Homework 3

Information, Impacts and Insights

E-FRI

Asha Mairh

Jeet Patel

Kavish Hukmani

Keshore Suryanarayanan

Neon Zhang

11-11-2021

Contents

1 Executive Summary

Online communities are perceived as a way to increase user engagement in video games. In this study, we examine the effect of the online community introduced by KygnaCell in their game Nicht-Soporific, by statistically evaluating its impact on metrics such as revenue, retention and Customer Lifetime Value (CLV).

There is data available on a set of users with information on their spends before and after the introduction of the online community, their age at the time of the introduction, and their average spend and churn status 3 months after the introduction. Additionally, we are also provided information on the acquisition channel of these users.

We employ various regression models for our analysis and infer that although the user spend increased by $29 in the short term, there was a significant uptick in churn which ultimately resulted in no significant difference in CLV among the joined users. We also observe that the method of customer acquisition has no statistically significant impact on churn or any of the aforementioned predictors used in the analysis.

The increase in churn due to the online community is attributed to negative user interactions and cyber bullying in the community. As a result, we recommend various methods to moderate toxic behavior to enhance user happiness [Martinez-Jerez et al., 2011], which could have a positive effect on retention and thereby the CLV.

2 Introduction

KyngaCell is a mobile gaming company that wanted to improve the revenue, customer retention and Customer Lifetime Value (CLV) for one of its games, Nicht-Soporific. In mid-2021, the company devised a new online community feature which provided the users with avenues to interact within the game both inside and outside the gameplay. There was a notion among the chief executives that this feature helped improve revenue, retention and thereby CLV.

This study aims to corroborate or refute the aforementioned notion with data. We do this by taking a cross-sectional chunk of data from the time when the feature was released and evaluate the spending habits prior to and after the release, among both customers who joined the community and those who did not. Since we also know whether the customers churned or not after 3 months of introducing the online community, we use that to devise a model that predicts the churn with the available information. Early prediction of churn based on customer

profiles is instrumental in devising retention strategies [Milošević et al., 2017]. Ultimately, we want to statistically determine the validity of the executives' notions, quantify the degree of validity, and explore the opportunities provided by the information in the dataset.

3 Problem Formulation

We want to determine the effect of introducing the online community and analyzing how it has affected downstream user behavior. We assess the impact by observing the change in revenue, user retention, and the Customer Lifetime Value (CLV). Additionally, we want to determine if customers onboarded organically have significant differences in the above assessments than the customers who were onboarded via marketing campaigns.

We use Difference-in-Differences (Diff-in-Diff) to determine the change in user revenue post introduction of the online community. Diff-in-Diff enables us to factor in the existing differences between the two user groups while quantifying the impact of the online gaming community on user revenue.

To assess the change in retention, we use logistic regression because the data only indicates if a customer churned within 90 days. Logistic regression considers the ground truth of a user's binary churn status and predicts the probability of churn [Ovchinnikov, 2017] (and thereby probability of retention), which is a continuous variable. We then use this estimated retention to calculate the aggregated CLV. Finally, we extend this framework to determine if there is a significant impact of the mode of onboarding the customers (campaign/organic) by introducing the same as a predictor in the logistic regression model.

4 Data Characteristics

To understand whether the online community feature impacted user revenue, we used the amount spent by each individual a month before and a month after joining the new feature. We are also able to bifurcate this information based on the binary field that indicates whether the user joined the online community or not. This information enables us to determine causality in a quasi-experimental setup with the power of creating a counterfactual situation.

Secondly, to determine impact of the community on retention, we have the ground truth of the churn status of users. Their age with the firm at the time of launching the online community is used as a predictor to create a level-playing field. Also, their average monthly spend in the

subsequent 90 days is available, which helps us determine if it aggravates or alleviates churn.

Finally, we are armed with the information on whether each user in the dataset was acquired organically or inorganically, to analyze if customer acquisition is related to customer retention. However, we are unable to determine the biases that the data could suffer from since we do not have visibility on how the users were sampled. We also do not have visibility of the breadth and depth of user engagement with the new feature prior to churn (if any), and their game activity before and after launching the online community feature. We also do not have information on acquisition costs and they could significantly alter CLV especially for inorganic users.

5 Model Development, Estimation and Results

To evaluate the change in revenue, we split the user spend data into two groups, those who were invited to participate and those who were not and into two time periods, the month before the introduction of the online community and the month after. Using the diff-in-diff model, we were able to infer that joining the online community resulted in a statistically significant increase in spend of $29.02.

We use logistic regression to estimate the customer churn. Customer age at the time of launching the online community, their average spend during the last 90 days, and a flag indicating if a customer joined the online community as explanatory variables are used to predict user churn propensity. We observe that only joining the online community makes a statistically significant impact on churn. The logistic regression model helps predict churn probability from which we obtain retention probability for CLV calculations. Following is the model:

$$log(odds_{churn}) = 0.92Joined - 0.06CustomerAge - 0.003AverageSpend \quad (1)$$

To estimate the difference in retention rates between the two groups, we run a two-sample t-test and observe that joining the online community reduces the retention rate by 19%, which is statistically significant. We use the individual retention rates and a margin of 50% of customer's spending in the time period to calculate each user's aggregated CLV, incorporating the discount factor of 0.1% as applicable in mid-2021, when the feature was introduced. We again run a two-sample t-test for differences in CLV between the two populations (joined and not joined) and observe no statistically significant difference.

The effect of mode of customer acquisition (organic/campaign) is assessed by evaluating its

impact on retention and CLV. Before forming the logistic model that incorporates customer acquisition channel in determining/predicting churn, multiple techniques were employed to check its independence with the existing variables. As a result, we see that the binary churn status, the average spend in 90 days after the introduction of the community, and customers' ages are independent of customer acquisition channel. In this way, we form the following equations to investigate the influence.

$$log(odds_{churn}) = 0.36 + 0.18Campaign - 0.05CustomerAge + 0.93Joined - 0.003AverageSpend \tag{2}$$

The updated CLV for each customer is calculated and upon the performance of two-sample t-test between the campaign and organic users, we observe no statistically significant difference.

6 Recommendations and Managerial Implications

Based on the results, we see that the customers who join the online community are willing to spend more in the short term but do not retain on the game in the long term. As a result, their aggregated value provided to the firm is not significantly different from the users who were not invited to the community. So, the business could focus on retaining these users over a longer period [Lee et al., 2020], which would in turn help derive more aggregated value from them.

This could be achieved by improving their experience in the online community. Online gaming communities are subject to cyber bullying [Kwak et al., 2015] [Fu,] and toxic interactions which lead to users churning. The business could implement moderation of online communities, and deploy content filtering mechanisms to increase accountability of users who partake in bullying. Additionally, KygnaCell could allow users to restrict interactions in the community to intra-team for inside the game, and within user-curated networks for outside the game. While we do not have the data to quantify or substantiate this recommendation, it is based on correlating observed results with proven research in the field of gaming.

Furthermore, since there are no significant differences between organic and inorganic users, the retention efforts need not be bifurcated based on method of acquisition.

7 Conclusion

We analyzed the effect of introducing an online community in a mobile game in three key areas - revenue, user retention and customer lifetime value. Based on the given data, we were able to infer that although user revenue increased in the short term, the user retention decreased. This led to no significant change in the CLV. We also noticed that the method of customer acquisition did not affect these changes in behaviour.

While the decrease in retention might seem odd at first glance, it could be explained by the fact that online communities, especially in video games, breed toxicity if not moderated properly. This toxicity leads to higher churn rates. Based on this we suggest taking measures such as moderation of the community and user-defined restrictions on on-game and off-game communication to promote healthier interactions among community members, which could enhance user experience and reduce churn.

References

[Fu,] Fu, D. A look at gaming culture and gaming related problems: From a gamer's perspective.

[Kwak et al., 2015] Kwak, H., Blackburn, J., and Han, S. (2015). Exploring cyberbullying and other toxic behavior in team competition online games. ACM.

[Lee et al., 2020] Lee, E., Kim, B., Kang, S., Kang, B., Jang, Y., and Kim, H. K. (2020). Profit optimizing churn prediction for long-term loyal customers in online games. *IEEE Transactions on Games*, 12(1):41–53.

[Martinez-Jerez et al., 2011] Martinez-Jerez, F. d. A., Steenburgh, T. J., Avery, J., and Brem, L. (2011). Hubspot: Lower churn though greater chi. *Harvard Business School Accounting & Management Unit Case*, (110-052).

[Milošević et al., 2017] Milošević, M., Živić, N., and Andjelković, I. (2017). Early churn prediction with personalized targeting in mobile social games. 83:326–332.

[Ovchinnikov, 2017] Ovchinnikov, A. S. (2017). Predicting customer churn at qwe inc. *Darden Business Publishing Cases*.

A Logistic Regression to predict churn

We use average customer spend for last 90 days, age of customer when the community was launched, and flag to indicate if a customer was part of the online community or not as explanatory variables to predict customer churn.

```
##
## Call:
## glm(formula = Churned ~ Avg_Spend + Customer_Age + Joined, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6753  -1.2113   0.7973   1.0979   1.2894
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.474797   0.523983   0.906  0.36487
## Avg_Spend    -0.002819   0.005655  -0.498  0.61815
## Customer_Age -0.055849   0.071598  -0.780  0.43537
## Joined        0.916584   0.355287   2.580  0.00988 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 268.95  on 198  degrees of freedom
## Residual deviance: 260.42  on 195  degrees of freedom
## AIC: 268.42
##
## Number of Fisher Scoring iterations: 4
```

Figure 1: Logistic Regression to predict Customer Churn

A.1 Model Evaluation

A.1.1 Confusion Matrix

	Predicted Retention	**Predicted Churn**	**Total**
Actual Retention	25	56	81
Actual Churn	18	100	118
Total	43	156	199

Table 1: Confusion Matrix of Churn vs Retention

Based on the above confusion matrix, we observe the precision of the model is 64.1%, the recall is 84.75% and the F-1 score of the model is 72.99%

A.1.2 ROC-AUC Curve

The ROC AUC for the logistic regression model is 62%. This implies that the model is about 12% better than a naive model baseline at predicting customer churn

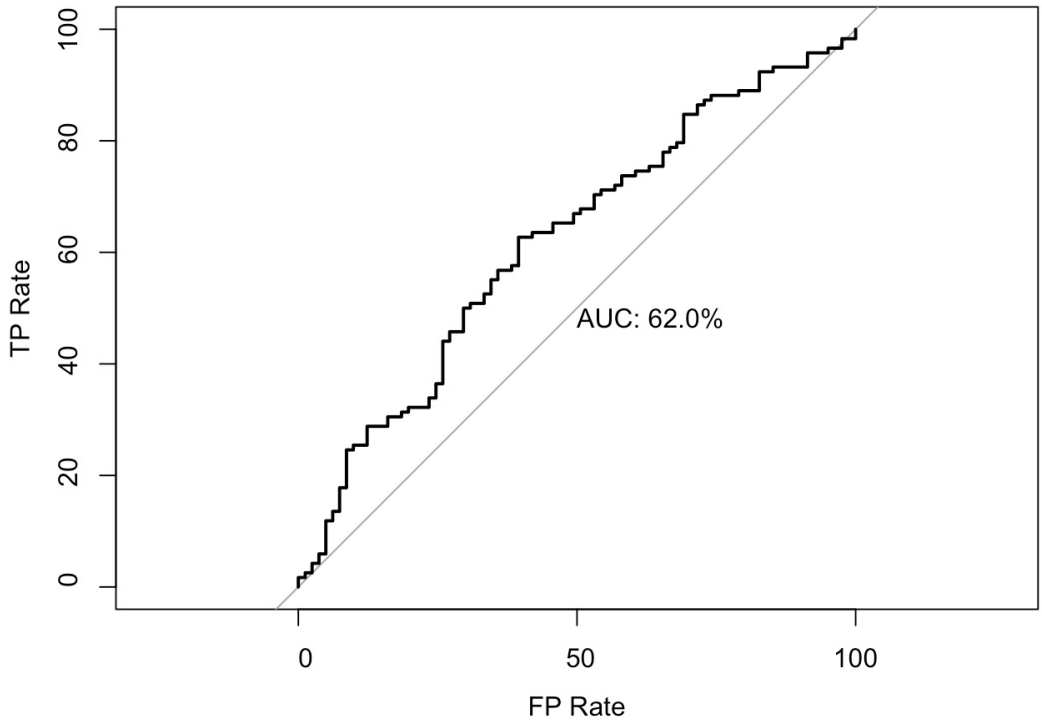

Figure 2: ROC AUC curve for logistic regression

B Difference in Population

The population under consideration for the subsequent analysis are customers who joined the online community and the customers who didn't join the online community.

B.1 Revenue

We used Diff-in-Diff to model the effects of introducing an online community on user spend. Tables 2, 3, 4 and 5 contain the relevant statistics from this model.

Regression Statistic	**Value**
Multiple R	0.583745137
R Square	0.340758386
Adjusted R Square	0.335738779
Standard Error	38.58521917
Observations	398

Table 2: Diff-in-Diff Regression Statistics

	df	SS	MS	F	Significance F
Regression	3	303207.6137	101069.2046	67.88548183	2.09234E-35
Residual	394	586594.7406	1488.819138		
Total	397	889802.3543			

Table 3: Diff-in-Diff ANOVA Summary

	Coefficients	Standard Error	t Stat	P-value
Intercept	70.37606838	3.567204774	19.72863147	9.38493E-61
Time Period	30.87179487	5.044789372	6.119540896	2.26723E-09
Intercept	17.75807797	5.557092842	3.195569782	0.001507997
Time Period x Joined	29.01844903	7.858916065	3.692423839	0.000253448

Table 4: Diff-in-Diff Summary

Group	Monthly Spend
Uninvited users pre online community launch	$70.37
Invited users pre online community launch	$88.13
Uninvited users post online community launch	$101.25
Invited users post online community launch	$148.02
Invited users post online community launch if it would have no effect	$119.01

Table 5: Diff-in-Diff Interpretation

B.2 Retention

We run a two sample t-test to estimate in the difference in retention rates among the population. We first check the equality of variances to determine the test to be used.

```
##
##  F test to compare two variances
##
## data:  online_retention and non_online_retention
## F = 0.77675, num df = 81, denom df = 116, p-value = 0.2272
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5225057 1.1717649
## sample estimates:
## ratio of variances
##          0.7767537
```

Figure 3: F-test for checking equality of variance in Retention Rate

The p-value on running the F-test is 0.2272. Hence we can conclude that the variance in the retention rates across both population are equal. We now proceed with conducting the t-test to determine if the differences in the retention rates are significant or not, if any.

```
##
##  Two Sample t-test
##
## data:  online_retention and non_online_retention
## t = -42.751, df = 197, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1869779
## sample estimates:
## mean of x mean of y
## 0.2926829 0.4871795
```

Figure 4: T-test for checking difference in Retention Rate

We observe that the the p-value of running the t-test is extremely significant and thus we can conclude that the retention rates for customers who joined the online community is on average 19.45% lower than the customers who didn't join the online community.

B.3 Customer Lifetime Value

We estimate the Customer Lifetime Value of the customers using the following equation.

$$CLV = \sum_{t=1}^{T} \frac{mr^{t-1}}{(1+i)^{t-1}} \tag{3}$$

On extending the T to a very large value ($T = \infty$), the above equation reduces to,

$$CLV = m\left(\frac{1+i}{1+i-r}\right) \tag{4}$$

We then run a two sample t-test to estimate in the difference in the customer lifetime value among the populations. We first check the equality of variances to determine the test to be used.

```
##
##  F test to compare two variances
##
## data:  online_clv and non_online_clv
## F = 0.4871, num df = 81, denom df = 116, p-value = 0.0007101
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3276611 0.7348089
## sample estimates:
## ratio of variances
##           0.487099
```

Figure 5: F-test for checking equality of variance in Customer Lifetime Value

The p-value on running the F-test is 0.0007101. Hence we can conclude that the variance

in the retention rates across both population are not equal and we use Welch's t-test to to determine if the differences in the customer lifetime value are significant or not.

```
##
##  Welch Two Sample t-test
##
## data:  online_clv and non_online_clv
## t = 0.98195, df = 197, p-value = 0.1637
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -2.309519       Inf
## sample estimates:
## mean of x mean of y
##  70.27333  66.89194
```

Figure 6: Welch's T-test for checking difference in Customer Lifetime Value

We observe that the the p-value of running the t-test is 0.1637. Hence we can conclude that no significant difference exists in the customer lifetime value of customers who joined the online community compared to the customers who didn't join.

B.4 Discount Rate Calculation

Calculating the CLV is another application of the discounted cash flow (DCF) analysis, assessing the current value of all the potential customer value. For the particular situation of the KyngaCell, the objective of this analysis is to evaluate the viability of online community features. Typically, the weighted average cost of capital (WACC), the historical average returns of a similar project, or risk-free rate of return can be the proxy of discount rate. Whereas, since the company information remains unknown, and the risk-free rate of return will underestimate the risks faced by the company when placed in the context of the feasibility of a commercial project, 3-Month Interbank Rates for the United States, as seen in table 6, that serves as the benchmark interest rate reflecting market conditions is applied to quantify the present value. According to the background of this campaign, the new online community feature was introduced several months before the mid-2021, in this way, the average interest rate for the first six months of 2021 was used as a substitute.

Month	Rate/%
2021-01	0.14
2021-02	0.11
2021-03	0.10
2021-04	0.11
2021-05	0.10
2021-06	0.09

Table 6: 3-Month Interbank Rates for the United States

C Effect of Acquisition Method

C.1 Independence Test

Before forming the logistic model assessing how people join the game touches on the intention of churning, prop-test, t-test and simple linear regression are employed to check whether there are dependencies between variables.

Acquisition method and community invited people- After the probability test was employed to check whether there is a difference between how users joined the game among those who were invited to the online community, we have insufficient evidence to reject the null hypothesis, which means that acquisition method and people who participate in the online community are independent.

Acquisition method and churning result- Probability test was applied to examine whether the acquisition method has an impact on the churning result. Under any confidence level, it cannot be concluded that such an impact exists.

Acquisition method and average spending in the last 90 days- Indicated by the simple linear regression, we find that either the model or the coefficient of the acquisition method is not valid in any confidence level. In this way, how people join the game does not affect how much they spend after 90 days introducing the online community.

Acquisition method and spending in 1 month before and after the community is launched- Using the t-test, the results are both insignificant, which means how people join the game and how much they spend before and after the initiation of the community are irrelevant.

Acquisition method and customers' ages- The insignificant result of the t-test of whether there is significant age gap between people who participate in the campaign and those who participate organically shows that these two variables are independent.

Eventually, we use the following model to calculate the churn rate:

$$Churned = \beta_0 + \beta_1 Campaign + \beta_2 CustomerAge + \beta_3 Joined + \beta_4 AverageSpend \quad (5)$$

C.2 Interpretation of Confusion Matrix

Because the loss to the company of giving up on the customers who would not churn is higher than the system maintenance cost of accommodating more people who will be churn, the cost of the false positives and false negatives are different. In this way, the accuracy is inappropriate to describe the functionality of the model. Besides, it is insufficient to only have the ability to recall the churning people, identifying the real churning people is also important, so we use the F1 score to balance it.

	Predicted Retention	**Predicted Churn**	**Total**
Actual Retention	22	59	81
Actual Churn	15	103	118
Total	37	162	199

Table 7: Confusion Matrix of Churn vs Retention

Among all the customers who are labeled as churning after 3 months after the launch of the community, 63.6% of them actually churned.

Among all the customers who truly churned after 3 months after the launch of the community, 87.3% of them are labeled correctly.

Eventually, our model has an F1 Score of .74.