


Marketing Strategy Optimization via User Behavioral Segmentation for Achieving Highest Expected ROI



Created by: Marin Stoytchev

Date: Oct. 27, 2021

Link: https://github.com/marin-stoytchev/data-science-projects/tree/master/user_behavioral_segmentation

Executive Summary

Problem Statement

- We are planning a targeted promotion with the objective to convert users who are currently using free app content to using paid app content.
- The question is: Can we determine the group of users that has the potential to yield the highest ROI?

Approach Used

- We treat this problem as a clustering problem to find the group of “non-converted” users (users who do not generate revenue) which overlap in behavior with “converted” users (users who generate revenue).
- Three different likely user segmentation results are used with realistic ranges of average user conversion rate to identify the target group with highest expected ROI.

Objective Achieved

- We find that promotions targeting 5% of the users who are not yet converted, but have strong overlap with converted users, has the potential of achieving highest expected ROI of up to 1300% with minimal risk for incurring losses.

Table of Contents

1. Problem Definition	4
2. Data Processing	6
3. Data Exploration	9
4. Data Modeling	13
5. ROI Optimization	17
6. Summary	20

II. Problem Definition



Define the Problem

Problem

- A mobile apps company has a very small percentage of users who generate revenue (converted users). The company is planning a series of promotions targeting users who currently do not generate revenue, but are likely to convert. The problem is to identify the right group of users. If the right users are not selected the company is going to lose money. Even if a group of users is selected which will yield profit from the campaign, can we be sure that our selection is optimal?

Question

- How to select the group of users to be offered the promotions which will yield the highest possible ROI with minimal risk for losses?

Risk

- If there is little to no overlap in the behavior of converted and non-converted users, it would be impossible to make a meaningful selection of users targeted with promotions.

Tasks

- Process the available data to create a dataset with most relevant features.
- Using this dataset, analyze feature behavior and correlations.
- Create a clustering model that will provide optimal user segmentation.
- Based on the clustering model results determine the range of expected ROI using realistic user conversion rate estimates and select the user segment with the highest expected ROI in most of the scenarios.

III. Data Processing



Available Data

- The available data consists of the following datasets:
 - User data – 22576 entries
 - Spending data – 107764 entries
 - Sessions data – 722955 entries
 - Revenue data – 6685 entries
- Samples of each dataset are provided for illustration of the datasets features

User Data Sample

	User_ID	Install_Date	Language	Country	Mobile_Device
0	0	2020-04-01	en	US	iPhone4,1
1	1	2020-04-01	en	IN	iPod5,1
2	2	2020-04-06	en	US	iPod7,1
3	3	2020-04-03	nb	NO	iPhone8,1
4	4	2020-04-03	en	GB	iPhone5,4
5	5	2020-04-07	en	US	iPhone5,3
6	6	2020-04-06	en	US	iPhone3,1
7	7	2020-04-06	en	US	iPhone6,1
8	8	2020-04-03	en	US	iPhone4,1
9	9	2020-04-04	tr	TR	iPhone6,2

Spending Data Sample

	User_ID	Time_of_Session	Spending	Amount
0	9829	2020-04-01 03:03:04	PointsEarned	-22
1	13757	2020-04-01 03:35:53	PointsEarned	-22
2	13757	2020-04-01 03:52:10	PointsEarned	-22
3	10009	2020-04-01 04:10:00	PointsEarned	-22
4	10009	2020-04-01 04:26:46	PointsEarned	-22
5	14151	2020-04-01 06:38:33	PointsEarned	-22
6	10799	2020-04-01 08:08:26	PointsEarned	-22
7	18706	2020-04-01 08:48:40	PointsEarned	-22
8	22489	2020-04-01 09:43:01	PointsEarned	-22
9	22489	2020-04-01 09:54:14	PointsEarned	-22

Sessions Data Sample

	User_ID	Time_of_Session	Session_Number
0	14067	2020-04-01 00:06:50	1
1	14067	2020-04-01 00:22:27	2
2	16275	2020-04-01 01:23:03	1
3	16275	2020-04-01 01:31:16	2
4	16275	2020-04-01 01:47:22	3
5	16275	2020-04-01 01:49:31	4
6	16275	2020-04-01 02:06:51	4
7	16275	2020-04-01 03:10:40	5
8	265	2020-04-01 03:55:56	1
9	12244	2020-04-01 05:23:50	1

Revenue Data Sample

	User_ID	Time_of_Session	Purchase	Revenue
0	7480	2020-04-04 08:15:49	Points	760
1	7480	2020-04-04 08:24:15	Upgrade	760
2	7480	2020-04-04 22:49:08	Points	410
3	2466	2020-04-06 00:16:48	Points	760
4	22001	2020-04-06 09:13:45	Points	760
5	19008	2020-04-06 09:44:30	Points	760
6	22001	2020-04-07 09:36:37	Points	760
7	9487	2020-04-07 15:57:34	Points	760
8	9487	2020-04-07 15:57:34	Points	760
9	11963	2020-04-07 20:51:03	Upgrade	760

Created Dataset for Analysis and Modeling

- From the available data, a dataset with the most relevant features is created for analysis and modeling
- Newly-created features in this dataset
 - User Group: 1 = Non-converted; 2 = Converted
 - Active Period: The time between first and last user session in hours

Dataset for Analysis and Modeling Sample

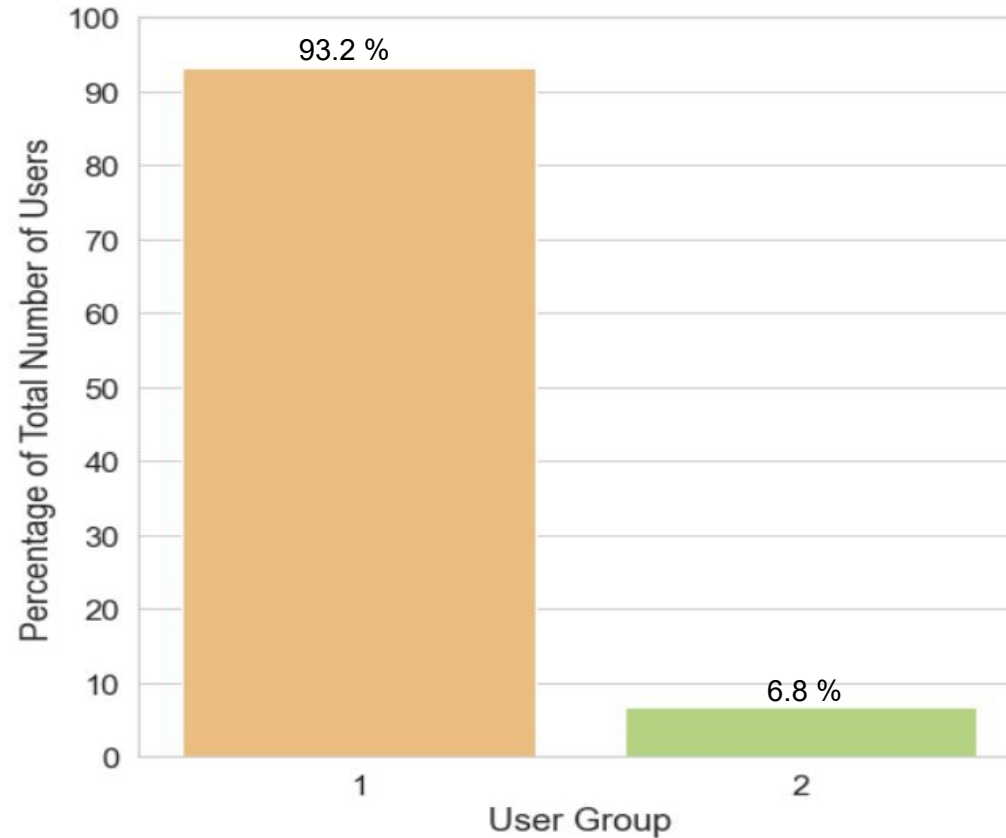
	User_ID	Install_Date	User_Group	Active_Period	N_Sessions	Spend_Points	Spend_Premium	Spend_App	Spend_VP
0	0	2020-04-01	1	89	12	-22	0	0	0
1	1	2020-04-01	1	511	33	-44	0	0	0
2	2	2020-04-06	2	1308	38	-44	74	-73	0
3	3	2020-04-03	1	2	3	-22	0	0	0
4	4	2020-04-03	1	74	3	-22	0	0	0
5	5	2020-04-07	1	126	11	-22	0	0	0
6	6	2020-04-06	1	25	5	-22	0	0	0
7	7	2020-04-06	1	553	44	-22	0	0	0
8	8	2020-04-03	1	626	29	-22	0	0	0
9	9	2020-04-04	1	0	6	0	0	0	0

III. Data Exploration



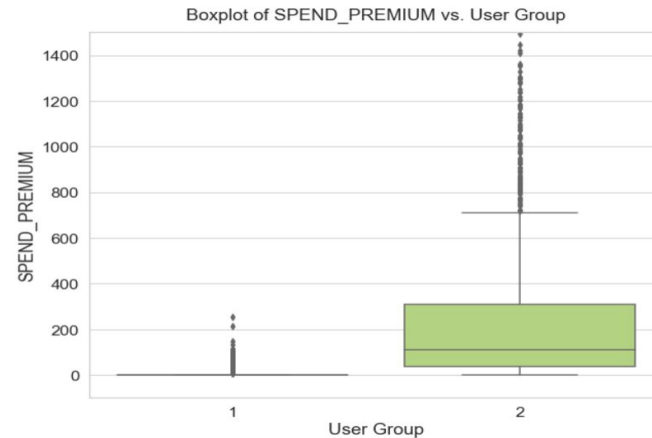
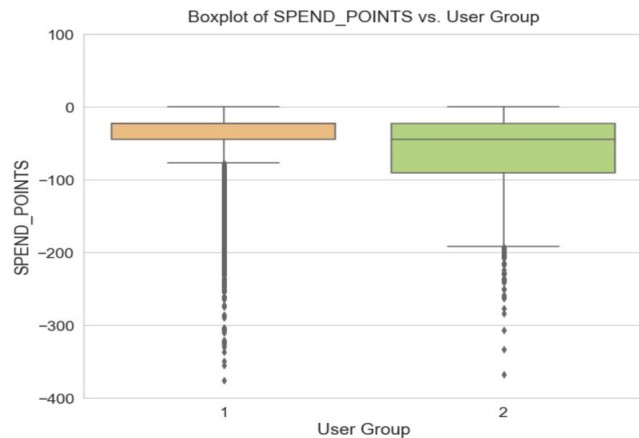
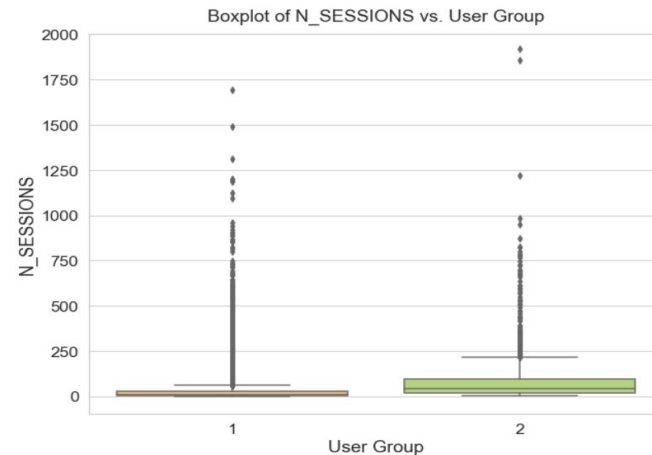
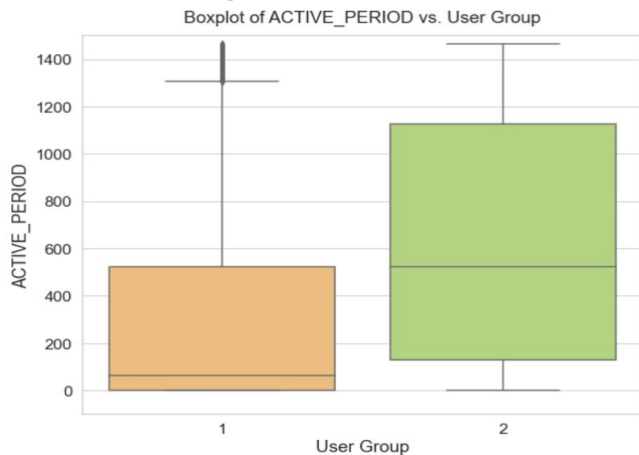
Users Partitioning by Group

- Currently, only 6.8 % of users are converted – generate revenue



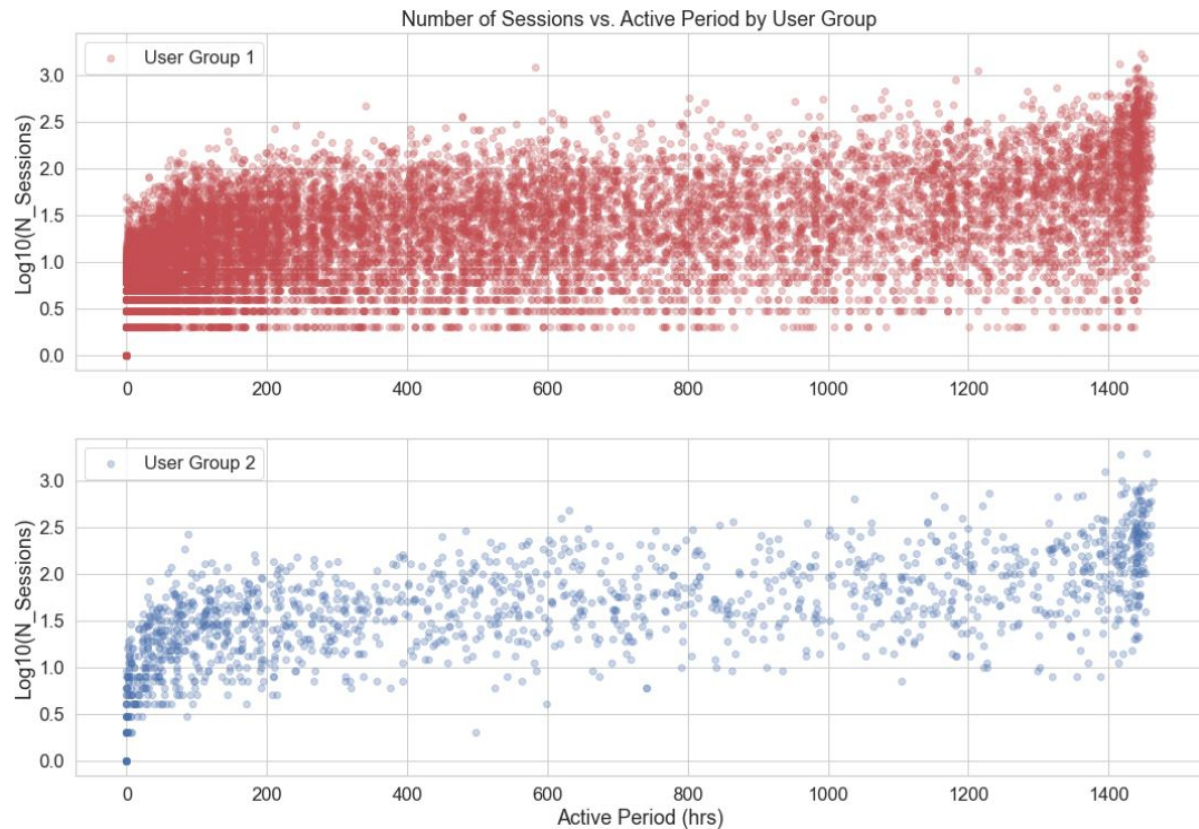
Feature Values Distribution vs. User Group

- Although, there are clear differences in the feature values distributions between Group 1 and Group 2 users, the plots also indicate certain overlap between the two groups



Examine Correlation: Number of Sessions with Active Period

- The two features that are most likely to be correlated are 'Number of Sessions' and 'Active Period'
- The motivation is to examine for type of correlation for some Users 1 similar to that observed for Users 2
- The data show that
 - There is no clear correlation between 'Active Period' length and 'Number of Sessions'.
 - Extremely wide range of 'Number of Sessions' values is observed across the entire 'Active Period' range – note that $\text{Log}_{10}(\text{N_Sessions})$ is used in the plots
 - There is no clear differentiation in the behavior of Users 1 and Users 2
 - In both groups, there is a clustering of users with small 'Active_Period' length, 0–200 hrs, and with large 'Active_Period' length, 1400–1500 hrs



V. Data Modeling

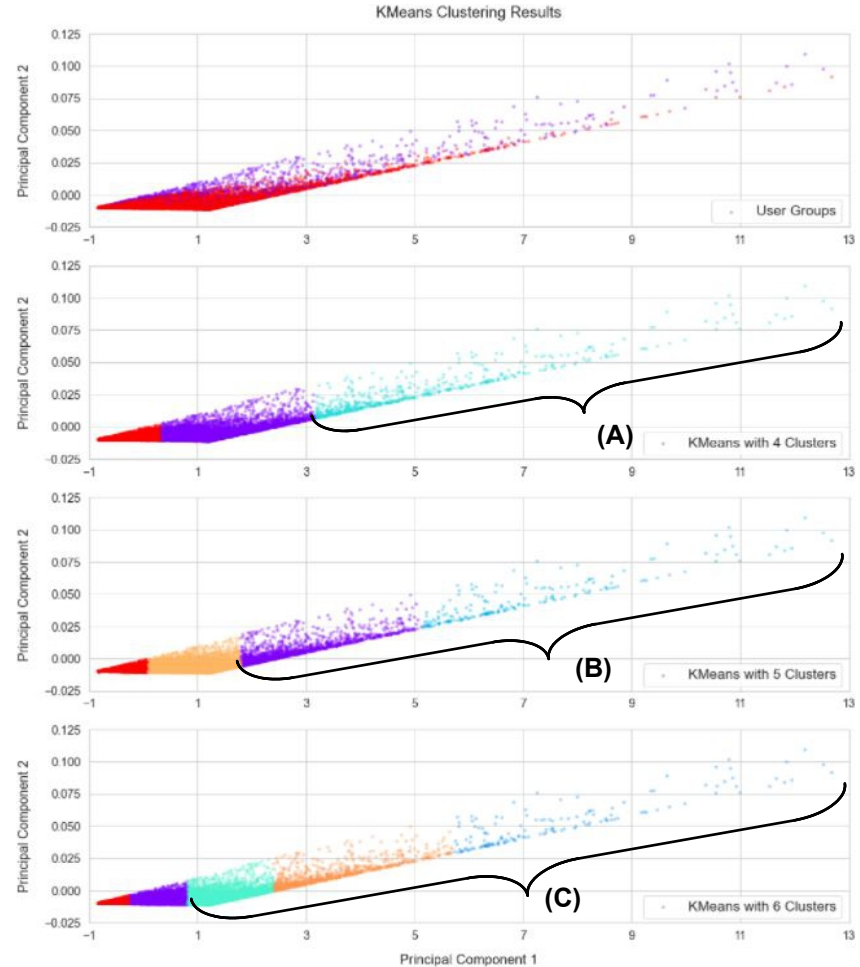


Modeling Process

- K-Means Clustering is used to determine the possibility of some of Users 1 overlapping with Users 2
- In order to gain insight and to build confidence in the model's results, we chose to work with two dimensions only which allows for informative visualization of the results. To achieve this, PCA with two principal components is used before applying the K-Means Clustering model.
- To obtain possible likely, but different, user segmentations we get results from K-Means Clustering model with four, five and six number of clusters

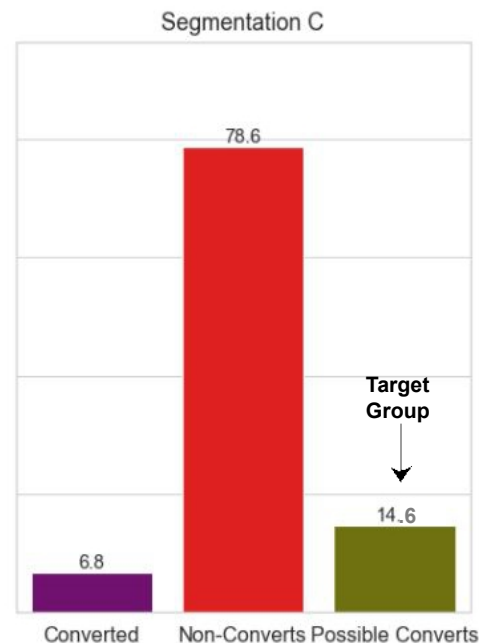
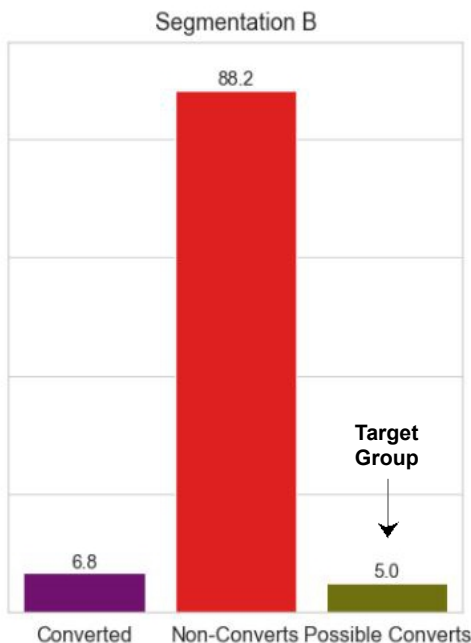
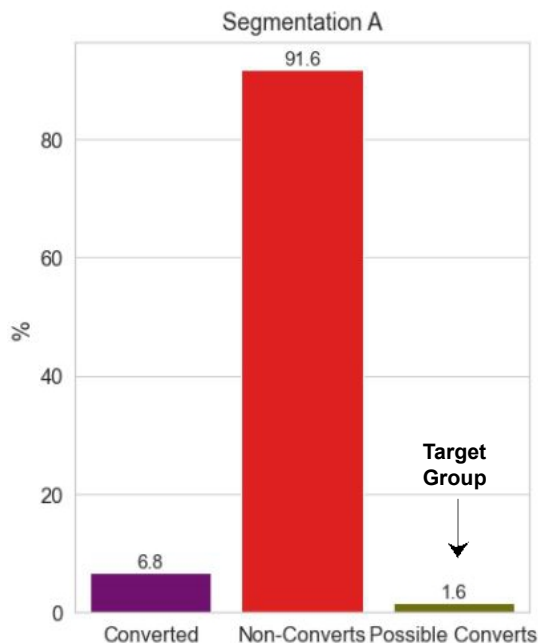
K-Means Clustering Results: Number of Clusters = 4, 5, 6

- Based on the K-Means Clustering results with different number of clusters we create three different user segmentations to determine the target user group
 - Segmentation A:** Most conservative targeting strategy based on model results with four clusters
 - Segmentation B:** Moderate targeting strategy based on model results with five clusters
 - Segmentation C:** Most aggressive targeting strategy based on model results with six clusters



User Partitioning Based on Different Segmentations

- Depending on the user segmentation, the target group for the promotions as percentage of the total number of users changes as follows:
 - Segmentation A:** Target group of 1.6 %
 - Segmentation B:** Target group of 5.0 %
 - Segmentation C:** Target group of 14.6 %



VI. ROI Optimization



Key Parameters and Parameter Ranges for Optimization

- Number of targeted users, n
 - **Segmentation A:** $n = 361$ (out of 22576)
 - **Segmentation B:** $n = 1128$ (out of 22576)
 - **Segmentation C:** $n = 3318$ (out of 22576)

These numbers are obtained from our clustering model results
- Average ROI per converted user over average user lifetime
 - Same for all segmentations
 - Here we use 200 %, 500 %, 1000 %, 2000 % (returns \$3, 6, 11, 21 per \$1 invested)

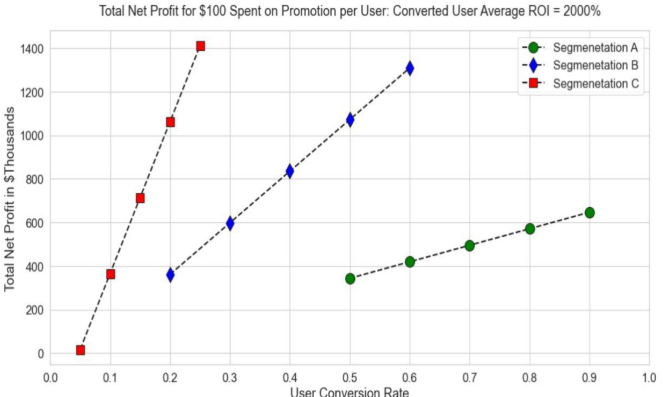
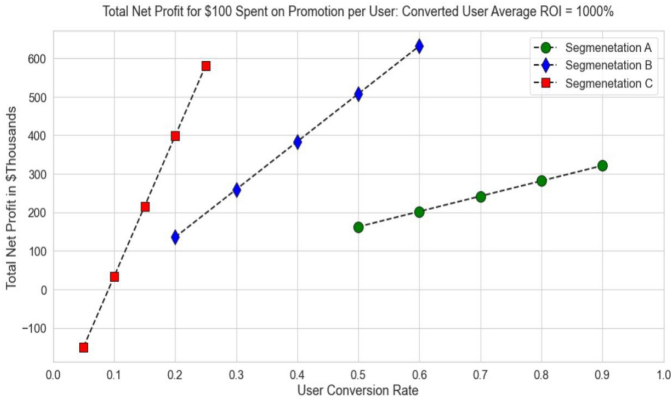
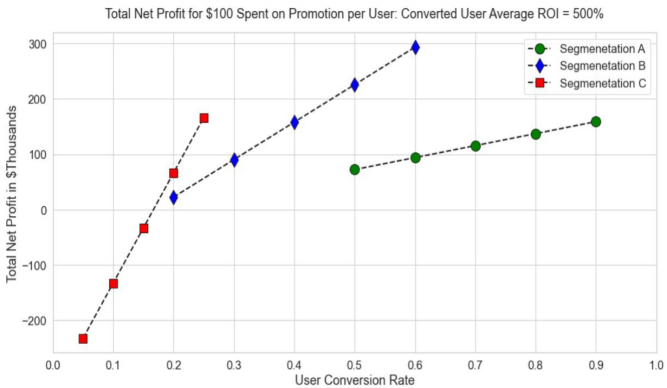
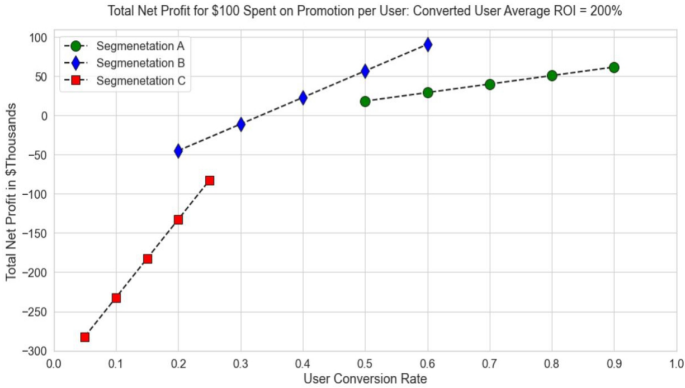
These numbers represent low, moderate and high ROI values to be used in our expected profit estimations
- Targeted users estimated conversion rate
 - **Segmentation A:** 0.50–0.90
This is a reasonable estimated range for users who are most likely to convert
 - **Segmentation B:** 0.20–0.60
This is the approximate estimated range assuming that the conversion rate range for the added users is 0.1–0.5
 - **Segmentation C:** 0.05–0.30
This is the approximate estimated range assuming that the conversion rate range for the added users is 0.0–0.1

Note that the ROI and User Conversion Rate ranges provided here, although realistic, are selected for demonstrating the process of selecting the targeted users for optimal profit strategy. For actual selection, values based on the particular company history and experience should be used.

Expected Profit for Different Parameter Values

Main Observations

- **Segmentation A**
 - Since the conversion rate is greater or equal to 50%, targeting this user segment never results in losses
 - However, it does not have large gains potential even at high ROI of converted users.
- **Segmentation C**
 - Due to small conversion rates and large number of users, losses are observed at low and moderate ROI.
 - At high ROI values, high profit is realized only at the high-end of its conversion rate range
- **Segmentation B**
 - This user segment appears to be the optimal target for promotions.
 - At low ROI, losses occur at the low end of its conversion rate range. However, such low ROI value is not realistic and should not be given great weight.
 - At moderate and high ROI, it outperforms both Segments A and C when comparing the corresponding ends of their conversion rate ranges.



Recommendation: Offer promotions to Segment B users since this is expected to achieve highest possible profit in most of the possible scenarios considered here.

VII. Summary



Summary

- **Data Processing**
 - From the available data, a dataset composed of the most relevant features is created to be used for analysis and modeling.
- **Data Analysis**
 - Data reveals that only 6.8 % of total users currently generate revenue.
 - Feature values distributions show differences between Users 1 (non-converted) and Users 2 (converted) behavior. Yet, some overlap is indicated.
- **Data Modeling**
 - PCA model is used to reduce the number of modeling features into two principal components.
 - K-Means Clustering model with 4, 5, and 6 number of clusters is used to determine three different, but likely, user segmentations.
- **ROI Optimization**
 - Using realistic parameter ranges for the three different user segmentations expected ROI values have been obtained.
 - The optimal marketing strategy is determined to be the one targeting 5 % of non-converted users (Segment B) which have good overlap with converted users and have the potential to generate highest profit with minimal risk of losses for the company.