# An Analytical Deep Dive: Decisions, Reflections, and Recommendations

Emily Barrett, Ellie Cummings, Sami Howa, and Emily Nguyen

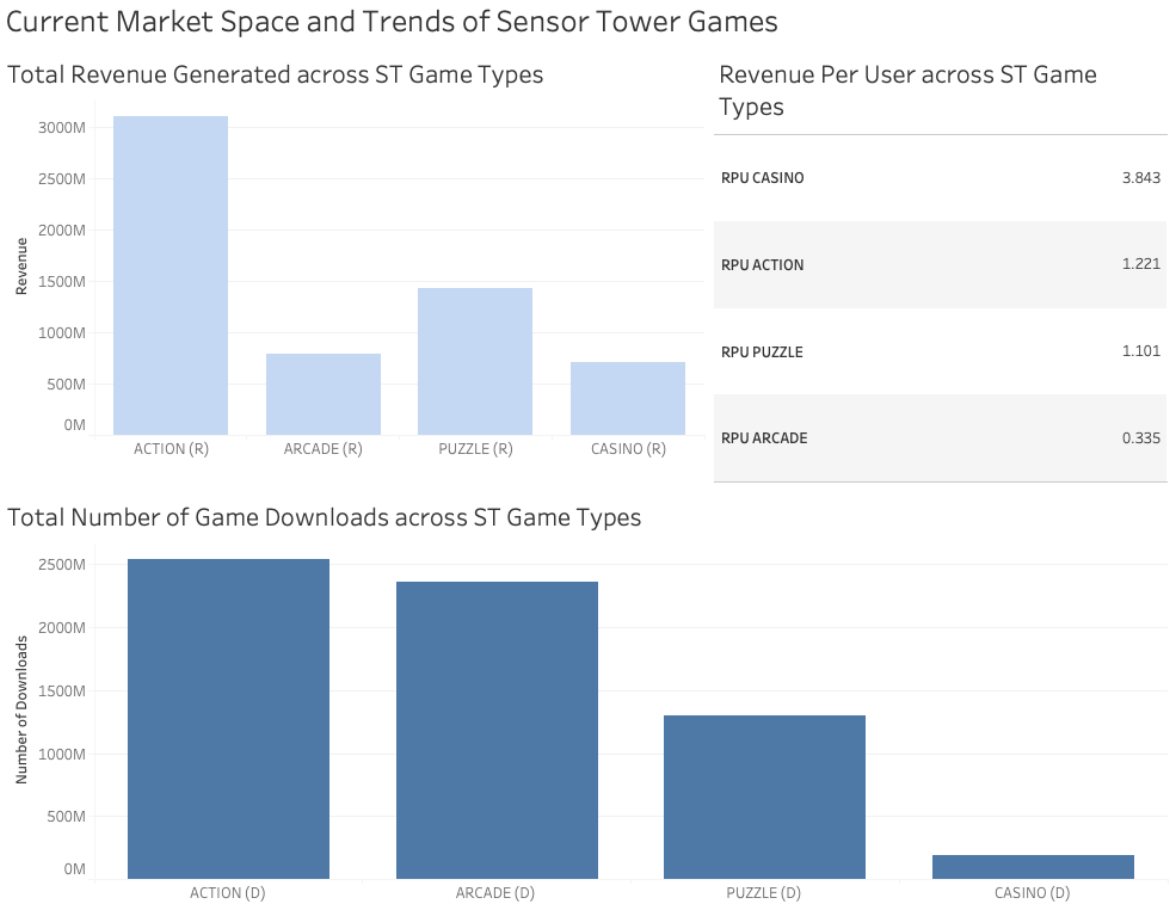
Table of Contents

## Executive Summary

Taking on the role of data analysts, we use statistical and qualitative analysis to provide our recommendation on whether or not to release a game in beta worldwide. By analyzing the overall market space, we were able to see that puzzle games do not perform as well as other types of games, like Arcade and Action. We then took a closer look at how individual regions and networks were performing to identify where problem areas are occurring. Then, we delved into forecasting efforts to predict where we will be by days 90 and days 180. Through these calculations, we analyzed our standing in comparison to successful competitors in the marketplace, which showed us the lack of success that our game has generated. An important component of making this final decision was looking at user engagement, and we developed multiple models to predict user churn. Our game showed high rates of churn, which when compared to mobile games worldwide we saw that our game is under performing in that area as well. As a final step in this process, we took a step back to analyze the completeness of our models, as well as determining if we can improve them by branching out from the traditional models created in the process. After looking at all of our calculations, analysis, and models we came to the final decision that this game should not be released into a worldwide market. The low predicted revenue and user accumulation, high churn rate, and lack of success compared to the games that have launched worldwide give us the evidence to not recommend this launch, and spend our time and money on improving the game to ensure success in the future.

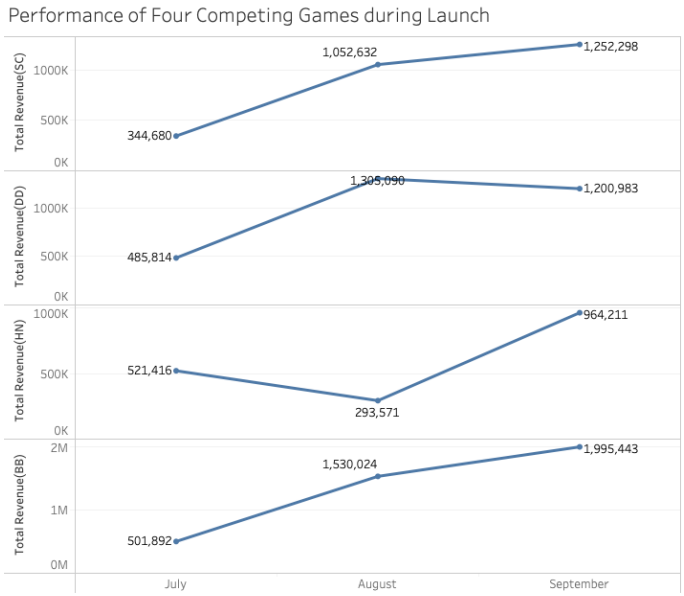
**Introduction to Business Analytics in Gaming**

At the beginning of this project the goal was to analyze the current market space, make an executive decision about whether a puzzle game should be released based on that data, and then further analyze how successful games in the market space are doing. From initial glances at the data, the metrics to use to measure the performance of the different games in the market space included revenue generation, number of downloads, and revenue per user based on game type. These calculations led to the creation of the dashboard below, which showed clear and interpretable data to answer the questions stated previously.




This data showed that action games generate the most revenue of all the game types, followed by puzzle, arcade, and casino respectively. Similarly, this trend is mirrored in the graph that shows the total number of game downloads across the game types, demonstrating that action games have the most downloads, followed by arcade, puzzle, and casino games respectively. When trying to answer the business question we defined, we realized that knowing how much revenue is generated per user would help us understand whether or not introducing a puzzle game into the current market space was a good idea or not. From the dashboard, the trend of RPU showed us that puzzle games are not revenue generators in the current market space, and that with the two graphs and the trends they described, we would not recommend the launch of a new puzzle game.

Continuing in the analysis of what the current market space looks like, we took the four successful games and compared the trend of their revenue generation overtime.




This graphic gave us substantial information that showed how successful games would perform in the market space. The gradual increase in Total Revenue generated was a testament to their success, and gave us baseline information that we needed to begin making recommendations for the game in beta. Based on this preliminary data, we concluded that a puzzle game would not perform well in this market space. The low revenue per user and low total downloads led us to that conclusion.

The four competitor games we want to measure ourselves against have proven that they are performing well in the marketspace. To show their overall statistics and create a baseline for comparison, we looked at their revenue per user and the total count of players over a period of 90 days. An important exception to this is the game Hello Neighbor, which had only 60 days of data in comparison to the three other games.

Simon's Cat has the highest revenue per user, with $.74 generated for each player. This declined moving to Diamon Diaries, with about $.60 generated for each user, followed by Brick n Balls with $.31 per user and lastly Hello Neighbor, with $.13 per user. These metrics give us a good idea of how we want the game in beta to perform in order to reach the benchmarks of success that these four games have reached after being launched worldwide. Refer to Appendix A to see Excel results of competing games RPU overall.

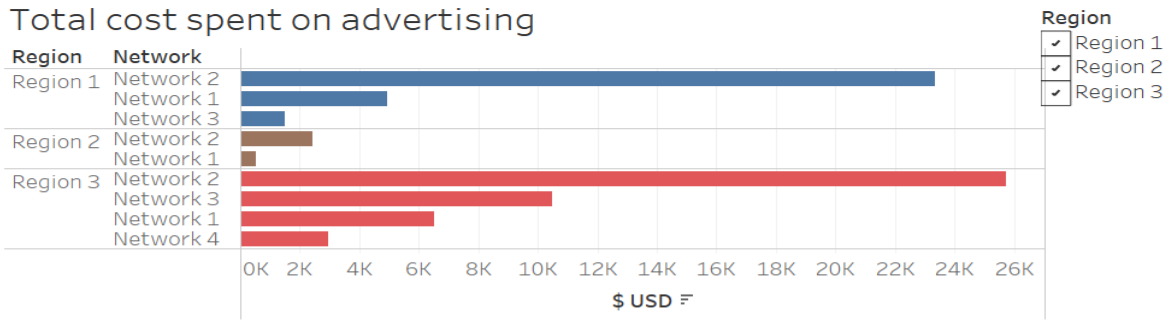
**Game Understanding**

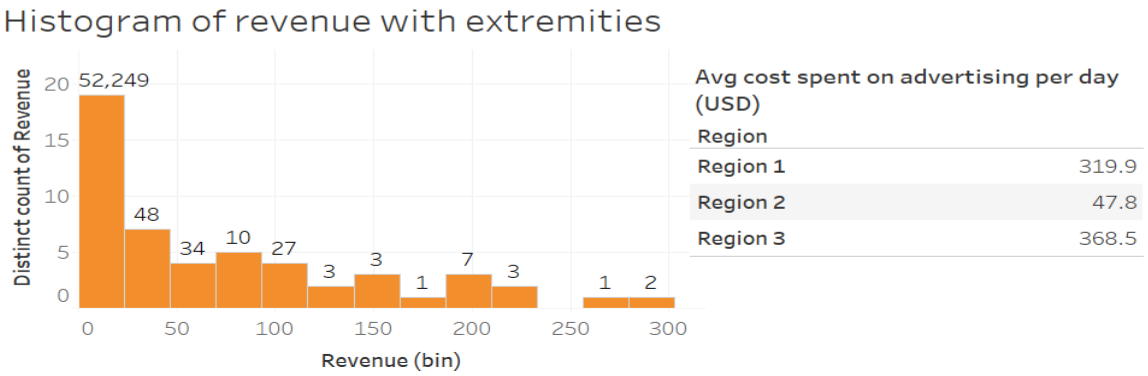
After our initial analysis on the general marketspace trends, we wanted to address areas where advertising cost is not spent enough with the goal of gaining more attention to our games. The question we wanted to address in this next section was: “In which region should we focus on spending more advertising costs so that we can attract more customers?”.

When cleaning the final_user_extract dataset, certain users that had no data on the amount of revenue spent were omitted from the data. There were a total of 720 cases that had no data for revenue out of 53,257 data points. This equated to 1.35 % of the data, which we omitted completely. If a user was missing data on either the network or region that they originated from, but there were multiple data entries for that user, then we filled in the data with either their region or network accordingly as these are case specific and can be derived from the surrounding data. This was the case unless a user only opened the app once in the 31 day

period, then there was no way to determine from which region or from which network they downloaded the game from. Data from these cases were omitted from the data set as well.

When looking at the ad spending data, we made the decision to omit all the advertising expense data for Region 2, Network 3. The average values were extremely small compared to the average spendings in other regions and networks (see chart below for a comparison between the other regions and networks.)




We included these graphs in our dashboard to display the costs that are going into advertising and what we are mostly receiving from these spendings. As seen in our graph at the top of the dashboard, region 3 is spending the most overall in advertising costs. They also have additional spendings in Network 4 (a network that region 1 and 2 do not have data for). Each region spends the most on advertising costs in network 2. Region 2 is spending the least amount of money in advertising costs but they also have the lowest RPU (which is displayed in the “Revenue per User” section). As seen in the table on the bottom right, they are spending a significant amount less than the other regions per day. Overall, our advertising costs are allocated more into region 1 and 3. The histogram on the bottom shows us that most of the revenue generated between the beta period is in between the range of $0-$50. This data resulted in a right-skewed histogram with few points falling between $125-$300. We can conclude from these findings that the majority of our advertising expenses are resulting in multiple small earnings.

The overall user retention rate during the beta period was 22.4%. In order to find the user retention rate, we calculated the number of users who returned to playing the game after the install date, i.e when the event_dt is greater than the install_dt, divided by the total number of users during the beta period.

To begin the calculation, we did an advanced filter on the user column by filtering by the unique values only in order to count how many users participated in the beta period of the game. This gave us a value of 27,856 unique users. Second, we needed to find out how many repeat users participated in the beta period. In order to find this, we calculated the number of times each unique user played the game after the date of install. In order to find these users, we counted how many users had an event_dt that equaled install_dt, meaning they only played the game on the day of install and never returned again during the beta period. The value for the users who only had an event_dt equal to instal_dt was 21,617. Next, we subtracted 21,617 from the total unique user value of 27,856 to get 6,239 users. The 6,239 users represent all the users who had event_dt greater than the install_dt, meaning they installed the game on a particular day and then returned to play the game on subsequent day(s) after install. Finally, we divided 6,239 by 27,856 to reach the 22.4% user retention rate for the beta period.

During the beta period, the revenue per user is $.56. We calculated this metric by filtering the list of users to show unique records only, then using COUNT to calculate the total number of distinct users. Then we used SUM to get the total amount of revenue spent by users on the game. To calculate Revenue per User, we divided Total Revenue Spent by Total Number of Users. Refer to Appendix B for RPU of the game during beta overall then broken up into 3 regions. Region 2 had the lowest RPU of $.31, while Region 1 and Region 3 were higher than average with RPU values of $.65 and $.60 respectively.

Below is the breakdown of the return on investment for each network. The top table is the total sum that each network spent on advertising the game in beta. This metric was calculated by sorting the data by network, then summing the revenue generated for each one. Below that table is the sum that users spent on the app based on where they downloaded the app from. When a user downloaded the app from an ad on Network 1, their contribution is included in the sum of the money users spent if they downloaded the app from that same Network. Using the same technique, we sorted the data by network, then summed the user's money spent to generate these numbers.

| Total Spent N1 | Total Spent N2 | Total Spent N3 | Total Spent N4 |
|---|---|---|---|
| 11954.3 | 51524.51 | 12003.11 | 2967.26 |

| Total User Spent N1 | Total User Spent N2 | Total User Spent N3 | Total User Spent N4 |
|---|---|---|---|
| 2560.49 | 6039.83 | 297.34 | 68.84 |

| | Return on Investment | Return on Investment % |
|---|---|---|
| Network 1 | 0.2142 | 21.42% |
| Network 2 | 0.1172 | 11.72% |
| Network 3 | 0.0248 | 2.48% |
| Network 4 | 0.0232 | 2.32% |

The bottom table shows the return on investment for each network. Network 1 had the highest return in the beta period, with 21.42% ROI after 31 days. This was followed by Network 2, having 11.72% ROI. Network 3 and Network 4 had very low ROI in the beta period with only 2.48% and 2.32% respectively.

It is difficult to compare our data to that of our competitors since they have launched worldwide while our game has only launched in 3 small regions. Additionally, we only have data that covers the 31 days of the game's beta period unlike our competitors who have data covering a quarter of a year. Given this limited dataset, we can deduce that our comparisons will

be skewed. These differences have been taken into consideration and we have made additional calculations to minimize discrepancies. To do this, we calculated the RPU for 3 small regions of our competitors during their first 31 days after launch (we excluded all the data following those first 31 days). The 3 regions we used to represent our competitors were Scandinavia, the British Isles, and Central America. We excluded the US since we have not yet launched there. We also compared our 3 regions' RPUs with the RPUs of individual countries of our competitors to gather a broader visual of how our game is performing. The results are shown below (cells highlighted in green represent the RPUs in regions where we have released our puzzle game-not our competitors).

**Revenue Forecasting**

The business question we are addressing in this section is: "Should we focus more on targeting networks and regions that bring in the most users or the most revenue? And, is an increase in user count per day directly correlated with an increase in revenue?" This question will call attention to the areas that are performing better in comparison to others and help us narrow down which areas we need to be targeting more for marketing purposes in order to improve our game performance in the future.

To calculate the predicted revenue for each network, we took the columns with event_dt, revenue, and network and combined them into a new spreadsheet. From there, we ordered the data by network, then separated the associated data into 5 excel sheets for each network which included the data on each day and how much revenue was collected that day. To calculate the total sum of revenue on a given day, we used the SUMIF() formula. From there, we partitioned the data using the time series partition, then performed linear regression to calculated D-90 and D-180 forecasts.

**Network 1**

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -503.7995484 | -679.8235445 | -327.7755522 | 86.06563905 | -5.8536665 | 2.3814E-06 |
| Day | 85.05845968 | 75.45556341 | 94.66135595 | 4.695265544 | 18.11579321 | 2.3552E-17 |

This output from time series analysis gives the the following regression equation:

$$Predicted\ Revenue\ =\ -503.80\ +85.06(day)$$

D-90 Predicted Revenue

$$-503.80\ +85.06(90)\ =\ \$7151.60$$

D-180 Predicted Revenue

$$-503.80\ +85.06(180)\ =\ \$14807.00$$

On day 90, users that joined from Network 1 are expected to generate $7151.60 in revenue, and on day 180, users are expected to generate $14807.00 in revenue.

**Network 2**

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -1146.072516 | -1414.038821 | -878.1062111 | 131.0201551 | -8.74729934 | 1.2541E-09 |
| Day | 226.0704194 | 211.4516589 | 240.6891798 | 7.147735458 | 31.62825774 | 4.9594E-24 |

This output from time series analysis gives the the following regression equation:

$$Predicted\ Revenue\ =\ -1146.07\ +226.07(day)$$

D-90 Predicted Revenue

$$-1146.07\ +226.07(90)\ =\ \$19200.23$$

D-180 Predicted Revenue

$$-1146.07 + 226.07(180) = \$39546.53$$

On day 90, users that joined from Network 2 are expected to have generated $\$19200.23$ in revenue, and on day 180, users are expected to have generated $39546.53 in revenue.

**Network 3**

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -49.14503226 | -63.51431779 | -34.77574673 | 7.025756538 | -6.99498082 | 1.0844E-07 |
| Day | 11.85991129 | 11.0760025 | 12.64382008 | 0.383286442 | 30.94268407 | 9.199E-24 |

This output from time series analysis gives us the following regression equation:

$$Predicted\ Revenue = -49.15 + 11.86(day)$$

D-90 Predicted Revenue

$$-49.15 + 11.86(90) = \$1018.25$$

D-180 Predicted Revenue

$$-49.15 + 11.86(180) = \$2085.65$$

On day 90, users that joined from Network 3 are expected to have generated a total of $1018.25 in revenue, and on day 180, users are expected to have generated a total of $2085.65 in revenue.

**Network 4**

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -18.5076129 | -25.8873097 | -11.1279161 | 3.608248506 | -5.12925118 | 1.771E-05 |
| Day | 2.702955645 | 2.300360179 | 3.105551112 | 0.196846094 | 13.7313146 | 3.2072E-14 |

This output from time series analysis gives the following regression equation:

$$Predicted\ Revenue = -18.51 + 2.70(day)$$

D-90 Predicted Revenue

$$-18.51 + 2.70(90) = \$224.49$$

D-180 Predicted Revenue

$$-18.51 + 2.70(180) = \$467.49$$

On day 90, users that joined from Network 4 are expected to have generated a total of $224.49 in revenue, and on day 180, users are expected to have generated a total of $467.49 in revenue.

**Organic**

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -1313.690323 | -1847.014514 | -780.3661313 | 260.7649431 | -5.03783333 | 2.2834E-05 |
| Day | 225.023125 | 195.927905 | 254.118345 | 14.22589393 | 15.81785482 | 8.4742E-16 |

This output from time series analysis gives us the following regression equation:

$$Predicted\ Revenue = -1313.69 + 225.02(day)$$

D-90 Predicted Revenue

$$-1313.69 + 225.02(90) = \$18938.11$$

D-180 Predicted Revenue

$$-1313.69 + 225.02(180) = \$39189.91$$

On day 90, users that joined Organically are expected to have generated a total of $18938.11 in revenue, and on day 180, users are expected to have generated a total of $39189.91 in revenue.

**Observations:**

Out of the five network types, users that joined through Network 2 are expected to have generated the highest amount of revenue by day 90 and day 180. The lowest predicted revenue comes from Network 4, with only a predicted total revenue of $224.49 by day 90 and $467.49 by day 180.

To calculate the predicted revenue per region, we first sorted the event_dt from day 1 through day 31 and then sorted by region. We then separated the data into three sections for regions 1-3. This showed us the revenue from each region on the given event_dt. Once each region was separated out, we used the SUMIF function to sum the revenue by each day so that we had a column for days 1-31 and the second column showed the total revenue for that day using the SUMIF function. Once this was completed for each region, we used the data mining time series partition to partition the day and revenue columns for each region into 60% training data and 40% validation data. After the partition was created for the regions, we ran the data mining linear regression function on the partitioned data, which resulted in giving us the coefficients for the linear regression equation for each region as seen below. Once the equation was made, we substituted 90 and 180 into the equation to arrive at the predicted revenue per region for day 90 and day 180.

**Region 1**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -2328.886774 | -3071.294742 | -1586.478806 | 362.9949189 | -6.4157558 | 5.1251E-07 |
| day | 409.2276613 | 368.7259862 | 449.7293364 | 19.80299634 | 20.6649365 | 6.7539E-19 |

This output from the time series analysis gives us the following regression equation:

*Predicted Revenue* = -2328.89 + 409.23 (*day*)

D-90 Predicted Revenue

-2328.89 + 409.23 (90) = $ 34,502.81

D-180 Predicted Revenue

-2328.89 + 409.23 (180) = $ 71,332.51

**Region 2**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -456.0535484 | -635.0726772 | -277.0344196 | 87.53008713 | -5.210249 | 1.414E-05 |
| day | 77.48139113 | 67.71509698 | 87.24768528 | 4.775157737 | 16.2259333 | 4.3533E-16 |

This output from the time series analysis gives us the following regression equation:

*Predicted Revenue* = -456.05 + 77.48 (*day*)

D- 90 Predicted Revenue

-456.05 + 77.48 (90) = $6,517.15

D-180 Predicted Revenue

-456.05 + 77.48 (180) = $13,490.35

**Region 3**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -284.0034194 | -370.582705 | -197.4241337 | 42.33230529 | -6.7089051 | 2.3253E-07 |
| day | 68.31479032 | 63.59150306 | 73.03807758 | 2.309416587 | 29.5809733 | 3.2637E-23 |

This output from the time series analysis gives us the following regression equation:

*Predicted Revenue* = -284.003 + 68.31 (*day*)

D- 90 Predicted Revenue

-284.003 + 68.31 (90) = $ 5,863.90

D-180 Predicted Revenue

-284.003 + 68.31(180) = $ 12,011.80

**Observations:**

According to our calculations, revenue is significantly higher for Region 1 than Regions 2 and 3 for both 90 and 180 day data.

**User Count- Network**

To calculate the user count for each network, we separated the data by network into 5 different excel sheets. To remove duplicate USER ID numbers, we used the Table Tools to Remove Duplicates from the table to accurately get a count of how many users joined each day. Then, using the COUNTIF() function, we performed the same process with the revenue of counting how many users joined each day for each network. After this, we proceeded to use a time series partition and perform linear regression to calculate predicted user count for D90 and D180.

**Network 1**

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -28.3483871 | -88.43435552 | 31.73758133 | 29.3785926 | -0.96493346 | 0.34255636 |
| Day | 163.7193548 | 160.4413965 | 166.9973132 | 1.602733624 | 102.1500719 | 1.2216E-38 |

This output from time series analysis gives us the following regression equation:

$$Predicted\ User\ Count\ = -\ 28.34\ +\ 163.72(day)$$

D-90 Predicted User Count

$$-\ 28.34\ +\ 163.72(90)\ =\ 14706.46$$

D-180 Predicted User Count

$$-\ 28.34\ +\ 163.72(180)\ =\ 29441.26$$

On day 90, total user count from network 1 is predicted to be approximately 14706 users, and on day 180, total user count is predicted to be approximately 29441 users.

**Network 2**

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -75.37419355 | -179.6517702 | 28.90338315 | 50.98575463 | -1.47833829 | 0.15009747 |
| DAY | 291.7391129 | 286.0503047 | 297.4279211 | 2.781500952 | 104.8854981 | 5.6882E-39 |

This output from time series analysis gives us the following regression equation:

$$Predicted\ User\ Count\ = -\ 75.37\ +\ 291.74(day)$$

D-90 Predicted User Count

$$-\ 75.37\ +\ 291.74(90)\ =\ 26181.23$$

D-180 Predicted User Count

$$-\ 75.37\ +\ 291.74(180)\ =\ 52437.83$$

On day 90, total user count from network 2 is predicted to be approximately 26181 users, and on day 180, total user count is predicted to be approximately 52437 users.

**Network 3**

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -11.43870968 | -24.35182682 | 1.47440747 | 6.313773711 | -1.81170726 | 0.08040003 |
| DAY | 43.31169355 | 42.60722525 | 44.01616184 | 0.344444595 | 125.7435716 | 2.9889E-41 |

This output from time series analysis gives us the following regression equation:

$$Predicted\ User\ Count\ = -\ 11.44\ +\ 43.31(day)$$

D-90 Predicted User Count

$$-\ 11.44\ +\ 43.31(90)\ =\ 3886.46$$

D-180 Predicted User Count

$$-\ 11.44\ +\ 43.31(180)\ =\ 7784.36$$

On day 90, total user count from network 3 is predicted to be approximately 3886 users, and on day 180, total user count is predicted to be approximately 7784 users.

## Network 4

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -7.393548387 | -12.41055578 | -2.37654099 | 2.453028889 | -3.01404864 | 0.00530959 |
| DAY | 8.786693548 | 8.512993356 | 9.060393741 | 0.133823697 | 65.6587266 | 4.2602E-33 |

This output from time series analysis gives us the following regression equation:

$$Predicted\ User\ Count\ = -\ 7.39\ +\ 8.79(day)$$

D-90 Predicted User Count

$$-\ 7.39\ +\ 8.79(90)\ =\ 783.71$$

D-180 Predicted User Count

$$-\ 7.39\ +\ 8.79(180)\ =\ 1574.81$$

On day 90, total user count from network 4 is predicted to be approximately 783 users, and on day 180, total user count is predicted to be approximately 1574 users.

## Organic

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -172.3935484 | -340.5203463 | -4.266750449 | 82.20436203 | -2.09713383 | 0.04481391 |
| DAY | 380.9600806 | 371.7880119 | 390.1321494 | 4.484615613 | 84.94821263 | 2.5226E-36 |

This output from time series analysis gives us the following regression equation:

$$Predicted\ User\ Count\ = -\ 172.39\ +\ 380.96(day)$$

D-90 Predicted User Count

$$-\ 172.39\ +\ 380.96(90)\ =\ 34114.01$$

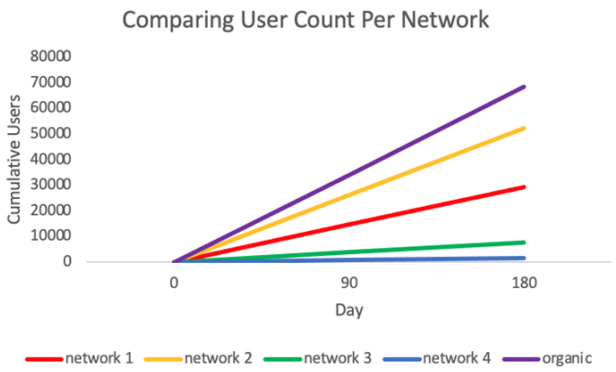
D-180 Predicted User Count

$$-\ 172.39\ +\ 380.96(180)\ =\ 68400.41$$

On day 90, total user count from organic users is predicted to be 34,114 users, and by day 180, total user count is predicted to be approximately 68,400 users.

**Observations:**

The graph below called “Comparing User Count Per Network”. In this graph, we can see that Networks 2 and Organic Network all have significant increases in user count from day 90 to day 180. Network 4 is the only network that barely increases in user count from day 90 to day 180. Network 3 has a slight increase in user count but Organic Network continues to take the lead.




**User Count - Region**

To calculate the user count per region, we first separated out user id, event_dt, and region from the initial data. Once the data was separated, we highlighted the user id column and went to excel data tab and then went to table tools and selected “remove duplicates”. Once all the duplicates were removed, we set up two new columns of day 1-31 and users. Under the user column, we used the COUNTIF function where we highlighted the event_dt column and the criteria was “1” for day 1 and continued to day 31. Once the users were counted for each day, we did a time series partition for the data and then ran linear regression to end up with the predicted user count equation.

**Region 1**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -269.2258065 | -496.6497539 | -41.80185902 | 111.1972674 | -2.4211549 | 0.02195806 |
| day | 578.171371 | 565.7643775 | 590.5783645 | 6.066308273 | 95.3086037 | 9.068E-38 |

This output from the time series analysis gives us the following regression equation:

*Predicted User Count* = -269.23 + 578.17 (*day)*

D-90 Predicted User Count

-269.23 + 578.17 (90) = 51,766 users

D-180 Predicted User Count

-269.23 + 578.17 (180) = 103,801 users

The total user count for Region 1 is predicted by the model to have 51,776 cumulative users on day 90 and 103,801 cumulative users on day 180..

**Region 2**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -123.9677419 | -230.6632252 | -17.27225865 | 52.16797228 | -2.376319 | 0.02430698 |
| day | 222.6975806 | 216.8768648 | 228.5182965 | 2.845996212 | 78.2494297 | 2.7041E-35 |

This output from the time series analysis gives us the following regression equation:

*Predicted User Count* = -123.97 + 222.70 (*day)*

D-90 Predicted User Count

-123.97+ 222.70 (90) = 19,919 users

D-180 Predicted User Count

-123.97 + 222.70 (180) = 39,962 users

The total user count for Region 2 is predicted by the model to have 19,919 cumulative users on day 90 and 39,962 cumulative users on day 180.

**Region 3**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -20.61935484 | -64.09840365 | 22.85969397 | 21.25876132 | -0.9699227 | 0.34010555 |
| day | 87.31491935 | 84.94294278 | 89.68689593 | 1.159760511 | 75.2870256 | 8.2373E-35 |

This output from the time series analysis gives us the following regression equation:

*Predicted User Count* = -20.62 + 87.31 (*day)*

D-90 Predicted User Count

-20.62 + 87.31 (90) = 7,837 users

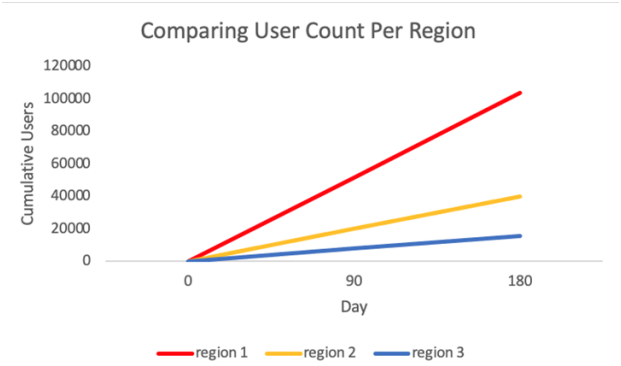
D-180 Predicted User Count

-20.62 + 87.31 (180) = 15,695 users

The total user count for Region 3 is predicted by the model to have 7,837 cumulative users on day 90 and 15,695 cumulative users on day 180.

**Observations:**

In the graph below called “Comparing User Count Per Region”, we can see that region 1 is doing significantly better than regions 2 and 3. Region 3 barely increased from day 90 to day 190 and Region 2 did slightly better.




**Revenue Overall**

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | 49.17083871 | -113.1622242 | 211.5039016 | 79.37155788 | 0.619501998 | 0.5404228 |
| day | 28.58101613 | 19.7250215 | 37.43701076 | 4.330073476 | 6.600584559 | 3.1108E-07 |

*Predicted Overall Revenue:* 49.17 + 28.58 *(Day)*

D-90 Predicted Revenue:

49.17 + 28.58 (90) = $2,621.37

D-180 Predicted Revenue:

49.17 + 28.58 (180) = $5,193.57

**Revenue Overall Cumulative**

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -3068.88671 | -3988.999975 | -2148.773445 | 449.8826177 | -6.82152764 | 1.7203E-07 |
| day | 555.3674355 | 505.1711416 | 605.5637294 | 24.54310892 | 22.62824312 | 5.6383E-20 |

*Predicted Overall Revenue* = -3,068.89 + 555.37 (*Day*) = $46,914.41

D-90 Predicted Revenue Cumulative

-3,068.89 + 555.37 (90) = $46,914.41

D-180 Predicted Revenue Cumulative

-3,068.89 + 555.37 (180) = $96,897.71

### User Count Overall

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | 668.5290323 | 548.4018339 | 788.6562306 | 58.73531064 | 11.38206345 | 3.2366E-12 |
| Day | 14.61008065 | 8.056604699 | 21.16355659 | 3.204273892 | 4.559560492 | 8.6117E-05 |

*Predicted Overall User Count* =688.53 + 14.61 (*Day*)

D-90 Predicted User Count
688.53 + 14.61 (90) = 2,003

D-180 Predicted User Count
688.53 + 14.61 (180) = 3,318

### User Count Overall Cumulative

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -907.5677419 | -1407.992211 | -407.1432728 | 244.6788658 | -3.70922 | 0.00087566 |
| Day | 885.9629032 | 858.6625103 | 913.2632962 | 13.34832647 | 66.37258276 | 3.1197E-33 |

*Predicted User Count Cumulative* =-907.57 + 885.97 (*Day*)

D-90 Predicted User Count Cumulative
-907.57 + 885.97 (90) = 78,829

D-180 Predicted User Count Cumulative
-907.57 + 885.97 (180) = 158,567

### Churn Classification

In this section, the business question we will address is: “Is there a trend of high churn and retention rate to help support whether we want to launch this game in the future?”. If the churn rate is higher than desired, we should work on improving user engagement and improving areas of our game that are contributing to user churn.

To calculate the logistic regression model to classify users as churned or not, we first had to decide whether a rule for deciding churn. The way we began this process was by establishing a rule which says if a user came back to the game more than 4 times from when they downloaded the game, then that user did not churn. Additionally, in order to create a model that factored each user individually, we derived a formula to calculate the total number of sessions a user logged on, which was summed across every time that the user logged onto the game. From there, we filtered out duplicate users from the data set in order to get only one record per user. Using Excel’s Data mining feature, we created dummy variables for the categorical variables, which were Network and Region. To create the model, we used Excel’s data mining platform to create the logistic model. We did not include the Region 3 or the Organic Network Data in the development of our model, to account for the n-1 rule with creating dummy variables.

Refer to Appendix C for an image of the coefficients for each predictor calculated on excel. In terms of odds, the logistic regression equation to predict whether a user will churn or not is:

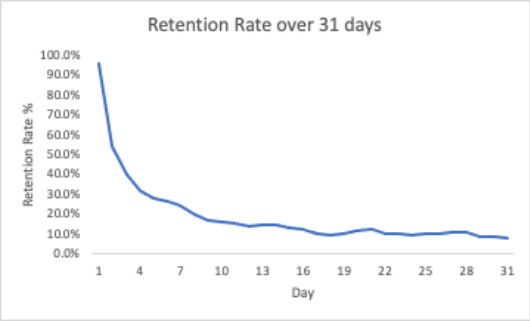
$$Odds(Churn = 1) = e^{5.07-.26(Number\ of\ Sessions)+.36(Region1)+.51(Region2)+.17(Network1)-.23(Network2)-.38(Network3)+.09(Network}$$

After these calculations were made, it is important to note that the p-values for Region 1, Region 2, Network 1, Network 2, Network 3, and Network 4 are all highly insignificant. These values indicate that these 6 predictors do not have a predictive relationship with deciding if a user will churn or not. However, the number of sessions that a user engaged in since they downloaded the game was highly significant, indicating good classifying power based on that metric. The odds associated with the number of sessions indicate that the chance that a user churns decreases by approximately 23.2% with each additional session played by a user.

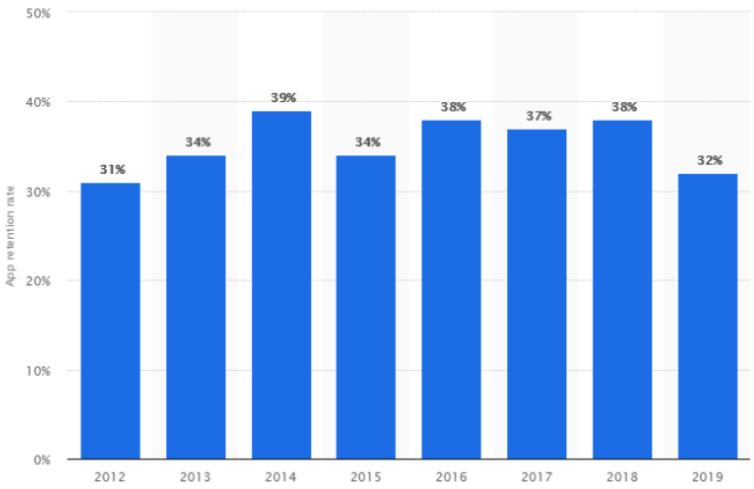
From the training data to the validation data, there was a .12 % increase in the error for predicting records. Since the increase in misclassification is so low, overfitting is not a problem for this data set. Refer to Appendix D for the percentage of error results of the training and validation data on Excel and the ROC Curve and Lift Chart.

Additionally, the ROC chart from the validation data shows that the model has outperformed the random classifier, as the blue curve lies above the red curve and the area under the curve is high. For the top 100 cases as ranked by the model for being highly probable to churn, the top 100 cases will yield 1.06958321 as many 1's as opposed to 100 cases chosen randomly.

We then wanted to calculate the overall retention rate, and the way this was calculated began by calculating the total number of users that engaged on a single day. This was calculated by using the formula, COUNTIF() with a criteria equal to the day. Then to show the trend of user retention, a running total was calculated, by summing entries consecutively for all 31 days. The retention rate was then calculated by dividing (# Number of users that played on day n) / (# total number of users that had downloaded the game by that day). Refer to Appendix E for the Excel results that were able to create the retention rate over 31 days graph.




The results showed a decrease in the number of people that played the game each day, indicating user engagement and interest decrease overtime. Additionally, the average retention rate across the 31 days was 19%, indicating the true loss of interest in this game after initial download. From the retention rate, the churn for this game is 81% in the first 31 days. Both churn and retention rates indicate a continual loss of interest by users in the game. According to a study produced by Statista, the average retention rates for mobile games worldwide across a 8 year time period did not drop below 30 percent. Their rule for determining churn was different than ours, saying a user churned if they visited less than 11 times in their yearly time period.




This study helps to put the results of our churn calculations in perspective, by providing a benchmark for where we would want our game to be once released to a worldwide market. With our average retention rate sitting at 19%, we can take a step back and start to analyze which parts of our game are contributing to such low user engagement, and start working on ways to improve these areas.

As Upland Localytics put it in a 2019 report, “retention and churn are some of the best ways to diagnose problems and successes in your mobile platforms.” They conducted a study on monthly retention rates for 37,000 mobile games over a span of 3 months, and by the end of month one, the average retention rate across all games was 42.29%, and by month three, that average was 27.24 %. While their rule for churn was not clearly stated, we can see our game is far behind the norm in terms of retention rates and churn.

To calculate the retention rate per region, we used a COUNTIF function to determine the unique users who installed the game on days 1-31. Then we calculated the cumulative users over the course of day 1-31. Next, we used a COUNTIF function to determine the total number of players who engaged with the app on each day during day 1-31. Once this was calculated, we divided the number of players who engaged on day n by the total number of users who had installed the game by day n.

Region 1:

| **region 1** | | | | | |
|---|---|---|---|---|---|
| day | users who logged in on day n | users who installed on day n | cumulative users | retention rate | churn rate |
| 1 | 592 | 618 | 618 | 95.79% | 4.21% |
| 2 | 635 | 547 | 1165 | 54.51% | 45.49% |
| 3 | 664 | 512 | 1677 | 39.59% | 60.41% |
| 4 | 681 | 505 | 2182 | 31.21% | 68.79% |
| 5 | 770 | 555 | 2737 | 28.13% | 71.87% |

For region 1, the data shows a high retention rate for day 1, however it quickly falls to below 50% in subsequent days. Region 1 has an average retention rate of 19.09% and an average churn rate of 80.91% which indicates users in this region did not have consistent interest in the game over the 31 day period.

Region 2:

| region 2 | | | | | |
|---|---|---|---|---|---|
| day | users who logged in on day n | users who installed on day n | cumulative users | retention rate | churn rate |
| 1 | 230 | 240 | 240 | 95.83% | 4.17% |
| 2 | 232 | 201 | 441 | 52.61% | 47.39% |
| 3 | 361 | 209 | 650 | 55.54% | 44.46% |
| 4 | 274 | 207 | 857 | 31.97% | 68.03% |
| 5 | 288 | 217 | 1074 | 26.82% | 73.18% |

For region 2, the data shows high retention for the first day and above 50% retention for the following two days, however retention quickly decreases after day 5. Region 2 has an average retention rate of 19.02% and an average churn rate of 80.98%, which indicates users in this region did not have consistent interest in the game over the 31 day period.

Region 3:

| region 3 | | | | | |
|---|---|---|---|---|---|
| day | users who logged in on day n | users who installed on day n | cumulative users | retention rate | churn rate |
| 1 | 99 | 102 | 102 | 97.06% | 2.94% |
| 2 | 108 | 94 | 196 | 55.10% | 44.90% |
| 3 | 112 | 78 | 274 | 40.88% | 59.12% |
| 4 | 126 | 83 | 357 | 35.29% | 64.71% |
| 5 | 132 | 91 | 448 | 29.46% | 70.54% |

For region 3, the data shows high retention for the first day, but quickly falls below 50% after the third day. Region 3 has an average retention rate of 19.31% and an average churn rate of 80.69%. Region 3 has a slightly greater average retention rate compared to the other two regions, but it still indicates that users did not have consistent interest in the game over the 31 day period.

**Retention Rate Per Network:**

To calculate the retention rate per network, we used pivot charts to determine the unique users who installed the game and the unique users who logged onto the game on days 1-31. We then calculated the cumulative users over the course of that month. Next, we calculated retention rate by dividing users who logged onto the game that day by the cumulative user count that same day. We calculated churn rate by minusing the users who logged on that day from the cumulative user count that day and dividing by the same cumulative user count number. To double check that these numbers were accurate, we added the churn rate and retention rate together to ensure the equalled 100%.

Network 1

| Event day | Logged On | Installed | Cumulative users | Retention Rate | Churn rate | Double check |
|---|---|---|---|---|---|---|
| 1 | 161 | 279 | 279 | 57.71% | 42.29% | 100.00% |
| 2 | 193 | 444 | 723 | 26.69% | 73.31% | 100.00% |
| 3 | 196 | 386 | 1109 | 17.67% | 82.33% | 100.00% |
| 4 | 204 | 295 | 1404 | 14.53% | 85.47% | 100.00% |
| 5 | 231 | 402 | 1806 | 12.79% | 87.21% | 100.00% |

For Network 1, the retention rate starts out higher than the churn rate, with a 58% to 42% ratio. The churn rate soon becomes higher than the retention rate on the second day and stays that way until the end of the month. The average retention rate is 9.15% and the average churn rate is 90.85%. These averages show us that the users were not as interested in the game as we'd hope.

Network 2

| Event day | Logged On | Installed | Cumulative | Retention rate | Churn Rate | Double Check |
|---|---|---|---|---|---|---|
| 1 | 320 | 971 | 971 | 32.96% | 67.04% | 100.00% |
| 2 | 311 | 572 | 1543 | 20.16% | 79.84% | 100.00% |
| 3 | 337 | 523 | 2066 | 16.31% | 83.69% | 100.00% |
| 4 | 353 | 656 | 2722 | 12.97% | 87.03% | 100.00% |
| 5 | 399 | 738 | 3460 | 11.53% | 88.47% | 100.00% |

For Network 2, the churn rate is higher than the retention rate for the entire 31 days. The churn rate percentages are pretty high for this network. The average churn rate is 92.15% while the average retention rate is 7.85%. These rates show us that customers are more likely to churn that come back.

Network 3

| Event Day | Logged on | Installed | Cumulative | Retention Rate | Churn Rate | Double Check |
|---|---|---|---|---|---|---|
| 1 | 51 | 131 | 131 | 38.93% | 61.07% | 100.00% |
| 2 | 38 | 63 | 194 | 19.59% | 80.41% | 100.00% |
| 3 | 52 | 123 | 317 | 16.40% | 83.60% | 100.00% |
| 4 | 54 | 73 | 390 | 13.85% | 86.15% | 100.00% |
| 5 | 64 | 117 | 507 | 12.62% | 87.38% | 100.00% |

For Network 3, the churn rate is consistently higher than the retention rate for the entire 31 days. They are around the same percentages as in Network two. The average retention rate is 8.23% while the average churn rate is 91.77%. It seems that users are more likely to churn in this region as well.

Network 4

| Event Day | Logged on | Installed | Cumulative Users | Retention Rate | Churn Rate | Double Check |
|---|---|---|---|---|---|---|
| 1 | 7 | 19 | 19 | 36.84% | 63.16% | 100.00% |
| 2 | 10 | 65 | 84 | 11.90% | 88.10% | 100.00% |
| 3 | 14 | 14 | 98 | 14.29% | 85.71% | 100.00% |
| 4 | 11 | 10 | 108 | 10.19% | 89.81% | 100.00% |
| 5 | 8 | 4 | 112 | 7.14% | 92.86% | 100.00% |

For Network 4, the churn rate starts at 63.16% (higher than the retention rate) and steadily increases from there. The average retention rate is 7.45% while the average churn rate is 92.55%. This is so far the highest churn rate but pretty close in comparison to the other networks. These values indicate that the users are not as interested in the game and wanting to return back to playing as we'd hope.

Organic

| Event Day | Logged on | Installed | Cumulative | Retention Rate | Churn Rate | Double Check |
|---|---|---|---|---|---|---|
| 1 | 382 | 929 | 929 | 41.12% | 58.88% | 100.00% |
| 2 | 428 | 816 | 1745 | 24.53% | 75.47% | 100.00% |
| 3 | 438 | 947 | 2692 | 16.27% | 83.73% | 100.00% |
| 4 | 464 | 715 | 3407 | 13.62% | 86.38% | 100.00% |
| 5 | 490 | 745 | 4152 | 11.80% | 88.20% | 100.00% |

For the Organic Network, the churn rate is consistently higher than the retention rate but starts out relatively close with a 59 to 41 percent ratio. Average retention rate is 8.49% while the average churn rate is 91.51%. These values show us that users in this network are more likely to churn than they are to return back to the game.

Overall, the average churn rate per network is higher than the average retention rate by a significant amount. This means that it is more common to see a user download the game and delete it (or become inactive) than it is to see them download the game and revisit for more play time. These are negative implications of the success of our puzzle game.

**Overall Performance**

Regarding regions, region 3 has the highest average retention rate while region 2 has the highest average churn rate. On average, the average retention rate lies around 19% which is lower than the ideal average at around 35% for startups, meaning people are not returning as much. The churn rates for all regions are mostly higher than average which suggests people are deleting the game more quicker after downloading and with little chance of them returning to the platform. For networks, Network 1 has the highest retention rate and Network 4 has the highest churn rate. Everything considered, it seems to indicate that the launch of the game may not have favorable results with low retention rates and high churn rates.

**Comparing Methods**

The purpose of this section is to take a step back from in depth analysis and examine the different ways we have tried to realize the game's value to determine which method will give us the most complete idea of if this game will be successful or not. By evaluating the models created through multiple linear regression to calculate expected LTV and the logistic regression model used to predict if a user will churn or not, we can take those results and propose better ways to improve our analysis in the places that it has lacked.

**Categorical Variables**

To predict the categorical variable about whether or not a user will churn, we first developed a Logistic Regression Model. This model did not show evidence of overfitting, as from the training data to the validation data, there was only a .12 % increase in the error for predicting records. Additionally, the ROC chart and lift chart showed that this model outperformed the random classifier in both instances with the validation data. It is important to note that the only significant predictor from the developed logistic regression model was the number of sessions a user engaged in since downloading the game. Region and Network of origination did not give evidence to suggest whether or not a user would churn.

In terms of odds, the logistic regression equation to predict whether a user will churn or not is:

$$Odds(Churn = 1) = e^{5.07-.26(Number\ of\ Sessions)+.36(Region1)+.51(Region2)+.17(Network1)-.23(Network2)-.38(Network3)+.09(Network}$$

The odds associated with the number of sessions indicate that the chance that a user churns decreases by approximately 23.2% with each additional session played by a user.

**Classification Tree**

To test the accuracy of our Logistic Regression model we also developed a Classification Tree model to predict whether or not a user will churn. This model was partitioned into training, validation, and test data. All three partition splits showed no evidence of overfitting. The error rate for training data was 1.899%, validation error was 2.135%, and test error was 1.808%. We can see a slight increase from training to validation data but a decrease from validation to test data. This shows that our model is doing very well. Additionally, the ROC chart and lift chart showed that this model outperformed the random predictor in both instances with the test data. It is important to note that cumulative sessions per user carries more importance than network and region when calculating whether or not a user will churn. This was shown in the Output Sheet from our model analysis. Refer to *Appendix F* for a picture of the Best Pruned Tree, Lift Chart, Decile-wise Lift Chart, and RROC Curve of the test data.

**Which Method is Best?**

According to our model results, we can conclude that both the logistic regression model and the classification tree model are equally successful in calculating our categorical target variable of whether or not a user will churn. They are both relatively similar in regards to reliability and accuracy considering the low error rate for each model and the fact that they both showed no signs of overfitting. The classification tree is not as effective in calculating the target categorical variable when it comes to region and network but since these variables were deemed highly insignificant in the logistic regression model, we can conclude that both models were able to accomplish the same goal. All in all, we accumulated the same results using both the logistic regression and classification tree model.

**<u>Continuous Target Variables</u>**

**Multiple Regression**

To predict lifetime value for the game in beta, we created multiple linear regression models separated by network, region, and overall data. These models gave us data on what the predicted revenue on day 90 and day 180 would be if this game is released, which can be broken down by network and region. Evaluating these models included looking at the p-values associated with predictors in the models and the regression statistics.

Regarding overall cumulative revenue, the p-value for “days” is less than 0.05, which means we have evidence to reject the null hypothesis. For overall cumulative user count, the p-value for “days” is less than .05 which means we can reject the null hypothesis. Both cumulative revenue and user count lift charts display how the model outperformed the random predictor as the lift ratios are above 1.0. However, the RROC charts are a little concerning as the line is not close to the top left corner as desired. Overall, it is relatively a good model. Please refer to Appendix G for charts.

**Regression Tree**

We conducted regression trees to see how much revenue or user count will be on Day 90 and Day 180. Looking at cumulative revenue first, the metrics from training to validation data increased drastically, indicating an overfitting model. For example, MAD and RMSE were 579.16 and 686.71 respectively for training data. MAD and RMSE for validation data are 5,824.8 and 6511.35 respectively. So if we want to identify what revenue would be for day 90 and day 180 using a regression tree, it would be $5,657.25 for both days. Similar results shown for cumulative user count as the metrics dramatically increased. The ROC charts for both revenue and user count show the fitted predictor is not close to the top left corner, indicating concerns for the accuracy of the model. Also, the lift ratios for both revenue and user count were under 1 which mean we are likely to identify the important class less. If we want to find out what the user count would be for day 90 and day 180, there would be 14,145 users for both days. To sum up,

there are concerns with overfitting in this model. Please refer to Appendix H for regression trees, metrics and charts.

**Double Exponential Smoothing**

Another method we used to predict lifetime value for the game in beta was creating a double exponential smoothing model for cumulative revenue and user count data. Due to limitations in excel, we could only use this model to predict day 90 cumulative revenue and user count. Evaluating this model included looking at average error, MAD and MAPE.

The MAPE for our cumulative revenue model was 6.4%, the average error was -76.2, and the MAD was 343.7. Since the full data set was used to create the predictions, we only have training metrics to reference. When we partitioned the data to see how the model would work for our full data set, the training data had a MAPE of 9.65% and the validation data had a MAPE of 11.34%.The double exponential smoothing model gave us a forecast of $23,468.59 of cumulative revenue at day 90. Overall the forecasted cumulative revenue trend using the double exponential smoothing model was increasing, which is the trend we wanted to see.

The MAPE for our cumulative user count model was 2.74%, the average error was -98.16, and the MAD was 278.10. When we partitioned the data to see how the model would work on the full data set, the training data had a MAPE of 3.59% and the validation data had a MAPE of 6.94%. The double exponential smoothing model gave us a forecast of 92,209 cumulative users by day 90. Overall the forecasted cumulative user count trend using the double exponential smoothing model was increasing, which is the trend we want to see.

**Which Method is Best?**

According to the models discussed, it is safe to conclude that the multiple regression and double exponential smoothing model are optimal for calculating D90 and D180 LTV for numerical target variables due to low concerns of overfitting and low errors. However, the double exponential smoothing model has limitations and could not make predictions for the desired metrics at D180. The regression trees did not give us a good picture of an ideal prediction because of the large generalization that the tree makes, saying that any day after day 17, the game is predicted to generate the same amount of revenue and the same amount of users. This lack of ability to predict growth in the game makes the regression tree the worst predictive model for our game. All components considered, the multiple linear regression model is the best predictive model out of the three that we applied to our data.

**D90 & D180 Revenue and User Count for our Competitors**

In order to more accurately compare the Day 90 and Day 180 Revenue and User Count results from our game in beta to our competitor's numbers, we pulled the data that would be the equivalent of 3 regions from each of our competitors in order to give a more fair comparison to our game in beta that has only been launched in 3 regions. To do this, we researched the countries listed in our competitors' data, researched different regions worldwide, and used the 3 regions of Scandinavia, British Isles, and Central America. This required us to filter out the countries that don't fall under these 3 regions in our pivot tables. Refer to Appendix I to see which countries from the data are applied to which regions.

**<u>Simon's Cat</u>**
**Revenue**
**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -25525.32609 | -28994.33724 | -22056.31494 | 1746.139632 | -14.618147 | 1.7E-25 |
| Day | 1741.861851 | 1677.079906 | 1806.643796 | 32.60823231 | 53.4178558 | 5.92E-70 |

Predicted Revenue = -25,525.33+1,741.86(Day)

D90 Actual Revenue= $139,727
D90 Predicted Revenue= -25,525.33+1,741.86(90)= $131,242.07
D180 Predicted Revenue= -25,525.33+1,741.86(180)= $288,009.47

**User Count**
**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | 30160.785 | 21047.23748 | 39274.33251 | 4587.337951 | 6.57479029 | 3.11E-09 |
| Day | 2319.306166 | 2149.115496 | 2489.496837 | 85.66610529 | 27.0737903 | 4.97E-45 |

Predicted User Count= 30,160.79+2,319.31(Day)
D90 Actual User Count= 214,978
D90 Predicted User Count= 30,160.79+2,319.31(90)= 238,898.69
D180 Predicted User Count= 30,160.79+2,319.31(180)= 447,636.59

**<u>Diamond Diaries</u>**
**Revenue**
**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -27647.92021 | -30488.7552 | -24807.08522 | 1429.944831 | -19.334956 | 8.39E-34 |
| Day | 4093.392861 | 4040.341773 | 4146.443949 | 26.70346196 | 153.290718 | 1.2E-110 |

Predicted Revenue= -27,647.92+4,093.39(Day)
D90 Actual Revenue= $344,343
D90 Predicted Revenue= -27,647.92+4,093(90)= $340,722.08
D180 Predicted Revenue= -27,647.92+4,093(180)= $709,092.08

**User Count**
**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | 210582.9128 | 183094.3201 | 238071.5055 | 13836.48512 | 15.2193936 | 1.27E-26 |
| Day | 4202.70641 | 3689.371508 | 4716.041312 | 258.3890274 | 16.2650344 | 1.55E-28 |

Predicted User Count= 210,582.91+4,202.71(Day)
D90 Actual User Count= 530,058
D90 Predicted User Count= 210,582.91+4,202.71(90)= 588,826.81
D180 Predicted User Count= 210,582.91+4,202.71(180)= 967,070.71

**<u>Hello Neighbor</u>**
**Revenue**
**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | 14829.25812 | 11803.65754 | 17854.85871 | 1515.403439 | 9.7856833 | 1.79E-14 |
| Day | 1366.894053 | 1290.668182 | 1443.119923 | 38.17851787 | 35.8027008 | 5.86E-45 |

Predicted Revenue= 14,829.26+1366.89(Day)
D90 Predicted Revenue= 14,829.26+1366.89(90)= $137,849.36 (Hello neighbor does not have data as far as the first 90 days hence a predicted value)
D180 Predicted Revenue= 14,829.26+1366.89(180)= $260,869.46

**User Count**
**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | 331992.1866 | 291269.1946 | 372715.1785 | 20396.53295 | 16.2768931 | 8.52E-25 |
| Day | 7801.532103 | 6775.571994 | 8827.492212 | 513.8627629 | 15.1821316 | 3.25E-23 |

Predicted User Count= 331,992.19+7,801.53(Day)
D90 Predicted User Count= 331,992.19+7,801.53(90)= 1,034,129.89 (Hello neighbor does not have data as far as the first 90 days hence a predicted value)
D180 Predicted User Count= 331,992.19+7,801.53(180)= 1,736,267.59

**Brick n Balls**
**Revenue**
**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -69142.69661 | -80754.6358 | -57530.75741 | 5844.912672 | -11.829552 | 4.98E-20 |
| Day | 4297.878235 | 4081.031424 | 4514.725046 | 109.1506468 | 39.3756552 | 1.6E-58 |

Predicted Revenue= -69,142.7+4,297.88(Day)
D90 Actual Revenue= $358,436
D90 Predicted Revenue= -69,142.7+4,297.88(90)= $317,666.5
D180 Predicted Revenue= -69,142.7+4,297.88(180)= $704,475.7

**User Count**
**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -147725.4066 | -168164.9264 | -127285.8868 | 10288.3081 | -14.358571 | 5.27E-25 |
| Day | 14014.39748 | 13632.70028 | 14396.09468 | 192.1287017 | 72.9427584 | 7.45E-82 |

Predicted User Count= -147,725.41+14014.4(Day)
D90 Actual User Count= 1,184,665
D90 Predicted User Count= -147,725.41+14014.4(90)= 1,113,570.59
D180 Predicted User Count= -147,725.41+14014.4(180)= 2,374,866.59

**Game in Beta**
**Revenue**
**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -3068.88671 | -3988.999975 | -2148.773445 | 449.8826177 | -6.82152764 | 1.7203E-07 |
| day | 555.3674355 | 505.1711416 | 605.5637294 | 24.54310892 | 22.62824312 | 5.6383E-20 |

Predicted Revenue= -3,068.89 + 555.37 (*Day*)
D90 Predicted Revenue= -3,068.89 + 555.37 (90) = $46,914.41
D180 Predicted Revenue= -3,068.89 + 555.37 (180) = $96,897.71

**User Count**
**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -907.5677419 | -1407.992211 | -407.1432728 | 244.6788658 | -3.70922 | 0.00087566 |
| Day | 885.9629032 | 858.6625103 | 913.2632962 | 13.34832647 | 66.37258276 | 3.1197E-33 |

Predicted User Count= -907.57 + 885.97 (*Day*)
D90 Predicted User Count=-907.57 + 885.97 (90) = 78,829
D180 Predicted User Count= -907.57 + 885.97 (180) = 158,567

**Summary:**

| 90 Day Stats | Predicted Total User Count | Predicted Total Revenue | RPU | % of Market Revenue | % of Users |
|---|---|---|---|---|---|
| Brick n Balls | 1,184,665 | $358,436.00 | $ 0.30 | 34.89% | 38.94% |
| Diamond Diaries | 530,058 | $344,343.00 | $ 0.65 | 33.52% | 17.42% |
| Simons cat | 214,978 | $139,727.00 | $ 0.65 | 13.60% | 7.07% |
| Hello Neighbor | 1,034,129 | $137,849.36 | $ 0.13 | 13.42% | 33.99% |
| Gamein Beta(Mult.Regression) | 78,829 | $46,914.41 | $ 0.60 | 4.57% | 2.59% |

| 180 Day Stats | Predicted Total User Count | Predicted Total Revenue | RPU | % of Market Revenue | % of Users |
|---|---|---|---|---|---|
| Diamond Diaries | 967,071 | $709,092.08 | $0.73 | 34.43% | 17.01% |
| Brick n Balls | 2,374,866 | $704,475.70 | $0.30 | 34.21% | 41.78% |
| Simons Cat | 447,636 | $288,009.47 | $0.64 | 13.99% | 7.87% |
| Hello Neighbor | 1,736,267 | $260,869.36 | $0.15 | 12.67% | 30.54% |
| Gamein Beta(Mult.Regression) | 158,567 | $96,897.71 | $0.61 | 4.71% | 2.79% |

Based on this table, the projected value for day 90 and day 180 revenue and user count is significantly lower for the game in beta than for the four successful games. To calculate the percent of market revenue and the percent of users in the market, we summed the predicted values for both metrics, but only included predicted revenue and user count generated from the multiple linear regression model in the total sum. To make comparison more accurate, we used the scaled data of the three specified regions to create these linear regression models. Using the predicted values from the multiple linear regression model, the total revenue between all five games showed that by day 90, the game in beta is projected to have generated only 4.57% of the market value in comparison to the other games. Similarly, by day 180 the game in beta is projected to only have generated 4.71% of the total revenue produced by these five games. This trend continues with user count, where by D90 the game in beta is predicted to have 2.59% of the total users, and by D180, this value slightly increases where the number of users for the game in beta only make up 2.79% of the total users. These tables are sorted by Predicted Total Revenue, which demonstrates that the game in beta is predicted to have the lowest user count and lowest revenue in comparison to the other games by D90 and D180. Although the predicted RPU is higher for the game in beta than Hello Neighbor and Brick n Balls on these days, the sheer difference in the total revenue and total user count demonstrates the lack of power this game has in the market. Overall, the game in beta is significantly underperforming compared to the successful games.

Looking at the different models generated and the comparison between the four successful games to predict user count and revenue, we predict that we will not be successful by D90 and D180. The difference in value for both metrics suggests that we will not ever be as successful as the other games in the future. The expected cumulative revenue on day 90 for our game in beta is $46,914.41 and the advertising cost is $228,932.67. We calculated this predicted advertising cost by running a logistic regression model and using the given formula: Predicted Cost=631.47+2,536.68(90). The expected cumulative revenue on day 180 for our game is $96,897.71 but the predicted advertising cost is $457,233.87. This comes to a -78.81% Return on Investment. From our calculations, our revenue is unable to cover even just our advertising costs for the first 90 days and we get even farther from breaking even on day 180. These calculations show that our business will suffer more in the long run when it comes to profitability if we launch this game.

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | 631.4660645 | 77.67946492 | 1185.252664 | 270.7698872 | 2.33211333 | 0.026845 |
| Days | 2536.675266 | 2506.46373 | 2566.886802 | 14.77170838 | 171.725247 | 3.59E-45 |

Based on these predictions, we should focus our investments on research and marketing. Investing in research allows us to make the accurate predictions we need in deciding whether or not a game should be launched and can help improve and better the state of our company. Without investing in research, we would not be able to gather, store, access, and analyze necessary consumer data in order to help our business make the most worthwhile decisions. Marketing is another area we should focus our investments because many regions and networks are lacking in revenue and user count compared to others. With a strong marketing strategy, we would be able to gather more meaningful data that would essentially assist us in making these big business decisions.

**Conclusion:**

At the beginning of analysis, one component that we wished we had was daily expenses of running each game. This data would give us the ability to calculate the net profit margin for each game, helping us to make bigger decisions about the success of the game, not just based on revenue and user count. Another piece of information that would have been beneficial when creating our analyses was knowing which regions our game was released in. The definition of “region” is very broad. There are regions within a state, regions within a country, and a region made up of multiple countries. This ambiguous term required us to make assumptions on regions in order to be as accurate as possible. If we knew the exact regions in which our game was launched, our results would be a lot more precise.

In conclusion, we would not recommend a worldwide launch of this game. Throughout the analysis, the game in beta showed user disengagement overtime, slow generation of revenue, and slow accumulation of users in comparison to the successful competitors. One of the first calculations we performed was the overall retention rate of the game, which showed that the game had close to an 80% churn rate, which is an undesirable trait when considering the launch of the game. This game spent $78,449.18 on advertising costs before launching the game across every region and network. By day 31, the game had generated a total of $15,700.48 in revenue. The return on investment for this game was -80.0 % by day 31, indicating that the game has not made any profit for the creators since launch. If the game were to stop investing money into advertising on all networks after the beta period, the game would potentially continue to see growth and profit from their games. However, if there is money put into advertising for each day following day 31, the calculations show that there is the possibility to not make a profit for a long period of time. Additionally, the calculations of predicted revenue and user count showed the sheer difference in impact that the game has compared to successful games in the marketplace. With all of the characteristics and calculations made through the course of this analysis, the game does not show signs of success in the competitive video game world, and we do not recommend that this game be launched.

***Appendix***

Appendix A

RPU Overall

| ST Game | TOTAL REVENUE | NUMBER OF DOWNLOADS | REVENUE per USER |
|---|---|---|---|
| SC | $ 2,649,610.00 | 3599790 | $ 0.74 |
| DD | $ 6,411,966.00 | 10657558 | $ 0.60 |
| BB | $ 8,879,440.00 | 28386572 | $ 0.31 |
| HN | $ 1,779,198.00 | 13574882 | $ 0.13 |

Appendix B

RPU of the game during beta overall

| | |
|---|---|
| Total Number of Users | 27856 |
| Total Revenue | $ 15,700.48 |
| Revenue per User | $ 0.56 |

| | |
|---|---|
| Region 1 RPU | |
| Total Number of Users | 18051 |
| Total Revenue | $ 11,758.29 |
| RPU | $ 0.65 |

| | |
|---|---|
| Region 2 RPU | |
| Total Number of Users | 6941 |
| Total Revenue | $ 2,172.93 |
| RPU | $ 0.31 |

| | |
|---|---|
| Region 3 RPU | |
| Total Number of Users | 2782 |
| Total Revenue | $ 1,658.29 |
| RPU | $ 0.60 |

Appendix C
Coefficients for each predictor

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Odds | Standard Error | Chi2-Statistic | P-Value |
|---|---|---|---|---|---|---|---|
| Intercept | 5.066001516 | 4.513964809 | 5.618038223 | 158.5391421 | 0.281656556 | 323.5123767 | 0.0000 |
| Cumulative sessions per user | -0.264017304 | -0.28250898 | -0.245525628 | 0.767960251 | 0.009434702 | 783.0841137 | 0.0000 |
| region_Region 1 | 0.360314876 | -0.175082785 | 0.895712537 | 1.433780807 | 0.273167091 | 1.739833262 | 0.1872 |
| region_Region 2 | 0.517224856 | -0.095070968 | 1.129520681 | 1.677366252 | 0.312401569 | 2.741147228 | 0.0978 |
| network_Network 1 | 0.168125523 | -0.33905652 | 0.675307565 | 1.183085105 | 0.258771103 | 0.422119817 | 0.5159 |
| network_Network 2 | -0.2305209 | -0.625923629 | 0.164881828 | 0.794119837 | 0.201739793 | 1.305682168 | 0.2532 |
| network_Network 3 | -0.375266261 | -1.110527123 | 0.359994601 | 0.687106305 | 0.375139986 | 1.000673327 | 0.3171 |
| network_Network 4 | 0.090232999 | -1.890448527 | 2.070914526 | 1.094429255 | 1.010570369 | 0.007972558 | 0.9289 |

Appendix D

**Training: Classification Summary**

**Confusion Matrix**

| Actual\Predicted | 0 | 1 |
|---|---|---|
| 0 | 475 | 139 |
| 1 | 50 | 9078 |

**Error Report**

| Class | # Cases | # Errors | % Error |
|---|---|---|---|
| 0 | 614 | 139 | 22.63843648 |
| 1 | 9128 | 50 | 0.547765118 |
| Overall | 9742 | 189 | 1.940053377 |

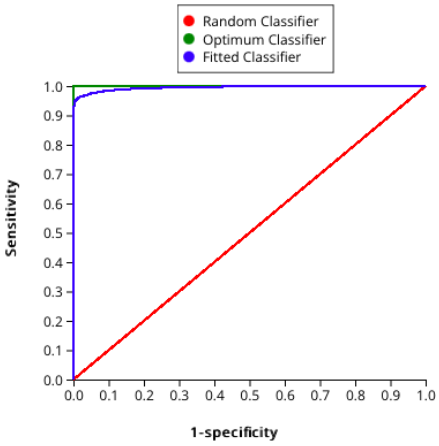
**Validation: Classification Summary**

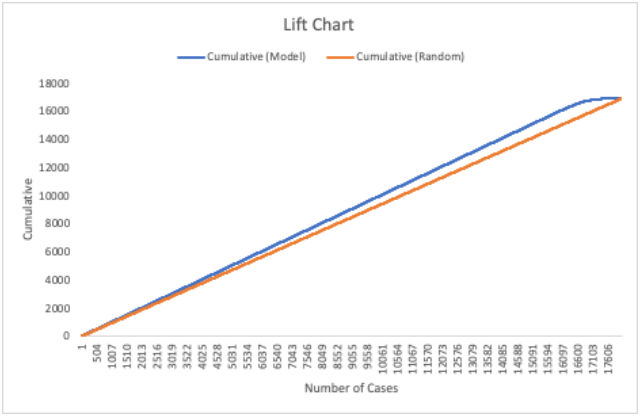
**Confusion Matrix**

| Actual\Predicted | 0 | 1 |
|---|---|---|
| 0 | 917 | 260 |
| 1 | 113 | 16802 |

**Error Report**

| Class | # Cases | # Errors | % Error |
|---|---|---|---|
| 0 | 1177 | 260 | 22.09005947 |
| 1 | 16915 | 113 | 0.668046113 |
| Overall | 18092 | 373 | 2.061684723 |



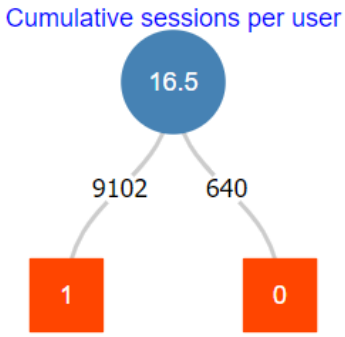




## Appendix E

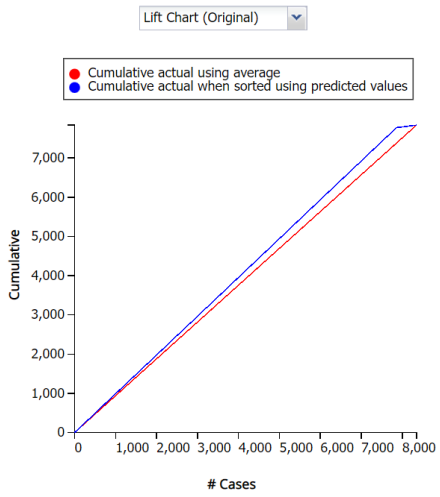
| Day | User Count | Cumulative | Users that used on Day n | Day n Retention Rate | CHURN |
|---|---|---|---|---|---|
| 1 | 961 | 961 | 922 | 95.9% | 4.1% |
| 2 | 846 | 1807 | 980 | 54.2% | 45.8% |
| 3 | 799 | 2606 | 1038 | 39.8% | 60.2% |
| 4 | 799 | 3405 | 1086 | 31.9% | 68.1% |
| 5 | 864 | 4269 | 1193 | 27.9% | 72.1% |
| 6 | 996 | 5265 | 1405 | 26.7% | 73.3% |
| 7 | 1010 | 6275 | 1505 | 24.0% | 76.0% |
| 8 | 847 | 7122 | 1429 | 20.1% | 79.9% |
| 9 | 751 | 7873 | 1345 | 17.1% | 82.9% |
| 10 | 762 | 8635 | 1378 | 16.0% | 84.0% |
| 11 | 783 | 9418 | 1436 | 15.2% | 84.8% |
| 12 | 770 | 10188 | 1439 | 14.1% | 85.9% |
| 13 | 856 | 11044 | 1569 | 14.2% | 85.8% |
| 14 | 1012 | 12056 | 1756 | 14.6% | 85.4% |

## Appendix F

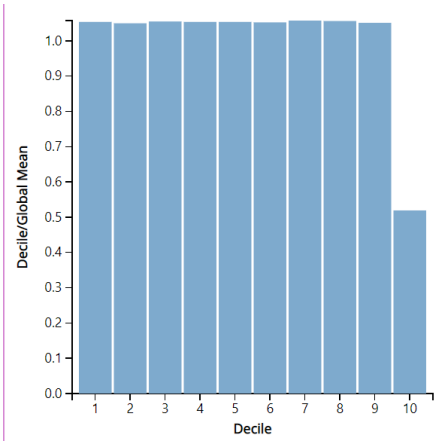
*Best Pruned Tree*



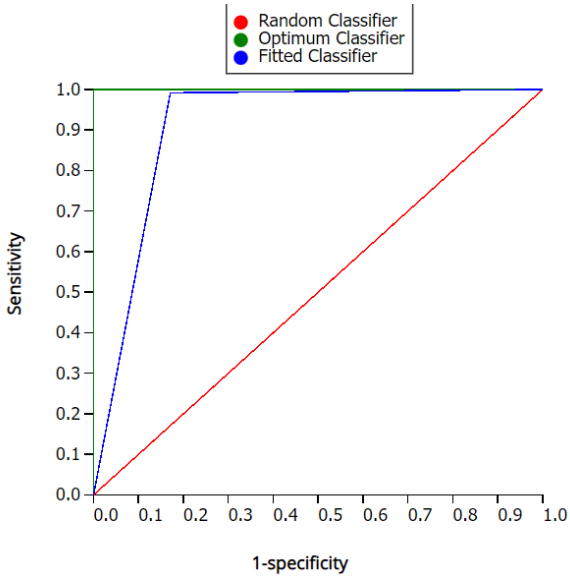

*Lift Chart for Test Data*




*Decile Wise Lift Chart for Test Data*




*RROC Curve for Test Data*




Appendix G

*Revenue Overall Cumulative*

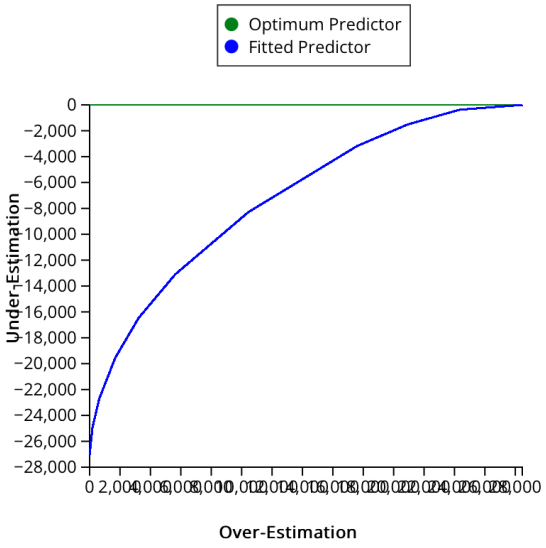
**Coefficients**

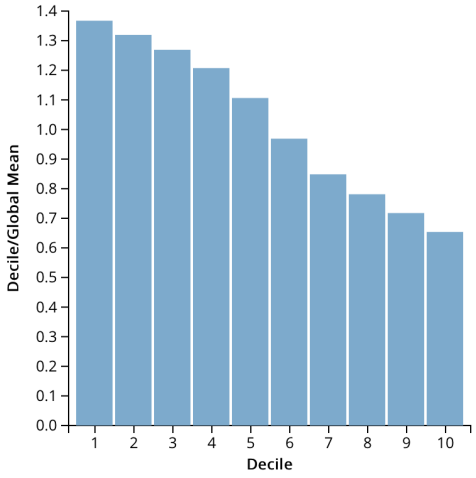
| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -3068.88671 | -3988.999975 | -2148.773445 | 449.8826177 | -6.82152764 | 1.7203E-07 |
| day | 555.3674355 | 505.1711416 | 605.5637294 | 24.54310892 | 22.62824312 | 5.6383E-20 |

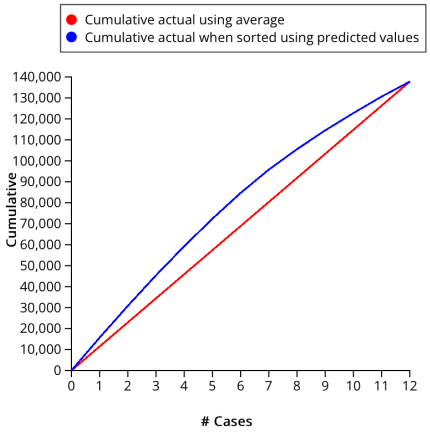
D90 Predicted Revenue Cumulative
Predicted Revenue= -3,068.89 + 555.37 (90) = $46,914.41

D180 Predicted Revenue Cumulative
Predicted Revenue= -3,068.89 + 555.37 (180) = $96,897.71

| Metric | Value |
|---|---|
| SSE | 1348466.651 |
| MSE | 43498.92424 |
| RMSE | 208.5639572 |
| Avg. Error | 1.57695E-13 |
| MAPE | 45.75575781 |
| MAD | 150.7041332 |
| R2 | 0.60037326 |






*User Count Overall Cumulative*

**Coefficients**

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Standard Error | T-Statistic | P-Value |
|---|---|---|---|---|---|---|
| Intercept | -907.5677419 | -1407.992211 | -407.1432728 | 244.6788658 | -3.70922 | 0.00087566 |
| Day | 885.9629032 | 858.6625103 | 913.2632962 | 13.34832647 | 66.37258276 | 3.1197E-33 |

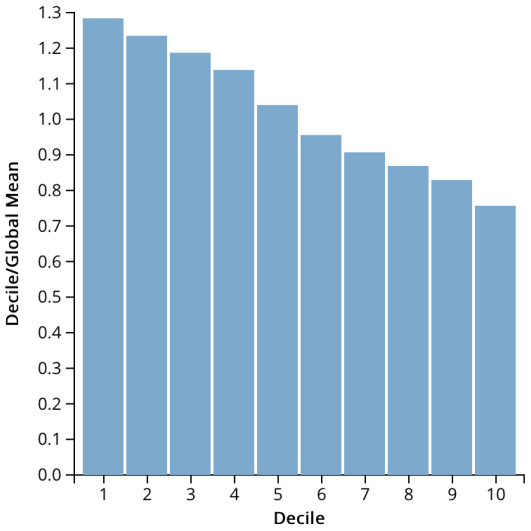
D90 Predicted User Count Cumulative

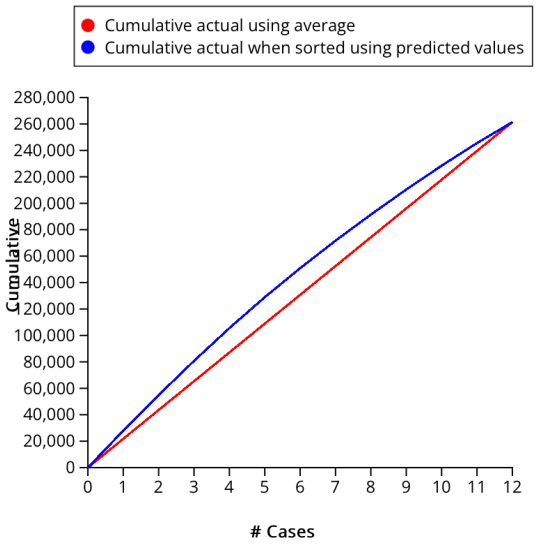
Predicted User Count= -907.57 + 885.97 (90) = 78,829

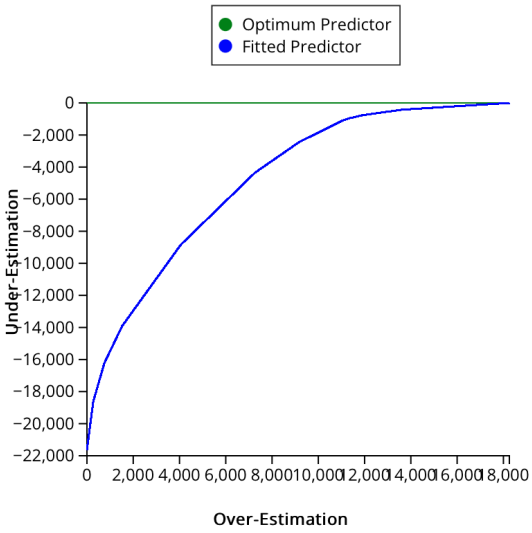
D180 Predicted User Count Cumulative

Predicted User Count= -907.57 + 885.97 (180) = 158,567

| Metric | Value |
|---|---|
| SSE | 738429.3351 |
| MSE | 23820.30113 |
| RMSE | 154.3382685 |
| Avg Error | -7.33463E-14 |
| MAPE | 12.89838207 |
| MAD | 113.3470083 |
| R2 | 0.417548951 |

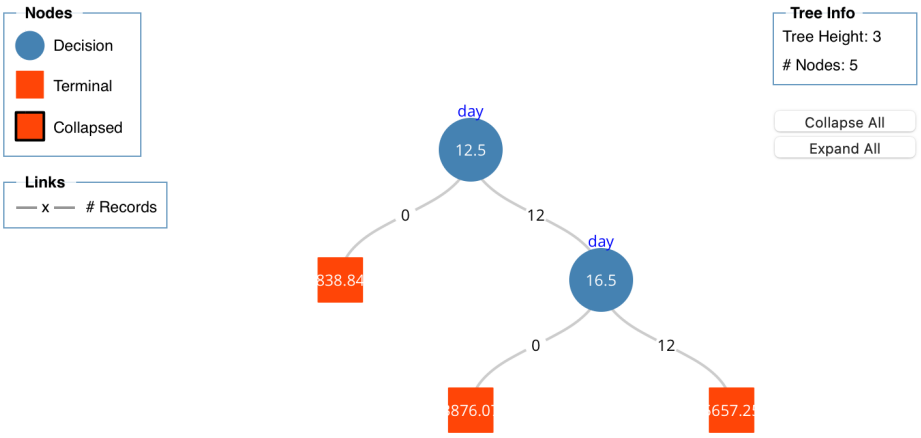



Appendix H

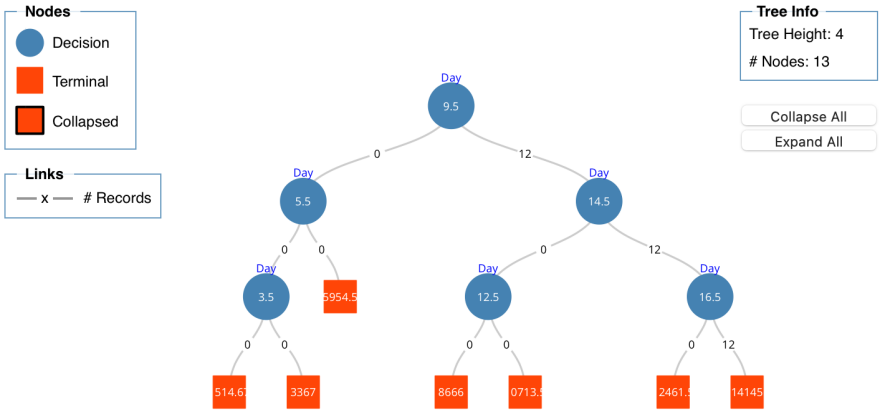
*Cumulative Revenue Regression Trees and Metrics*




| Metric | Value |
|---|---|
| SSE | 5427046.332 |
| MSE | 285634.0175 |
| RMSE | 534.4473945 |
| MAD | 444.646759 |
| R2 | 0.930652326 |

| Metric | Value |
|---|---|
| SSE | 198709943.1 |
| MSE | 16559161.93 |
| RMSE | 4069.295016 |
| MAD | 3702.424211 |
| R2 | -0.955173399 |

*Cumulative User Count*




| Metric | Value |
|---|---|
| SSE | 246057.326 |
| MSE | 12950.3856 |
| RMSE | 113.799761 |
| MAD | 93.4094183 |
| R2 | 0.999321 |

| Metric | Value |
|---|---|
| SSE | 43213420.9 |
| MSE | 3601118.41 |
| RMSE | 1897.6613 |
| MAD | 1542.32281 |
| R2 | 0.75537484 |

Appendix **I**

Assumptions:
Region 1=Scandinavia (DK, FI, SE, NO)
Region 2=British Isles (IE, GB)
Region 3= Central America (CR, GT, SV, MX, PA)

*Project Log*
**Introduction to Business Analytics in Gaming**

| Date & time | Activity | Sami | Emily B | Emily N | Ellie |
|---|---|---|---|---|---|
| 2/11/22<br><br>2 hrs | Worked on the first project deliverable. Created shared documents and collaborated on which methods to approach this project would be best. Wrote up first bullet point | X | X | X | X |
| 2/12/22<br><br>1.5 hr | Worked on graphs and other bullet points. Came up with suggestions for new graphs | X | X | X | X |
| 2/13/22<br><br>2 hrs | Met over zoom and fixed up graphs. Came to a consensus on whether puzzle games should or shouldn't be released. Finished the last bullet point and turned in the project. | X | X | X | X |
| 2/16/22<br><br>1.5 hrs | Met over zoom to go over professor's suggestions and fix up our lab 1. Created new graphs. Made calculations in | X | X | X | X |

| | | | | | |
|---|---|---|---|---|---|
| | excel. Added new graphs to dashboard. | | | | |
| 2/20/22<br><br>3 hours | Made final revisions to project, added graphs, revised commentary. | | X | | |

**Game Understanding**

| Date | Activity | Sami H | Emily B | Ellie C | Emily N |
|---|---|---|---|---|---|
| 2/21/22<br>2 hrs | Worked on cleaning the data and organizing Excel sheets for Tableau. Made about 7 graphs in Tableau. Made an outline in google doc for how we want this lab to play out. | X | X | X | X |
| 2/23/22<br>1.5 hrs | Created more graphs, made a histogram, worked on retention rate calculations, worked on ROI calculations. Discussed on what data points should be kept and what should be discarded. | X | X | X | X |
| 2/25/22<br><br>2 hrs | Discussed where we are progress wise and how we should go about splitting the rest of the lab deliverables. Worked together on our own separate duties. | X | X | X | X |
| 2/27/22<br><br>2 hrs | Calculated the rest of the user retention rates on excel and wrote up explanations in google doc | | | X | |
| 2/27/22<br><br>2 hrs | Worked on calculating the rest of ROI and RPU in Excel and included that in the doc with additional write up | | X | | |
| 3/2/22<br><br>2.5 hrs | Started pulling past data from Excel to create a more valid comparison between us and our competitors. Retrieved the data for the first 31 days of our competitors and calculated the RPU of individual countries and regions | X | | | |
| 3/3/22<br><br>1.5 hrs | Worked on coming up with a Business question and communicating with Sami on which dashboards could be | | | | X |

| | | | | | |
|---|---|---|---|---|---|
| | added/edited. | | | | |
| 3/4/22<br><br>1.5 hrs | Finished pulling RPU data from competitors and made graphs on it to be put in the google doc. Wrote some additional paragraphs and cleaned up the sheet. | X | | | |
| 3/6/22<br><br>2 hrs | Ellie, Emily B, and Emily N read over Sami's comments on the google doc and gave their ideas/recommendations and fixed up small errors. Sami fixed up the dashboard according to recommendations. Emily N and Sami reviewed the doc before turning it in. | X | X | X | X |

**Revenue Forecasting**

| Date | Activity | Sami H | Emily B | Ellie C | Emily N |
|---|---|---|---|---|---|
| 3/22/22<br>45 min | Met to discuss plan for Lab 3. Split work into sections and decided who would work on what section. Made a plan on when all of our parts should be done by. | X | X | X | X |
| 3/23/22<br>2 hours | Worked on organizing data per network in order to partition and run regression for each network to find predicted revenue. | | X | | |
| 3/24/22<br>2 hours | Worked on organizing data per network in order to partition and run regression for each network to find predicted user count. | | X | | |
| 3/24/22<br>1.5hrs | Worked on organizing data per region in order to partition and run regression for each region to find predicted revenue. | | | X | |
| 3/25/22<br>1.5hrs | Finished running regression for each region. Worked on calculating the user count per region and then running regression to find predicted user count. | | | X | |
| 3/25/22<br>1 hour | Put all information, visuals and interpretations into google documents. | | X | | |
| 3/26/22<br>1.5<br>hours | Found the overall revenue and user count regression equation to find the predicted user count and revenue. | | | | X |

| | | | | | |
|---|---|---|---|---|---|
| 3/27/22<br>3.25 hours | Went through the Google Doc and read up on everybody's calculations. Created a business question that we briefly discussed and went into detail on how our calculations will help answer that question. Wrote observations for each section (revenue per network, revenue per region, user count per network, and user count per region) and also created graphs for each section to give more visual observations. Wrote the "final observations" section and inserted two more graphs displaying overall results. Formatted google doc and turned in lab. | X | | | |

**Churn Classification**

| Date | Activity | Sami H | Emily B | Ellie C | Emily N |
|---|---|---|---|---|---|
| 3/29/22<br><br>2 hrs | Met to discuss plans for Lab 4. Split work into sections and decide who would work on what section. Made a plan on when all of our parts should be done by. | X | X | X | X |
| 4/3/22<br>2 hours | Worked on creating and evaluating logistic regression models. Figured out a way to calculate overall, and daily retention rates for the game. | | X | | |
| 4/5/22<br>3 hrs | Calculated overall retention and churn rate of game and reorganized data | | X | | |
| 4/10/22<br>1.5 hrs | Input all findings into document and made final comments on the model and the churn rate over time | | X | | |
| 4/1/22<br>1 hr | Started working on organizing data and setting up Excel sheet in a way that would make sense for this lab. | X | | | |
| 4/4/22<br><br>2 hrs | Worked on creating pivot tables for each region (a pivot table for unique users installing and pivtot table for unique users logging into the game). | X | | | |
| 4/7/22<br>1 hr | Did additional research on churn and retention rate. Watched youtube videos, read articles, and tried out different methods on Excel sheet | X | | | |

| | | | | | |
|---|---|---|---|---|---|
| 4/8/22<br>15 min | Sami and Ellie met to discuss their progress and analyze results | X | | X | |
| 4/10/22<br>1 hour | Designed a business question and reported overall performance, finalized everything | | | | X |
| 4/10/22<br><br>2.5 hrs | Finished up calculating cumulative users, churn rate, and retention rate for each network. Organized the Excel sheet to be presentable for submission. Worked on google doc and inputting and interpreting my findings. | X | | | |

**Comparing Methods**

| Date | Activity | Sami H | Emily B | Ellie C | Emily N |
|---|---|---|---|---|---|
| 4/12<br>2 hours | Met to discuss plans for final lab. Assigned sections to each group member and went through which calculations we needed to make to make our final conclusion. | X | X | X | X |
| 4/12<br>1 hour | Formatted Google Document, calculated regression models for competitor games, and made conclusions in google documents. | | X | | |
| 4/13<br>1 hr | Wrote final assessment of MLR, Logistic Regression Model | | X | | |
| 4/15<br>3.5 hrs | Answered Greenlight questions, created visuals to summarize the calculations for D90 and D180 comparisons, and went over all questions to be sure of submission. | | X | | |
| 4/12<br><br>2 hrs | Started working on other models for categorical data. Worked on researching other models to use besides classification trees | X | | | |
| 4/13<br><br>2hrs | Continued working and finishing up on calculating and writing categorical target variable section. | X | | X | |

| | | | | | |
|---|---|---|---|---|---|
| 4/14<br>2 hrs | Caught up on where we were at and then planned the next half of the project. Worked together for a couple of hours. | X | X | X | X |
| 4/16<br><br>2 hours | Redid multiple regression models, created regression trees for continuous variables and evaluated those two models and wrote most of the best method for continuous | | | | X |
| 4/17<br><br>3.5 hrs | Finished up my part of the competitors section (brick n balls and hello neighbor). Wrote additional sections. Created appendix. Cleaned up document. | X | | | |

**Final Report and Presentation**

| Date | Activity | Sami H | Emily B | Ellie C | Emily N |
|---|---|---|---|---|---|
| 4/17<br>3 hours | Created Final Presentation in Google Slides. Started the formatting and layout of what to present in class on 4/21. Went and fixed MLR calculations to get more accurate predictions for D90 and D180. Fixed analysis in report 3 for the final submission. Created table for comparison of competitors and game in beta for report 5. | | X | | |
| 4/18<br>3 hrs | Fixed up lab 5 for final report where we should've compared to our competitors using 3 regions & not all the data. Added a paragraph explaining it and more to appendix. Fixed up additional sections that needed editing for final report. | X | | | |
| 4/18<br>3 hours | Fixed feedback from lab 5 to update findings for final presentation. Updated tables with scaled data from smaller region comparison. Did calculations for ROI for overall comparison and added into the final report, as well as lab five. | | X | | |
| 4/19<br>2.5 | Worked on creating my slides and writing up a paper on what to discuss for final presentation. Looked at other | X | | | |

| | | | | | |
|---|---|---|---|---|---|
| | slides to think about discussion topics for when we meet later tonight. | | | | |
| 4/19<br>1 hour | Met to practice our presentation as well as finish our slides before practice presentation Wednesday morning. | X | X | X | X |
| 4/19<br>2 hrs | Worked on trying to figure out how to calculate churn using classification trees for competitors. Continued my presentation prep sheet. Looked over presentation slides and animations. | X | | | |
| 4/20<br>.5 hrs | Met in the morning to discuss and give practice presentation to Dr. Gudigantala | X | X | X | X |
| 4/20<br>2 hours | Did outside research for comparison of churn and retention rates for final presentation. Wrote this research into the final report as well as finding graphics and putting them into slides. Redid the churn and retention rates for overall data and created new graphs to put in the presentation. | | X | | |
| 4/20<br>30 min | Looked over my slides and transitioned them to Powerpoint for Moodle Turn in. Turned in slides | X | | | |
| 4/21<br>2 hrs | Practiced Presentation. Created Slideument. Printed off copies of it for presentation. | X | | | |
| 4/24<br>2 hours | Made edits to the final report and reworded sections to create a comprehensive flow. Updated all graphs and wrote in new sections that resulted from feedback from the presentation. Updated project log and worked on formatting for submission. | | X | | |
| 4/24<br>2 hfs | Wrote executive summary and fixed sections from feedbacks. Extracted information from previous labs to implement in final report. | | | | X |
| 4/24<br>3 hrs | Added an updated graph photo to one of my sections. Worked on formatting the final report (with photos, appendix, and project log). Created final appendix and updated references to refer to the | X | | | |

| | | | | | |
|---|---|---|---|---|---|
| | right letter. Went through old emails to find everyone's excel file. Emily N and I turned in final report. | | | | |