

II1
II2

How can we increase revenue from Catch the Pink Flamingo?

Luis A. Romero Gamarra

Final project for Big Data Specialization 2017 through Coursera. This project use Big Data tools for create new opportunities to Eglence Inc. with objective of increase revenue through Catch the Pink flamingo game.

Problem Statement


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

We have the follow data sets:

Flamingo-data, contains data about the game, user information, purchasing activities, teams activities and chats informations. Then, we can analysis this data for obatain data insight and create data-driven knowledge.

Combined-data, contains all information in one data set, this is important to create new summaries that show currently information before analysis in mode general.

Chat-data, contain information about user activities in context of group activities like chats, new teams and interaction between users into each chatroom. Analysing this data we can know the user behaviorial in the game.



The data set is composed of three files in csv format with simulated gaming and chat data. This information is a point of start for analysis and understand the model data.

Flamingo-data contain information about users, game register, in-app purchase and different team users. Combined-data contains aggregated data from all dataset in one archive that use to create summaries with objective of understand it and create aggregate filters and obtain new data insight before the analysis. Chat-data contains simulated data related chat activities in the game, analyzing this data we can see user and teams behavioral, communities behavioral and found what they are more attractive to help with objective of increase revenues.

Diapositiva 3

Data Exploration Overview

- Between user platforms, Iphone users have most purchased products, follow for Android, Mac and Linux. Windows user are purchase less.
- The two most coomonly clicked categories fo ads are computer with 2638 and games with 2601 hits respectively.
- The total amount of money spent buying for all user was \$21,407
- There are 6 items available to be purchase and their cost si between 1\$ and 20\$.
- Two most commonly purchased products are item 2 with 714 and item 5 with 610 purchased respectively. Items 5 have an revenue total of 12,200\$ and items 4 have 4250\$
- The top 10 user spend range more money is between 172\$ and 223\$, the userId 2229 is the most spend, and use Iphone platform.
- The average teams is composed for 78 users.

Understand the data, exploring data exhaustively is the most important. Splunk have all tool for exploring data and create different filters, statistics, format and aggregations to understand the information and create new insight.

We found different platforms, founding that user use more iPhone platform than others. User most commonly click on ads for computers and games categories. There are six items for purchase, of they only two (items 2 and 5) are most commonly purchased. Each item has a range between 1 and 20 dollars.

We found the total purchase in-app was \$21,407, and the top user is 2229 and spend around \$200.

What have we learned from classification?

Through the Decision Tree analysis we can identify who are big spenders and little spenders, in this case, HighRollers and PennyPinchers respectively.

Then, we can classify this kind of user through average price, creating a new attribute depends of range of Price (more 5\$ for HighRoller and less or equal 5\$ for PennyPinchers). For that KNIME offer a set of options for create different types of partitions.

We learn to create two partitions, for one side, we create a train data used to create a new model. For another side, the test data that together with the trained model will be applied to create a final results.

We obtain that users that use iPhone platform are most likely to be a HighRoller, iPhone users spend more money than others. The user that use linux, android, windows and mac have most probability to be a PennyPincher because spend less in purchase.



Decision Tree analysis in KNIME help us to classify different information, in this case we identify who are big spenders and little spenders. We learn to load data in KNIME and work with different tools.

Preparing data dropping null values before analysis is essential. Create new categorical attribute from numeric attribute through numeric binner that help to define new category which it will be the target for the prediction. This enable analysis for users are big spender (HighRoller) and little spender (PennyPinchers).

This data was partitioned into train and test dataset. Train data was used to create trained model that applied with test data set. This is important because this two partitions allow define the different parts to achieve our prediction. In this case, train data allow to learn of input data and create a model that, in next step, it will be apply to test data through decision tree predictor. For last, the result is compared for perform the accuracy through error measure.

We concluded that users that use iPhone platform are most likely to be a HighRoller, iPhone users spend more money than others. The user that use Linux, android, windows and mac have most probability to be a PennyPincher because spend less in purchase.

What have we learned from clustering?

Create new attributes for generate new data insight through PySpark, where we analysis this data using Kmeans Model.

In this part we create three attributes: totalBuyId, totalHits and revenue. The objective was relationed this attribute for create a training data that show groups of clusters.

We obtain tree cluster where for each one we obtained a cluster center that suggest differences between each cluster.

So, First number (field1) in each array refers to scaled version of the number of buy-Id and the second number (field2) is the scaled version of the totalHits, and the third number (field3) is the scaled version of the revenue per user.

```
In [18]: centers = model.clusterCenters()
         centers
Out[18]: [array([-0.64645597,  0.12239214, -0.61965848]),
         array([ 0.43584599, -0.25797935,  0.34764005]),
         array([ 2.24868783,  0.14530832,  2.38374319])]
```

In this analysis we using Kmeans model for partitioned users into groups based on characteristics such as their game playing behavior, purchase behavior, inclination to click on displayed ads.

For this analysis we create three attributes:

- **totalBuyId**, attribute that count total item buying for each user. This attribute is from buy-clicks.csv columns and it is important for analysis in-app purchase.
- **totalHits**, attribute that count total hits clicked for each user, if user was click their value is 1, otherwise is 0. This new attribute is from game-clicks.csv file.
- **revenue**, it is the aggregation of the total spending by user.
The analysis tried to find out if this number is related to total number of hits and buyIds by users.

First number (field1) in each array refers to scaled version of the number of buy-Id and the second number (field2) is the scaled version of the totalHits, and the third number (field3) is the scaled version of the revenue per user.

These clusters can be differentiated from each other as follows:

Cluster 1 is different from the others in that it has lower values than clusters 3. All fields have lowest values, so this show that few total buy items means user is new or has no interest in the items.

Cluster 2 is different from the others in that it has high values than clusters 1. All fields have highest values, so this show that more total hits means user tend to purchase more in the app, except in the field 2 (center) since the values is lowest, this mean that currently exist

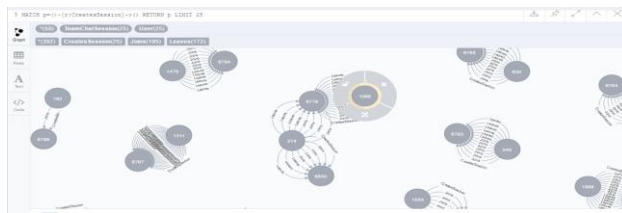
Cluster 3 is different from the others in that this cluster have high values than other cluster except in the field 2 that mean the user can be click hits more times but not help to increment revenue. Otherwise, the highest values in the fields 1 and 3 show that user with buy more item tending to increase the revenue.

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

Chat data set bring knowledge to know behavioral players and their with other group or teams.

The analysis focused in this information, we have found the user most active in longest conversations or interactive neighborhoods and team closely communicate examining different nodes and edges in a graph, this information is essential to create new business opportunities.

Thank to this, we know the target users or teams where create advertising campaigns and offers new products.



We use Neo4J tool for analysis this part of project. We load chat-data files and create constraints information about data model. This data contain chat session of team users and log from each event create.

We create the database through input sentence to configure the schema, after this execute the script for generate tables and contents for each node and edges. This part is very important because it is point of start to analysis.

After tables are created and the edges are specify, we proceed to perform various queries to find longest conversation chain using edge label and participant in this conversation.

Analyzing who chattiest user from each chat team show each user that most active through cluster coefficient. This is important to know users as target for increment ads or create campaign and offer new products.

Recommendation

After the exhaustive analysis carried out in the different phases, we conclude that to increase revenues the company should pay attention to the following points:

Focuses in iPhone user, this user spend more money in purchase than other user in the different platforms. It will be important create new products or promoter the currently offer because through analysis we know that this user spend most money than others.

Make new promotion and focused in the items, we can make new item according level of user, for each historical of item, we can down their price or we can sell it, create determinate item for a determinate kind of user for increase their value. For this last sentence, it is important know the user behavioral, in this case, we know user with more clicks (hits) and many session are good target to show more ads than other users. We recommends create new items to new users for increase engagement and adding more ads for user more actives.

For finished, focuses in the user most active in each chartroom, they are potential consumers for ads and new items. With this recommends we are sure that Eglence Inc. will increase revenues and engaged new users.

Focused in iPhone platform because their user are big spenders, create more offers and products for this platform, it is important to increasing revenue and improve the engagement.

For user and teams more active, create gift for each achievement or level gained, remember that team more active are more attractive to click in more ads, for this reason it is important encourage the engagement focused in iPhone platform. With this advice, Eglence Inc. can increase the revenue and engage its users only if care users.