Successful social games and their super-power: Big data analytics

Ioana Roxana STIRCU
The Bucharest University of Economic Studies
roxana.stircu@gmail.com

This article is a short presentation of big data analysis and game analysis. The paper describes the case of social games, and observes the huge improvement that big data analysis has on social games and their success. It also contains a presentation of Pokémon GO, and its evolution on the market, from launch until today. A set of metrics and algorithms are proposed, that can be used to improve game features and monetization. In the last section, I apply a Naive Bayes classifier, using WEKA, on a set of data collected from social media networks, to predict how using a game that implies walking influences the amount of daily steps a player makes.
Keywords:*big data analysis, Naive Bayes classifier, WEKA, monetization, Pokémon GO case study, social networks, social games*

1 Introduction

In the recent years, together with other industries that have experienced an exponential growth, the gaming industry has also evolved incredibly well. One of the main reasons, aside introduction and fructification of the social gaming on social networks, using new platforms (such as consoles and mobile cell phones), has been introduction of game analytics. There are many gaming companies on the market using game analytics in the most important steps of a gaming project. The most frequent cases are: define user behaviour, identify patterns in the process of acquisition and retention, test and optimize social media ads, calibrate player experience in game, improve game design in order to be suitable for every type of player, predict and optimize players' retention and monetization.

1.1 Big Data and Analytics

Big data is a term generally used to denote a huge amount of data, of a large variety of types, which are processed and analysed to gain all the information and insights resulted from it. One of the defining characteristics of big data is the fact that, in order to get the value from data fast enough to make the corresponding decisions and improvements, all the existing platforms and tools have to be updated and re-designed in order to meet the new business needs. The four V's of big data are the main properties and problems in the field ([1], [2]):

- volume – the needed space capacity is incredibly big, and is exponentially growing;
- variety – data is structured and un-structured, from a lot of different sources, like applications, websites, social networks, emails, news, etc. to video, audio, texts, logs;
- velocity – big amounts of very variate data can be processed in a long time; sometimes only the process of gathering, cleaning and formatting data from different sources can take too long, add the analysing time and the decisions are made too late;
- veracity – when exploring the universe of big data, errors and mistakes can be very hard to track and fix, so the process should be made very carefully.

In order to meet the requirements of big data, has been developed new solutions for storage, called NoSQL (not only SQL) ([3],

[4]). Document storage is a NoSQL method, and is made using some types of documents, such as XML, YAML, JSON, BSON. Another storage method is graphs database, where the graphs properties like nodes, edges and orientation are used to describe data and relations between data nodes. One of the widely used solutions is columnar-based storage, where data is represented based on columns (or lines), that can be grouped logically [5].
The standard solution for big data management is represented by Apache Hadoop, a software platform coded in Java, which runs on a cluster of servers. This properties explains the performance and efficiency of Hadoop, as it is scalable according to business needs, due to MapReduce ([6], [7]).

2.1 Gaming Industry

The gaming industry can be divided into three big categories, according to the type of games developed within that category. The traditional video and MMO games doesn't seem to be the most important category of games anymore, and is represented predominantly by male players, with payments made via cash and credit card. Social games have the advantage that are offered at smaller prices via social networks, and are also played on mobile devices. This category include approximately 54% female players, using mobile payments and PayPal accounts to make purchases, and are playable across multiple platforms and devices. The mobile games represents a fundamental change in gaming industry, with all new methods to redefine the landscape: with mobile gaming, now anyone can play games, anywhere they want, and having a huge list of games to choose from, as the majority of games had adopted freemium monetization.

2.2 Game Analytics

Some very big and successful gaming companies understood that game analytics, and gaining insights from players' generated data, is a very powerful tool, which can ease their path to success in their project. Many companies have a lot of projects based entirely on game analytics, pattern recognition or predictions of the best mix to use in different business processes. This step forward also changed the project of making games, from the structure of the previous projects where a big teams worked at a game for a few years, and then released the disc/CD on the market, waiting for the sales numbers. The game has changed: now, all the smart gaming companies make brainstorming and prototypes for a few games then, using game analytics with targeting and player segmentation, analysis and prediction, they can decide which of the projects ideas are more successful, and develop only the selected projects. From this point, all the features added to the initial projected are tested, optimized and designed to meet the desired properties, according to patterns and insights gained during testing and prototyping process.
The trend of using data analysis to improve the game development, combined with the importance and usage of social factor and social networks in some types of games and, most important, the huge development of mobile cell phones in the last few years, has projected the game analytics in the direction of big data analysis. Mobiles casual games or MMORPGs with millions of players, with hundreds of missions, maps, puzzles and sessions, with dedicated players that access the app on a daily basis, with years of history and progress, all together represent the growing picture of big data in gaming.
For example: every day, *Pokémon GO* have over nine millions of daily active users only from the USA, who are logged into their preferred social game [8], continuously generating data, while social and mobile gaming is still a fraction of the $80 billion video game market. The entire industry is based on data and insight. Because a big part of the revenue comes from a small percent of the initial players, is crucial to understand player preferences and behaviour when even

the smallest improve has a fundamental impact on monetization and revenue.

2.3 Monetization and KPIs of Gaming

The process of game monetization [9] represents the moment when a video game starts producing money for the company and for the people involved in the creation and development. Monetization should start long before this moment, and the monetization strategies and models should be created at the beginning of the project. This process had a great evolution in the past decades. In the initial model of monetization, in the 80-90's, the user simply bought the disc or the console from the store. The developers followed the sales and fans, and had no other information about the game's success chances before launching the disc. In the last few years the monetization changed radically, as people are increasingly logging into social gaming apps through mobile phones and Facebook. The new projects are developed having a monetization model in mind, and the developers launch a small test part of the game. Using test results and insight gained, the company develops the next features and calibrate their monetization methods to be successful in the final part of their project. Monetization can be made by the application, but also by events related, campaigns, products, and merchandise.

For the game monetization, also called game subscription, a lot of subscription models are developed by gaming companies, according to their game types and platforms. Most important models are *pay-to-play* (players have to pay a monthly subscription fee to be able to play the game), *free-to-play* (usually involves an upfront cost and no additional payments), and *freemium* (game access and game content is granted for free for all users, but offers the option to pay for additional access and content). This models of monetization are applied via retail purchase, in-game micro transactions, digital download, indirect monetization (like product placement, banner advertisement or commercial breaks), and even the newest trend of monetization, called crowdfunding (funding independent gaming projects – the developers can raise enough money before the development process starts).

Zynga is credited for starting social gaming in 2009, when they introduced *Farmville* for Facebook. From this moment, users started spending real money to gain credits in the virtual world of social gaming in no time. Other projects that successfully combined social gaming and great monetization models are *Candy Crush*, *Clash of Clans* and *Clash Royale*, *Pokémon GO*.

In social games, the most used metrics in analysing players' activity, also known as KPIs (key performance indicators) are:

- *DAU* is the number of daily active users, *MAU* is the number of monthly active users, *DAU* divided by *MAU* ratio;
- *sessions* are determined every time the player opens the game;
- *retention* represents the number of players still retained in game after a number of days, most used types are *second day retention*, *seven days retention*, *twenty-eight days retention*;
- *engagement* can be defined as the amount of time spent playing a game;
- *K-factor* measures the growth and "going viral" property of a game, based on virality actions and their success;
- *ARPU* is the average revenue per paying user;
- *LTF* is the lifetime value, which can be calculated based on in-game purchases, but also other

monetization-related actions, such as influence on virality, shares.

3.1 Case Study: Pokémon GO

Pokémon GO success started last summer, in July 2016, right after its launch on USA market and a few other selected countries. In the first week of availability, the game became the most downloaded mobile application ever, surpassing *Candy Crush* and *Clash Royale*. On February 27, 2017, Niantic announced that the game has surpassed 650 million downloads and that players have walked from here to Pluto since launch [11]. The app is a location-based augmented reality game, developed by the studio called Niantic – in collaboration with Nintendo, both for iOS and Android devices.

The players use the mobile device' GPS to locate virtual creatures, called Pokémon (pocket monsters), who appear on the screen as if they were in the same real-world location as the player. Once located, the player has to walk to reach the selected Pokémon, and then he can capture, battle, and train the virtual creatures. The main goal is to catch as many monsters as you can, and to complete your Poke dex, as you level up and are able to catch more complex and powerful creatures (the dex can be completed by hatching Pokémon eggs, and a Pokémon can also evolve by catching more of the same type). This way you can become a Pokémon trainer, take on Gym battles and defend your Gym, for the prestige and ownership of Gyms. By visiting Poke stops, located in interesting places like museums, bars, parks or coffee shops, the player can stock up on Poke balls and other very interesting and helpful items.

3.2 Gameplay and Game Mechanics

After creating a new account, the player choses his customized avatar. This avatar is then displayed on the game's map – generated and updated based on player's current geo-location. The geo-map is updated with Poke stops, which can provide eggs, Poke balls, berries, potions or lure modules (used to attract rare Pokémon, and Poke Gyms, which are battle locations, displayed in places of interest (like parks, coffee shops, bars, public institutions). The main feature is that player's movement in real life makes his avatar progress on the game's map. Every time a player encounters a Pokémon, viewed in augmented reality mode or on a generic background, and generated using the camera and gyroscope – displayed as it were in real world, he can take screenshots or catch it directly with a Poke ball by flicking it from the bottom of the screen up toward the monster. This way the Pokémon is caught and added to the inventory, but the player also gets two type of in-game currencies, candies and stardust, that are used to evolve a Pokémon. Depending on the success of the action is calculated the capture rate, determined by the timing and the type of the Poke ball used.

The monetization model implemented in this project is free-to-play, but the game also supports in-app purchases for additional in-game items, where players can purchase additional Poke balls or other in-game items (like incense – attracts Pokémon to you as you move for 30 minutes, lure models – attracts Pokémon to a fixed location, lucky eggs – double XP points gained for 30 minutes).

3.3 Proposed Metrics and Data Structures

In order to define and optimize the monetization model, we have to first define a set of metrics that can be used in the case of Pokémon GO. The data system should contain at least the following elements:

- user and initial setup: unique user ID (numeric), account created date (date and time)
- game version (numeric/string) and GPS related data: current location (GPS coordinates - string), city

(string), country (string), time zone (string)

- responses to pop-ups: allow access to location (true/false) and camera (true/false) to be able to play the game, allow notifications (true/false), accept terms of service (true/false))
- birth date (date)
- account type (string): sign-up with Google or just access the "Pokémon trainer club"
- avatar selection (current avatar configuration, may be string) and updates (new avatar configuration, updated string)
- nickname (string)
- actions in game: like avatar update, Pokémon catch, check out a Poke stop or visit a gym, change location, check out menu (Pokémon, items, Poke dex, Shop, Settings, Tips, etc) and actions done in sections of the menu, like change settings, check out information, buy a package from the shop, level up, and sessions.

To the data described above, generated from triggers implemented in game code, we can add external information, such as data imported from Facebook (posts, photos articles about Pokémon GO). If validated and successfully used, Facebook information can be very important to test and measure virality and general feelings and trends about the game. Very important are the articles from the media related to the game, Google searches and all the emails and messages addressed to the development team via customer support or the social networks. All this data is very important, even though we cannot do a direct correspondence between data and players IDs in the database. Using the information gathered from this channels, the company can improve the current game build, and update it with fixes mentioned by users, or even implement new features that seem to be much needed; and all this based on players' feedback and news related to the game.

For data storage, most probably various types of NoSql solutions could be used, like document storage, but it seems more suitable to use columnar based storage. Considering that fact that new data will be added continuously for every player, probably best choice is to process streams of data. This can be done by processing "time windows" of data, to get more velocity in the process, or use the similar approach of managing data at rest using Hadoop.

3.4 Architecture for Data Processing

After defining the data structure and a proposal of metrics that could be used for game analytics, we can define the process as seen in **Fig.1**. In the previous section, we defined the main sources as game data (defined by triggers implemented directly in game), and the data that can be imported from other external sources, such as Facebook, media and news.

Fig.1. Game analysis with Big data

After gathering data from the available sources, the process continues with Big data management. This part represents the processing and preparation of data, and in the case of the metrics described above, this can be made by splitting some columns,

in order to make some information more accessible and clear. Another processing step can be adding new data calculated as the age of players, using their birth date known in game and the current date. The data discovery is made by exploring the available data and understand it better before the data preparation step. During this part the data is prepared to be used in analysis, which can help us make predictions of successful monetization mix. The last step of the process is visualization of the data and the results of predictions algorithms.

3.5 Data Analysis and Prediction Algorithms

In the data analysis process, considering the data structures described above, we can use various analysis tools and prediction algorithms to make the best monetization for the game. For players' segmentation, we can use clustering or classification algorithms, such as k-means algorithms or neural networks and decision trees, to make sure we address similar players with the corresponding methods.

For monetization, prediction algorithms can be used to optimize the monetization mix according to each of the players' segmentations described above, or to estimate future sales based on the previous purchases made in game. For example, based on evolution of the player in game, the number of sessions and the current number of in-game coins, we can predict using regression, when will the player remain without any coins and action accordingly. At that point, in-game store offers and a friction area could be introduced in order to make the player need to buy more items in order to pass that point.

All these are small differences in implementation but can truly change the monetization of a product, and help the company to get the most profit out of its products and services.

4.1 Data Analysis using Pokemon Data

The data set used in this section was gathered from the social media networks, using articles and posts from the last months, created by players of Pokemon GO. The user specific data (such as age, gender, etc) was approximated using player's looks, in case this information wasn't available on their public profile.

The data set has 200 instances, with the following values and attributes, which were analysed during the pre-processing step, using WEKA:

- *player_id,* given by the order of the list;
- *name*, representing their name used on this particular social network;
- *gender*, which was setted according to each player's profile; according to pre-processing steps, data is distributed as 98 females and 102 males (see **Fig.2**):

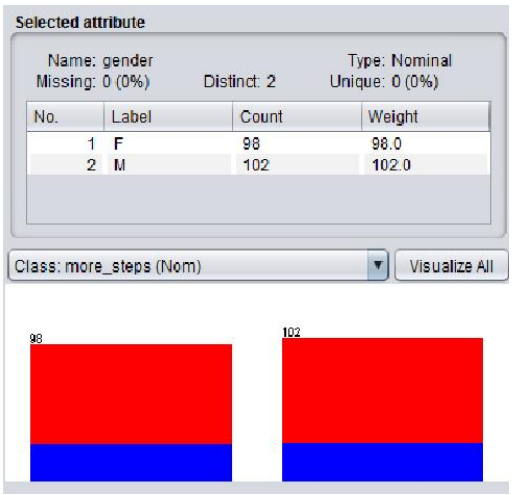

Fig.2. Gender distribution

- *age*, calculated using their birth year – when it was available on the social network, or approximated using their looks, and according to pre-processing step, the data is distributed as follows, in **Fig.3**, with users having an average age of 31.805 years:

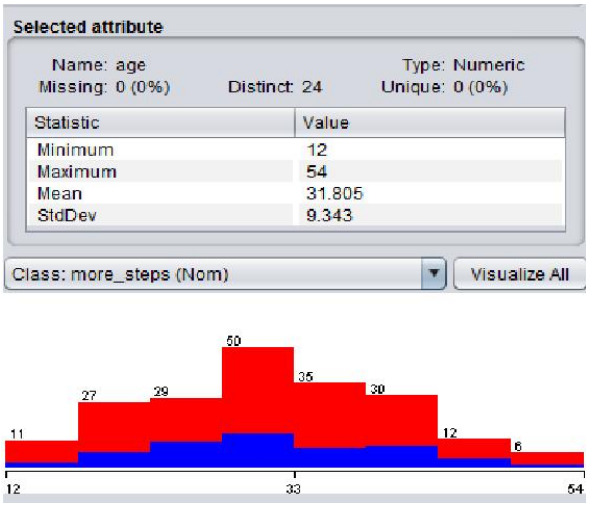

Fig.3. Age distribution

- *height*, approximated using player's looks;
- *weight*, approximated using player's looks;
- *BMI or the body mass index*, calculated field, using the values of player's height and weight, and the BMI formula:

BMI=weight/height²,

where weight was transformed in kilograms and height was transformed in meters, and according to pre-processing steps, having the following distribution, as seen in **Fig.4**:

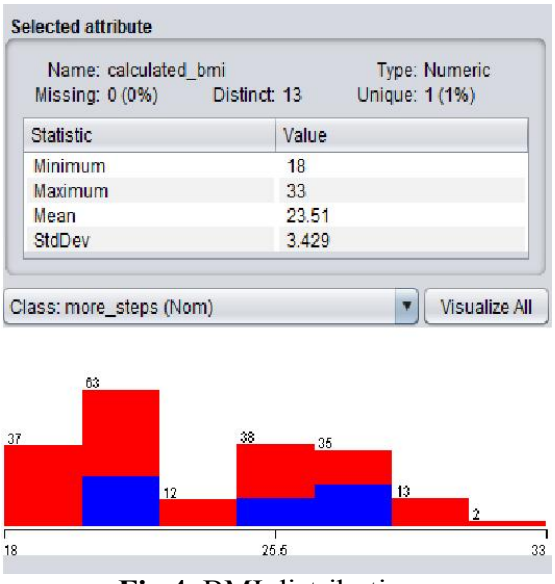

Fig.4. BMI distribution

- *days_since_playing, daily_avg_steps_before, daily_avg_steps_after, more_steps*, information specific to this case, of a game that implies to walk in order to proceed.

4.2 Naive Bayes Classifier

The Naive Bayes Classifier represents a supervised learning method, and a statistical method for solving classification problems. The classifier assumes an underlying probabilistic model and allows to capture uncertainty about the model by determining probabilities of the outcomes. This model can be used to solve diagnostic problems, and also predictive problems.

This classifier is called after Thomas Bayes (1702-1761), who proposed the Bayes Theorem, with the "naive" assumption of independence between every pair of features, as follows: given a class variable y, and a dependent feature vector x_1 through x_n, Bayes' then:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Using the naive independence assumption that

$$P(x_i | y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i | y),$$

for all i, this relationship is simplified to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

and can use Maximum A Posteriori estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class y in the training set.

The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i \mid y)$.

In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They require a small amount of training data to estimate the

necessary parameters. Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution.

Naive Bayes classifiers are mostly used for

– text classification, where the Bayesian classification is used as a probabilistic learning method;

– spam filtering- the best known use, classifier is used to identify spam e-mail;

– recommendation system - filtering unseen information and predict whether a user would like a given resource;

– online applications - using a training set of examples which reflect nice, nasty or neutral sentiments.

4.3 Classification of social media data

After pre-processing, calculating and formatting the needed variables, a Naive Bayes classifier (scheme: `weka.classifiers.bayes.NaiveBayes`) is applied using 12 attributes, in WEKA on the 200 instances of data gathered from the social network, using as test option the cross-validation with 10 folds:

```
=== Run information ===

Scheme:
weka.classifiers.bayes.NaiveBayes
Relation:     data_csv
Instances:    200
Attributes:   12
player_id
name
gender
age
height
weight
height^2
calculated_bmi
days_since_playing
daily_avg_steps_before
daily_avg_steps_after
more_steps
Test mode:    10-fold cross-validation
```

After running the model, the results are as follows:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly      Classified      Instances
175               87.5    %
Incorrectly     Classified     Instances
25                12.5    %
Kappa                          statistic
0.6951
Mean            absolute            error
0.1141
Root       mean       squared       error
0.301
Relative          absolute          error
28.5286 %
Root     relative      squared      error
67.3965 %
Total      Number      of      Instances
200

=== Detailed Accuracy By Class ===

                 TP   Rate      FP   Rate
Precision    Recall      F-Measure    MCC
ROC Area  PRC Area  Class
                 0.818    0.103    0.750
0.818      0.783         0.696      0.968
0.918     N
                 0.897    0.182    0.929
0.897      0.912         0.696      0.968
0.989     Y
Weighted Avg.    0.875    0.160    0.879
0.875      0.877         0.696      0.968
0.969

=== Confusion Matrix ===

a   b   <-- classified as
45  10 |   a = N
  15 130 |   b = Y
```

The correctly and incorrectly classified instances show the percentage of test instances that were correctly and incorrectly classified. The raw numbers are shown in the confusion matrix below, with a and b representing the class labels. The percentage of correctly classified instances is called accuracy or sample accuracy. Given this data set, the accuracy of classifier is about 87%. To get better results will have to try different classifiers or preprocess data even further. Kappa is a chance-corrected measure of agreement between the classifications and the true classes. It's calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum

possible agreement. A value greater than 0 means that the classifier is doing better than chance.The paper shall include an introduction on the current research in the papers field, original solutions, experimental results analysis, conclusions and references.

References

[1] Martin Hilbert, Internet: "Big Data for Development: A Review of Promises and Challenges. Development Policy Review.", martinhilbert.net, Retrieved 2015-10-07

[2] Mark Beyer, Internet: "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data", Gartner, Archived from the original on 10 July 2011, Retrieved 13 July 2011

[3] C. Mohan, Internet: "History Repeats Itself: Sensible and NonsenSQL Aspects of the NoSQL Hoopla", http://openproceedings.org/2013/conf/edbt/Mohan13.pdf, 2013

[4] Internet: "Amazon Goes Back to the Future With 'NoSQL' Database", WIRED, https://www.wired.com/2012/01/amazon-dynamodb/, 2012-01-19, Retrieved 2017-03-06

[5] Stephen Yen, Internet: "NoSQL is a Horseless Carriage" (PDF), NorthScale, Retrieved 2014-06-26

[6] Jeffrey Dean, Sanjay Ghemawat, Internet: "MapReduce: Simplified Data Processing on Large Clusters", https://static.googleusercontent.com/media/research.google.com/ro//archive/mapreduce-osdi04.pdf

[7] Jeff Bertolucci, Internet: "Hadoop: From Experiment To Leading Big Data Platform", Information Week, 2013, Retrieved on 14 November 2013

[8] Kurt Wagner, "How many people are actually playing Pokémon Go? Here's our best guess so far", Internet: http://www.recode.net/2016/7/13/12181614/pokemon-go-number-active-users, 2016

[9] Jeremy Liew, Internet: "29 business models for games", http://lsvp.com/2008/07/02/29-business-models-for-games/, July 2, 2008

[10] Anders Drachen, Christian Thurau, Julian Togelius, Georgios N. Yannakakis, and Christian Bauckhage, Internet: "Game Data Mining", http://julian.togelius.com/Drachen2013Game.pdf, 2013

[11] Samit Sarkar, Internet: "Pokémon Go hits 650 million downloads", http://www.polygon.com/2017/2/27/14753570/pokemon-go-downloads-650-million, 2017

Ioana Roxana STIRCU, graduated the Faculty of Mathematics and Informatics of the University of Bucharest in 2010, and gained the Master title in Cryptography and Codes Theory at the same university.