A Data Mining solution for detection of potential buyers on MMORPGs

Átila Valgueiro Malta Moreira* Danilo Rafael de Lima Cabral Gabrielle Karine Canalle
Paulo Jorge Leitão Adeodato

Universidade Federal de Pernambuco, Centro de Informática, Brazil

ABSTRACT

This research aims to increase the target group through the qualification after the acquisition, reducing the cost of the campaigns through the increase of Conversion Rate.

Keywords: MMORPG, Decision Tree, Logistic regression, CRISP-DM, Conversion rate.

1 INTRODUCTION

The main revenue channel for the games is through transactions of purchase of virtual goods. Since players are not necessarily required to pay to play, it is very important to understand quickly and accurately the quality of the users being acquired, since the profit of gaming companies is directly associated with the CPI(Cost per Install) and the LTV(Life Time Value).

This article aims to develop a predictive approach to identify possible paying players for this way improve the billing of games that follow the Freemium model. To that end, difference between non-payer player and payer player behaviour need to be discovered in a short period of time after the sign up. This tool will be an invaluable ally to the game marketing.

This research follows the methodology CRISP-DM[1], for managing data mining processes, that can be used by professionals with different levels of knowledge. This model split a data mining project in six steps. There are: business understanding, data understanding, data preparation, modeling, evaluation and deployment.

For organizational reasons this article has been split up into sessions based on CRISP-DM followed by conclusions and potentials improvements.

2 BUSINESS UNDERSTANDING

This initial phase focuses on understanding the project goals and requirements from a business perspective. This step is responsible by convert this knowledge into a data mining problem definition.

In figure 1, one can note the decision-making process to payer player prediction. In this diagram can be noted that the flow of decision starts on player's sign up. From that moment, player's logs are stored during 3 days. If the player leave the game or convert to payer before third day, this player is removed of the decision-making process, due is out of the project scope. Else a decision is taken and the player is classified as possible payer or non-possible payer.

This research had defined objective as a binary decision and stipulates the follow KPI(Key Performance Indicator) [2]:

- CR (Conversion Rate): try to reach 10% of CR using predictive tools. The game is currently reaches 6%.

*e-mail: atila@bighutgames.com

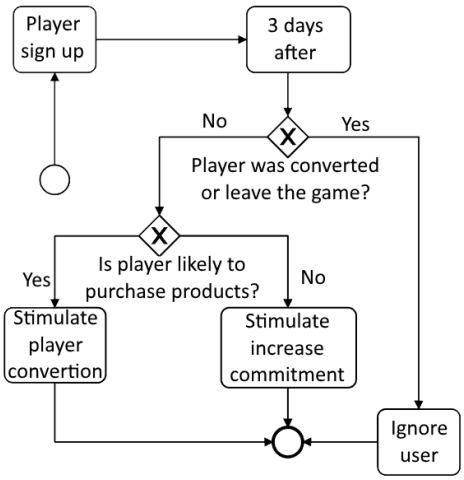

Figure 1: Decision-making process.

3 DATA UNDERSTANDING

Since the data mining problem was defined the data understand step begins, at this step the raw data are collected. A qualitative study is done to understand the meaning and detect potential quality issues with the data.

This step was a great challenge due the massive amount of data structured and unstructured divided on three servers. In total, there are about 13 TB of raw data scattered on two databases and one log files system. It was necessary a joint analysis with the GM(Game Master)[1] and the live operation team. For a better understanding some samples of the three cases are presented bellow at Table I.

Table 1: Log samples.

Log from	Log
DeltaDna Samples	C54B5,clientDevice,2017-09-09 08:22:47.937 C54B5,gameStarted,2017-09-09 08:22:48.463 48D05,itemActioned,2017-09-09 08:38:49.148
Native Samples	09/09/17,00:00:03,IP,LOGIN,tra,npc 0 09/09/17,00:00:05,IP,LOGIN,get 0867 09/09/17,00:00:05,IP,LOGIN,err,hitposition
GA Samples	1504909344,Game:Play,design,BR,57000 1504911376,Buy:NP:Andarilho,design,BR,90.0 1504912547,Buy:NP:Emblema,design,BR,75.0

The next steps for this work were executed a data selection, data cleansing and data transformation. Distinct treatments were utilized for each origin with the objective of generate a communication protocol between the three origins by a unique key. Each line in the file

[1]The responsible professional for guiding the game, being the biggest expert of the domain.

system contains the follow information: *userId*, *logTime*, *logType*, *sessionId*, *platform* e *bruteLog*.

The database was not structured and the research problem was not formally characterized. That way, the majority of variables detailed above represent a scientific contribution to a data-mart creation focused on detect payer players. This data view, mainly in the decomposition of the time series, embed the domain expert knowledge. This transformations are part of an approach called Domain-Driven Data Mining [3]. This approach has been successfully applied on data mining practices[4].

4 Data Preparation

After running the user scope filter of this research, the final sample consisted of 49.556 logs of players. In the final sample only 6% of the players are the target class(payers). This final sample was divided in two parts: training and test. The training sample contains 2/3 of final sample and the test sample contains about 1/3. The partitioning were randomly allocated, stratified by target class, keeping 6% of target class on both samples. The model performance was validated on test sample.

Since keeping unnecessary attributes implies in more noise to the learn process [5], the attribute selection is an important phase of data preparation. A logistic regression[6] was executed to identify most important attributes according beta coefficient and p-value.

Table 2: Final attributes after logistic regression.

Attribute	Coefficient	p-value
duration	-6.87	0.00
message_d1	-0.18	0.00
sessions_d0	-0.14	0.00
sessions_d1	-0.11	0.00
missionComplete_d1	-0.09	0.00
message_d0	-0.08	0.00
missionComplete_d0	-0.07	0.01
crash_d0	-0.05	0.00
crash_d1	-0.05	0.00
useItem_d1	-0.04	0.02
missionComplete	0.08	0.01
transaction	0.11	0.00
channel	0.18	0.00
daysPlayeds	0.20	0.00
useitem	0.24	0.00
message	0.25	0.00
sessions	0.33	0.00
mail	0.38	0.00
duration_d1	0.60	0.00
duration_d0	2.61	0.00
Constant	-1.96	

The table 2 presents the coefficient (β) and the *p-values* for each attribute considered as significant to this research (attributes with *p-value* > 0.05 was removed from the sample).

Having selected all the more significant attributes, some modifications were implemented to embed more expert knowledge:

- *channel* (acquisition channel): the domain expert identified that googleAds channel bring to the game players with more disposition to conversion. Consequently, a category variable was created which category 1 was assigned to players that was brought by googleAds. Player that was from another acquisition channel were assigned as -1 and player without acquisition channel log were assigned as 0.
- *mail*: The mail domain was extracted from the raw mail and then a category based on the sign-up complexity on domain was created. Paid domains and more restrictive domains was assigned as 1. Volatile mail domains were assigned as -1 and all players without mail log was assigned as 0.

Subsequently a distinction on data treatment was realized according the modeling technique applied. To logistic regression[6], all variables (except *duration*, *duration_D*0 and *duration_D*1) were adjusted according *z-score* [7]. To the Decision Tree [8] and Rule Induction [9] this variables were transformed in binomials. The variables *duration*, *duration_D0* and *duration_D1* were discretized by quartiles with values 1,2,3 and 4.

5 Modeling

The data modeling phase involve a selection, application and configuration of algorithms that extract knowledge of sample. Logistic regression[6], decision tree[8] and rule induction[9] were executed to extract knowledge on this research.

Our first algorithm for extraction was logistic regression [6]. This model was applied aiming to classify the data. Aiming that goal, the logistic regression was training with samples of the training data set and it was evaluated with the test data set. In this set, we generated a propensity score of the potential paying users. According to the business, we defined the propensity point, where 10% of the best ranked by the algorithms of learning were classified as paying users(target class 1)

The figures 2 and 3, respectively, the ROC curve and the KS curve generated by the logistic regression. This model got a $0,734$ of AUC_ROC and $0,333$ of MAX_KS2.

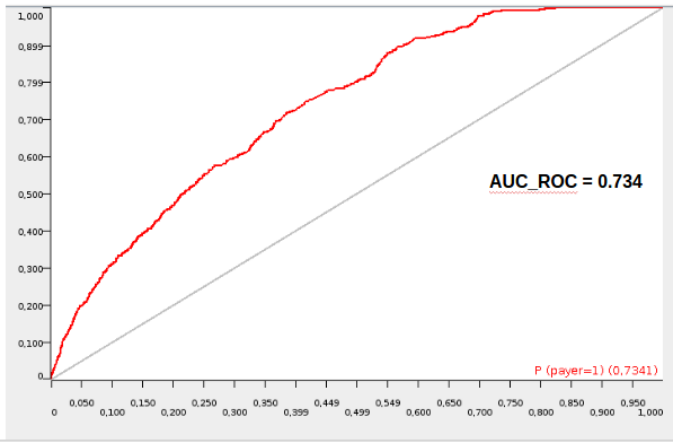

Figure 2: Logistic regression ROC curve.

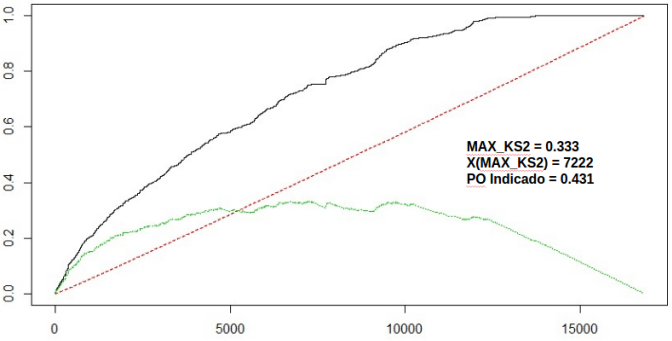

Figure 3: Logistic regression KS curve.

The confusion matrix of the logistic regression to the operation point of 10% is presented on table 3. From these data, we can con-

clude true-positive has 30% of accuracy and false-positive has 91% of accuracy to the operation point defined on business understanding. The general accuracy of this system is 88.499%.

Table 3: Logistic regression confusion matrix.

Class \Prediction	0	1
0	**14724 (91%)**	1496
1	442	**189 (30%)**

The second adopted classifier to extract knowledge was the decision tree *C*4.5 [8]. At first time this algorithm was used to generate rules on unbalanced training sample (without oversampling) to validate the decision model. The decision tree ROC curve and the KS curve are represented on figures 4 and 5. The AUC_ROC has value equals = 0.709 and the MAX_KS2 has value equals = 0.355.

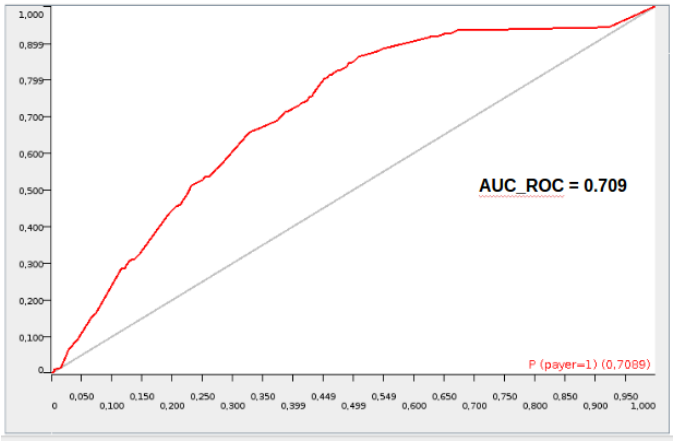

Figure 4: Decision tree ROC curve *C*4.5.

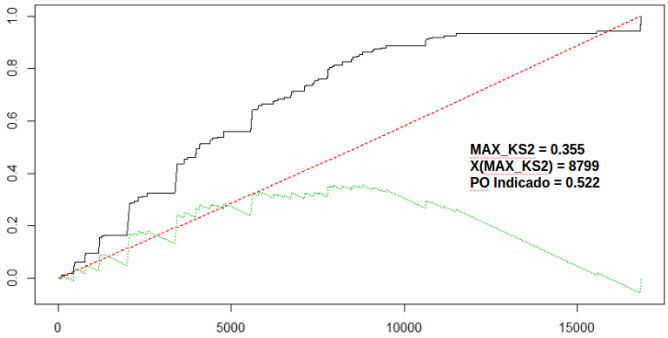

Figure 5: Decision tree KS curve *C*4.5.

The rights and wrongs of the decision tree *C*4.5 – on negatives and positives classes – are represented on confusion matrix on table 4 (with operation point equals 10%). Based upon this table we can note a 16% of accuracy to true positive and 90% of accuracy to true negative and a general accuracy of 87.49%.

As previously mentioned on this paper, the decision tree *C*4.5 was used on this work mainly to create classification rules able to provide a better insight of the decision model. The table 5 present two rules generated by decision tree.

The first rule refer to target class, with a support of 0.015, that represent 25% of target class elements on training sample and confidence equals 0,646. The second rule refer to no target class and has a support of 0.224 and confidence equals 0.974.

Table 4: Decision tree *C*4.5 confusion matrix.

Class \Prediction	0	1
0	**14639 (90%)**	1581
1	527	**104 (16%)**

Table 5: Sample of rules generated by decision tree *C*4.5

Rule	Corrects	Total	Support	Confidence
duration = 2 AND channel = -1 THEN 1 (2 attributes)	317	490	0.015	0.646
duration = 3 AND channel = -1 THEN 0 (2 attributes)	7121	7313	0.224	0.974

These two rules shown a great influence of the attributes *duration* and *channel* on decision model. These influences were discussed on session 6. The last technique applied on this research was the rule induction. The algorithm used to induct rules was the *JRip* [9]. The table 6 present a list of rules extracted of the algorithm output over unbalanced training sample.

Table 6: Rule induction main rules

Rules	Support	Confidence
duration = 1 AND daysPlayed >= 3 AND duration_d0 >= 2 AND mail = 1 AND duration_d1 <= 2 THEN 1 (5 attributes)	0.000	1
duration = 2 AND useItem = 0 AND message_d1 = 0 AND mail >= 1 THEN 1 (4 attributes)	0.002	0.759
duration = 2 AND daysPlayed = 3 AND duration_d1 = 3 AND mail = 1 THEN 1 (4 attributes)	0.005	0.663

The first two rules on the table 6 present confidence 1 but has an insignificant support to the analyzed set. The other rules present support between 0.002 and 0.007 being more significant to decision model. In addition, we can note that all rules reinforce the *duration* relevance. Another relevant attributes are *daysPlayed* and *channel*. These analysis were discussed on session 6.

6 ANALYSIS

In this section the model technique results were evaluated according to precision on solving the problem. Thus, it analyses the results that more converge and diverge of the expert knowledge.

Among the results obtained on logistic regression, it is important to analyze the follow items:

- useItem(+0.247), transaction(+0.11) and message(+0.257): these logs are typical actions of engaged players and deep knowledge of the game.
- crashD0(-0.055) and crashD1(-0.051): The expert hoped that crash and conversion was directly proportional since players that explore more the game are more vulnerable to crashes.
- mail(+0.383) and channel(+0.18): this values were expected.
- duration(-6,873): The expert expected a positive value to duration, but this variable presented the most negative value. After a more detailed analysis joint with the rule induction and decision tree analysis the expert concluded that this value happened due of the existence of auto-selling accounts [2].

[2]Accounts that player create to sell items in games, usually this kind of account is the secondary account of the player and remain constantly online

- sessions(+0.333) and daysPlayed(+0.204): this values were expected.

Due the great similarity between results of decision tree and rule induction the analysis of both was made in conjunction and can be viewed bellow:

- duration: Higher quartile is less prone to conversion. However, the decision tree presented third quartile is more prone to conversion. The rule induction assumes the second third quartile are more prone to conversion. The expert already expected that the fourth quartile(higher duration) players were not prone to conversion due auto-selling accounts, but decision tree and rule induction pointed the first quartile is not prone to conversion too, after some analysis again the expert arrive in the following conclusion: the first quartile there are no engaged player and chest accounts[3].
- channel: Both logistic regression and decision tree presented that games on channel -1 and duration on second quartile are prone to conversion. This rule evidence a fault on expert analysis of channel, after a more detailed analysis the expert noted facebookAds should not be assigned as -1 due the look-a-like facebook algorithm[4].

7 Final Remarks

This paper presents, beyond a prediction system for potential paying users with a good accuracy, a big amount of tips to improve the understanding of the company business.

For future work, we have listed below our new goals for this research: Improve the category channel to players brought by FacebookAds; Exclude all the auto-selling and chest account; and, Optimize the analysis time from three to two days, since the logistic regression show us that there is no meaningful variable in the third day.

Acknowledgments

The authors would like to thank firstly the Centro de Informática - UFPE for his academic support. Also, we would like to thank Raid Hut, Big Hut Games and Manifesto Games for allow us to use their data in this research.

[3] Accounts that players create to store few used items.

[4] Lookalike Audience is a way to reach new people who are likely to be interested in your business because they're similar to your best existing customers.

References

[1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0 step-by-step data mining guide," 2000.

[2] B. Marr, *Key Performance Indicators (KPI): The 75 measures every manager needs to know*. Pearson UK, 2012.

[3] L. Cao, "Domain driven data mining (d3m)," in *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*. IEEE, 2008, pp. 74–76.

[4] P. J. Adeodato, P. L. Braga, A. L. Arnaud, G. C. Vasconcelos, F. Guedes, H. B. Menezes, and G. O. Limeira, "Domain driven data mining for unavailability estimation of electrical power grids," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2010, pp. 357–366.

[5] B. Ratner, "Variable selection methods in regression: Ignorable problem, outing notable solution," *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 18, no. 1, pp. 65–75, Mar 2010. [Online]. Available: https://doi.org/10.1057/jt.2009.26

[6] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, no. 1, pp. 3–14, 2002.

[7] R. Larson and B. Farber, *Estatística Aplicada*, 4th ed. São Paulo, Brasil: Pearson Prentice Hall, 2010.

[8] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[9] W. W. Cohen, "Fast effective rule induction," in *Twelfth International Conference on Machine Learning*. Morgan Kaufmann, 1995, pp. 115–123.