

# How can we increase revenue from Catch the Pink Flamingo?

MOLDOVAN DORIN-VASILE

## Problem Statement

How can we use the following data sets to understand options for increasing revenue from game players?

- 8 CSV files that contain the simulated game data and log data for the game
- 1 CSV file which aggregates data from several data files
- 6 CSV files that represent the simulated chat data

My opinion regarding why the different kinds and sources of data are so important for Eglence to be able to identify new revenue opportunities is that the different kinds and sources of data can provide information and insights that can be hardly observed in a single source of data. Each kind and source of data corresponds to a different aspect of the game.

The analysis of this data makes the process a data science story as it involves:

- Different sources of information which are generated in real time
- Large quantities of information that are stored in the simulated game data and log data files
- The possibility to explore, classify, cluster and analyze data

The first 8 CSV log files contain information about the users, the teams, the team assignments, the user sessions, the buy clicks, the ad clicks, the game clicks and the level events.

Next, a single CSV file is created which aggregates data from several game data files.

Finally, the last 6 CSV files represent the simulated chat data which is related to the game.

## Data Exploration Overview

The data exploration uses 8 CSV files which contain the simulated game data and log data for the game. The operations performed on this data are of two types:


- Aggregation
- Filtering

The most important thing(s) to convey from my experience with exploring the data using the aggregation and the filtering operations are:

- The items should be promoted in increasing order of price so that the products with higher prices are purchased by more users than the products with lower prices. The data exploration results show that the most purchased item is associated with the third amount of money spent on products.
- The iphone platform should be promoted more than the other platforms because the top three users used iphones

## What have we learned from classification?

The classification was performed on a CSV file that aggregates data from several game data files. The following operations were performed:

- The selection of the samples
  - The creation of a new attribute which has 2 categories (HighRollers & PennyPinchers)
  - The partition of the data in train and test datasets
  - The application of the Decision Tree algorithm
- 

The most important or remarkable points from my classification analysis are summarized next:

- From the total number of samples only approximately one third made purchases
- The creation of the new attribute which has two categories (HighRollers and PennyPinchers) was necessary because using this approach it is possible to apply a classification algorithm
- The trained dataset was used to create a decision tree model and the test data set was used to test the algorithm
- The application of the Decision Tree algorithm shows that the difference between a HighRoller or a PennyPincher is influenced by the type of the platform that is used. The users that use iphones are more likely to be HighRollers while the users that use android, linux, windows or mac are more likely to be PennyPinchers
- The users should use platforms that produce high revenue such as iphones

## What have we learned from clustering?

- The clustering results differ significantly depending on the selected attributes and the number of clusters
- The dataset used in the experiments was the same dataset used for classification and the number of clusters was set to 2
- The results showed that the teams with higher levels had lower values for the number of game clicks and for the number of hits




The dataset used for clustering is the same dataset as the one used for classification. The dataset contains information about users, buy clicks and game clicks.

Next are summarized two recommended actions:

- The difficulty of the game should be decreased for the levels that have high values
  - The cluster analysis showed that the highest values for the number of clicks and the number of hits were obtained when the team level had low values.
- The conditions to pass from one level to another level should be relaxed – The conditions to pass from one level to the next level should be relaxed to maximize the number of users that participate in the levels that have higher values. Also more users generate more clicks and consequently more revenue.

## From our chat graph analysis, what further exploration should we undertake?

- The graph exploration was performed on the 6 CSV files that represent the chat data
  - Several operations can be performed on the graph such as:
    - The determination of the longest chain and its participants
    - The analysis of the relationship between the chattiest users and the chattiest teams
    - How active are the users
- 


The most important or remarkable points from my graph analysis are:

- The length of the longest chain (9) and the number of distinct users that participate in the longest chain (5) can provide information about the users that have the initiative to start and to maintain discussions
- The analysis of the relationship between the chattiest users and the chattiest teams shows that only one user from the top 10 chattiest users is a member of one the top 10 chattiest teams
- The analysis of the clustering coefficients shows that the top chattiest users do not have the highest clustering coefficient

## Recommendation

This recommendation is a result of the analysis performed in the data exploration, classification, clustering and graph analysis:

- The users should be grouped in teams according to how chatty they are. The chattiest users should be grouped in a single team. This way they should participate in longer conversions and if these conversations contain a large number of ads then the company can improve the value of the revenues.



The results from the chat graph analysis show that the chattiest users are not necessarily members of the chattiest teams. Only one user from the top 10 users is the member of a team that is in the top 10 chattiest teams. Also, the chattiest users do not have the highest values for the clustering coefficient. My opinion is that if the chattiest users are grouped in the same team then it is possible to generate more discussions. From time to time it should be possible to change the teams according to how “chatty” the users and the team are. Longer conversations should be a great opportunity for the inclusion of more ads in the chat. More clicks can generate more revenue for the company.