE* = 0.002419981

Therefore:

- 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**

*This model predicts $0.002419981 revenue per user, more or less on average than the actual value

Research Question 3

Taking both models into account, we can infer that:

Given that the two cohorts behave differently regarding LTV, we can assume that having two predictive models would be better than having a one-fits-all model

Given that our data showed a negative slope, the predicted value will become smaller once it passes certain days. To counteract this, we can set the floor value as 0

Research Question 4 (Discussion)

Summary:

- 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.
- 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 our product in this use case is operating at a loss.
- 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.

Potential analysis improvement and recommendation:

- We can improve our analysis by adding other data points related to users' activity within our app.
  - One simple way we can determine whether a given data point can predict our user's level of quality is by running a correlation analysis in addition to the business context.
  - Several data points might have predicted the user's level of quality, such as their time spent on the app, net promoter score, etc.
- If we have a better understanding of how the app works and access to other data points, we can transform this use case into a time-series problem and fit predictive models such as ARIMA, SARIMAX, and Prophet by Facebook
  - However, it is essential to note that each model requires assumptions to be satisfied.
  - In other words, these models are highly specific compared to general predictive models such as linear regression.

Thank you!

Contact:
Email: mangara@sas.upenn.edu