An index to measure the engagement of Globoplay users – Globo's OTT Platform

Felipe Parpinelli Constâncio
Gabriel Sodré Belém
Karla Klahold de Souza Biscaro

CITE THIS ARTICLE

Constâncio, Felipe Parpinelli; Belém, Gabriel Sodré and Biscaro, Karla Klahold de Souza;2019. An index to measure the engagement of Globoplay users – Globo's OTT Platform. SET INTERNATIONAL JOURNAL OF BROADCAST ENGINEERING.
ISSN Print: 2446-9246 ISSN Online: 2446-9432. doi: 10.18580/setijbe.2019.4. Web Link: http://dx.doi.or- g/10.18580/setijbe.2019.4

An index to measure the engagement of Globoplay users – Globo's OTT Platform

Felipe Parpinelli Constâncio, Gabriel Sodré Belém and Karla Klahold de Souza Biscaro

Abstract — **Engagement is a very abstract and very important metric to measure the evolution of the product and understand the users. In this study, we will show how, through data science techniques our team has created an index of engagement, what we call IEN, for our users base crossing data and providing different weights for playtime, diversity of watched programs and frequency of use. And from this index, we generate segments of users. We will also show how we create interventions for these user segments, allowing you to use more complex business rules in the product.**

***Index Terms* — Globoplay, VOD, OTT, Engagement index**

I. INTRODUCTION

GLOBO is the largest media group in Brazil and the largest Latin America content producer. We reach 100 million people daily on our properties.

Globoplay is the group's OTT platform, with a wide variety of AVOD, SVOD and Live Streaming content, including international tv series, movies, reality shows, telenovelas, local news and so on.

Understanding and measuring the relationship between the engagement of our users with their advertising and subscription lifetime value including their propensity to churn is a rather complex task due to this wide variety of content, seasonality, forms of payment and consumption, such as VOD and catch-up.

We had a number of challenges in this task of understanding and pursuing a single user engagement metric, the first one was knowing which variables to choose and which best define user engagement. The advantage of dealing with a single index to accurately describe engagement is that we avoid dealing with different metrics at the same time to do this kind of assessment, as well as makes easier to rank or categorize the users.

Globoplay has three types of users: subscribers, free and anonymous. We understand that for this work, the ideal would be to split and create a specific index for each of the types of users. In this study, we focus on subscribers and free users.

After some testing with possible variables, we came up with three product metrics for the subscriber: playtime in hours, user frequency on the platform (user sessions), and diversity of watched titles. And for the free user, we use the combination of three other variables: number of videos watched, playtime in hours and frequency (distinct days watching videos). To understand the importance and define the weights of each of these variables, we use the PCA technique, Principal Component Analysis.

II. METHODOLOGY

Principal component analysis (PCA) is a multivariate technique which objective is to reduce dimensionality, each component is a linear combination of the features of interest. In our case, we want to develop an index that translates the engagement of the users, meaning we want to reduce a set of information about video consumption into one metric.

The first principal component is a linear combination of the features with maximal variance, so considering the first principal component as the engagement index is very interesting because maximize information content considering the variance of the features as well as the correlation among them. Consider the first principal component, **Y**, as a weighted average of p features ($\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_p}$) that has the largest variance, in this case the variance of Y is:

$$Var(\boldsymbol{Y}) = \boldsymbol{a'Sa}$$

Where a is the vector of optimal weights and **S** is the variance-covariance matrix of **X** (the set of p features).

The first principal component maximizes the Var(**Y**) subject to the normalization constraint **a'a** = 1. Notice that features with larger variance are given larger weights, the same occurs to correlated variables [2].

III. ENGAGEMENT INDEX FOR SUBSCRIBER USERS

As previously mentioned, for the subscriber user case, we consider the following variables: Playtime in hours, frequency in the platform and variety of titles. For each user, we calculate weekly the sum of each of these variables. Using the PCA technique, we had 80% of the variability explained by the first principal component. From the first main component, we can obtain the weights of each of the variables used: weight frequency ($\boldsymbol{w_f}$), weight playtime ($\boldsymbol{w_p}$) and weight title diversity ($\boldsymbol{w_t}$).

Through the PCA, the following weights were determined for each variable, obtained from our development sample:

a) $\boldsymbol{w_f}$ = *0.88*

b) $\boldsymbol{w_p}$ = *0.45*

c) $\boldsymbol{w_t}$ = *0.11*

http://www.set.org.br/ijbe/ doi: 10.18580/setijbe.2019.4 Web Link: http://dx.doi.org/10.18580/setijbe.2019.4

Therefore, the Engagement Index (IEN) is calculated from the multiplication between the weights obtained by the sum of the weekly consumption of these variables by the user, from the following formula:

$$IEN = \sum_{d=1}^{7} \boldsymbol{F_d} \times \boldsymbol{w_f} + \sum_{d=1}^{7} \boldsymbol{P_d} \times \boldsymbol{w_p} + \sum_{d=1}^{7} \boldsymbol{T_d} \times \boldsymbol{w_t}$$

Where:
$\boldsymbol{P_d}$ = user playtime in hours;
$\boldsymbol{F_d}$ = user frequency;
$\boldsymbol{T_d}$ = variety of user title;
d = day;
$\boldsymbol{w_f}$ = weight frequency;
$\boldsymbol{w_p}$ = weight playtime;
$\boldsymbol{w_t}$ = weight title diversity

This generates a continuous value that expresses the engagement of each user on the platform. We then get the engagement rates corresponding to the 20th, 40th, 60th and 80th percentiles, which allows us to classify our users into 5 engagement groups as shown in table I:

Table I
Relationship between the defined classes and the engagement index

Engagement class	**Engagement index**
marathoner	$IEN \geq 9.84$
assiduous	$5.13 \leq IEN < 9.84$
regular	$2.76 \leq IEN < 5.13$
eventual	$1.31 \leq IEN < 2.76$
accidental	$IEN < 1.31$

In the table II, we show a lift analysis of each user segment:

Table II
Lift analysis between engagement classes and their consumption

Engagement class	**Hours Playtime (%)**	**Video views (%)**	**Distinct Days (%)**
marathoner	0.00	0.00	0.00
assiduous	-65.83	-59.29	-30.56
regular	-83.33	-78.57	-44.44
eventual	-92.67	-89.29	-61.11
accidental	-98.47	-92.86	-72.22

With the classes defined from the engagement index, we focus on identifying in this sample the percentage of unsubscriptions in each of these groups and seek to understand the relationship of engagement with churn, we show this correlation in the table III:

Table III
Relationship between the defined classes and Churn

Engagement class	**Churners (%)**	**Lift (%)**
marathoner	16	0
assiduous	14	-12.50
regular	19	18.75
eventual	22	37.50
accidental	29	81.25

Notice that 51% of churns in this sample were concentrated in the two lowest engagement profiles.

IV. Engagement Index for Free users

After getting a numerical index that explained subscriber engagement, the next challenge was to study the "free" user behavior. This is further subdivided into two groups, the "free logged in" and the "anonymous" ones. The first subgroup is formed by those who have registration and, consequently, have a unique identification. The second is a more complex case to study since it is not possible to observe the same individual in different accesses. Therefore, anonymous users will be a challenge for future discussions.

Although the "Free logged User Engagement Index" used a methodology similar to the "Subscriber Engagement Index", it's important to realize that they are two distinct behavior groups. While subscribing users have access to all product content, logged in free is limited to watching only excerpts and a few full contents, not having access to the main international content, for example.

In addition, both groups present their value to the product differently. While the subscriber generates revenue by paying monthly for unlimited access, the free user watches more publicity. Given the business need of this group of users, three attributes have been determined for index composition. The first of these is the amount of watching videos ($\boldsymbol{w_v}$), which is directly related to advertising consumption since on Globoplay an advertisement is displayed at the beginning of each video. Another parameter is the video playtime ($\boldsymbol{w_p}$). Also present in the subscriber's index, representing the volume of hours consumed by users. Finally, the distinct days that users consumed video on platform ($\boldsymbol{w_f}$). This time is represented by the number of different days on which the user had video consumption. Once again, the attributes are consolidated weekly, with values weighted by the PCA algorithm, implying the following weights:

a) $\boldsymbol{w_v}$ *= 0.89*

b) $\boldsymbol{w_p}$ *= 0.39*

c) $\boldsymbol{w_f}$ *= 0.23*

Therefore, the engagement index (IEN) for free users is formed by the sum of weight products by the sum of weekly consumption of each variable, resulting in the following formula:

$$IEN = \sum_{d=1}^{7} \boldsymbol{V_d} \times \boldsymbol{w_v} + \sum_{d=1}^{7} \boldsymbol{P_d} \times \boldsymbol{w_p} + \sum_{d=1}^{7} \boldsymbol{F_d} \times \boldsymbol{w_f}$$

Where:

$\boldsymbol{V_d}$ = Total amount of videos watched by user;
$\boldsymbol{P_d}$ = user playtime in hours;
$\boldsymbol{F_d}$ = user frequency;
d = day;
$\boldsymbol{w_v}$ = weight of amount videos watched;
$\boldsymbol{w_p}$ = weight user playtime;
$\boldsymbol{w_f}$ = weight user frequency

The principal component analysis method was reliable when it reached 79.25% of explanatory variability. The boxplot shown in Figure 1 illustrates the variability of the attributes.

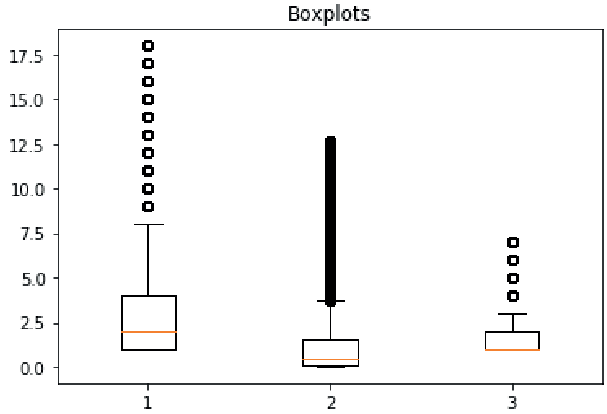

Figure 1 - Boxplot with the used variables

For the development of this index, was considered the period of May 2019, a period that is less impacted by anomalies of consumption and seasonality. And in order to use values that most closely match the actual behavior of the user group, the percentile method [4] was used to remove data with values above the 95th percentile for the "number of videos" and "video playtime" attribute, since that these data have low representativeness in the sample. As the frequency has low variability in relation to the other parameters, it was not necessary to apply any method to remove outliers.

The percentile method was also applied to define the engagement profiles, where the first quintile determined the upper limit value for the least engaged group and the fourth quintile the lower limit for the most engaged users. The profiles were defined as shown in Table IV, from the most engaged to the least engaged users:

Table IV
Relationship between the defined classes and the engagement index for free users

Engagement class	**Engagement index**
marathoner	IEN ≥ 5.74
assiduous	2.96 ≤ IEN > 5.74
regular	1.74 ≤ IEN > 2.96
eventual	1.16 ≤ IEN > 1.74
accidental	IEN < 1.16

The first analysis considered the consumption behavior of each profile. Calculating the average value of each attribute for the five groups, we bring in table V a lift analysis of each user segment:

Table V
Lift analysis between engagement classes and their consumption

Engagement class	**Video Playtime (%)**	**Video views (%)**	**Frequency (%)**
marathoner	0.00	0.00	0.00
assiduous	-60.74	-59.79	-43.40
regular	-80.37	-77.76	-60.69
eventual	-87.80	-88.48	-68.55
accidental	-99.20	-88.48	-68.55

Notice that accidental and eventual users differ only in the amount of "video playtime".

The marathoner group has a subscription propensity that is 54% higher than the accidental group as illustrates Table V. It noticed that 47% of sales are concentrated in the two most engaged profiles. The following table illustrates how free user engagement correlates directly with the propensity to subscribe.
The Table VI illustrates how free user engagement correlates directly with the propensity to subscribe.

Table VI
Lift analysis between engagement classes and their consumption

Engagement class	**Lift of the propensity to subscribe (%)**
accidental	0
eventual	+ 5.6
regular	+ 14.8
assiduous	+ 31.5
marathoner	+ 57.4

V. USE CASE

Due to the high seasonality and success of some TV shows we sometimes face a high churn rate after the end of this content.

Recently the telenovela “A Dona do Pedaço” is very popular among the Globoplay subscribers and it is responsible for a representative amount of video playtime in the platform. In order to avoid churn after the end of this content, we clustered the subscribers into groups of higher and lower risk of churn and described them in terms of consumption and demographic information, for example what are they favorite content and etc.

The audience selected for this case study were consumers of “A Dona Do Pedaço” from August 26, 2019, to October 31, 2019. We used a clustering technique called K-means [1], a simple and very popular unsupervised algorithm. To do that we used 4 variables: engagement index (IEN), engagement index (IEN) disregarding the consumption of “A Dona do Pedaço”, subscription time and distinct days of consumption of video. Through the elbow method [3], we identified that we can discriminate our sample into 4 groups, as shown in the figure 2:

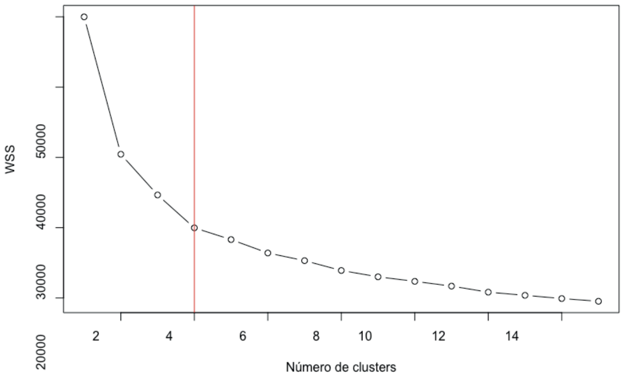

Figure 2 - The elbow method

The groups were named as: A, B, C and D and are presented in Table VII:

Table VII
Distribution of users by clusters

Cluster	**A**	**B**	**C**	**D**
Usuários (%)	14,47%	31,94%	38,85%	14,74%

To facilitate the understanding of segmentation analysis we use the PCA, this time for the purpose of reducing dimensionality. The graph in the Figure 3, shows the groups by principal components 1 and 2, which together add up to 79% of data variability. Principal component 1 (Dim1) describes user engagement and principal component 2 (Dim2) is represented by the subscription time:

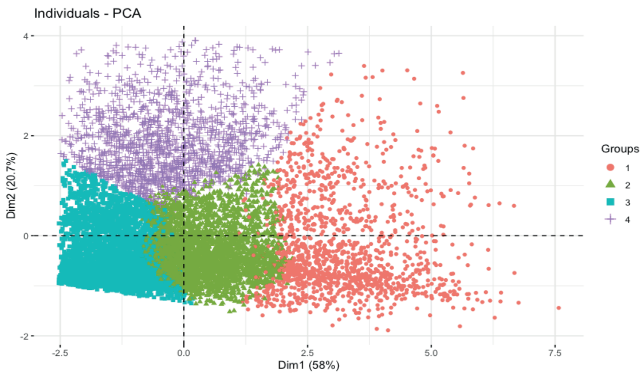

Figure 3 - dimensionality reduction

$$Dim1 = 0.54 \times \boldsymbol{e_1} + 0.53 \times \boldsymbol{e_2} + \boldsymbol{d} \times 0.47 + \boldsymbol{s} \times 0.02$$

$$Dim2 = -0.08 \times \boldsymbol{e_1} + (-0.12) \times \boldsymbol{e_2} + \boldsymbol{d} \times 0.21 + \boldsymbol{s} \times 0.97$$

Where:

$\mathbf{e_1}$ = IEN;
$\mathbf{e_2}$ = IEN disregarding “A dona do pedaço”;
d = distinct days of video consuption;
s = subscription time;

From the four clusters generated, we observe the following behaviors:

1. Cluster **C** presents the highest risk group, because has a lower IEN and a shorter subscription time if compared to the others.

2.Cluster **B** has a high IEN, but not so old subscribers.

3. Clusters **A** and **D** were considered lower risk groups, where in cluster **A** there is a high IEN and in **D** not having such a high engagement, they are old subscribers, which greatly reduces the likelihood of churn

Table VIII
Relation of variables means with clusters

Cluster	**A**	**B**	**C**	**D**
Mean of iEN	56.4	25.3	9.8	17.3
Mean of iEN without “A dona do pedaço”	48.4	19.7	7.3	12.6
Mean of Subscribe Days	254.3	148.1	143.0	772.8
Mean of Distinct Days	25.7	21.0	8.5	17.1

As a result of this work, different types of A / B testing were created, enabling the Globoplay business team to create push notifications, email marketing and various product-specific approaches within these user groups.

VI. Conclusion

The engagement index seems to be a reliable metric to describe the engagement of Globoplay users, as was shown is this paper we could validate the abstract concept we wanted to measure with this index.

The engagement index (IEN) increases as the video playtime, diversity of titles, quantity of videos and frequency increases. Furthermore, the engagement index is correlated to the inclination of being a subscriber and to canceling the subscription as well. And it is an efficient method to rank and categorize our users, which is very useful in many ways.

REFERENCES

[1] HASTIE, TREVOR, TREVOR HASTIE, ROBERT TIBSHIRANI, AND J. H. FRIEDMAN. 2001. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

[2] BISWAS, BASUDEB & CALIENDO, FRANK. (2004). A Multivariate Analysis and Extension of the Human Development Index. Utah State University, Department of Economics, Working Papers.

[3] GARETH JAMES, DANIELA WITTEN, TREVOR HASTIE, ROBERT TIBSHIRANI. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.

[4] BELÉM, GABRIEL; SILVA, RAYSSA; PERES, RODRIGO (Coord.). Utilização De Kernel E Percentis Para Identificar Outliers E Observações Pouco Representativas. Rio de Janeiro, 2018. 58 p. Trabalho de Conclusão de Curso (Engenharia Eletrônica) - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (cefet/rj) , 2018.

Felipe Parpinelli Constâncio was born in Rio de Janeiro, Brazil, in 1989. Senior Software Developer working in Globoplay at Globo.com. Received the B.S. in Information Systems from Pontifical Catholic University of Rio de Janeiro (PUC-Rio), in 2013. Currently is MSc. degree candidate in Computer Science with emphasis in Data Science at the same university.

Gabriel Sodré Belém was born in Rio de Janeiro, Brazil, in 1994. He received the B.S. degree in Electronic Engineering from Federal Center for Technological Education Celso Suckow da Fonseca (CEFET), Rio de Janeiro, in 2019. He joined Globo.com in 2018, as an intern, and currently works as a Data Engineer at Globoplay.

Karla Klahold de Souza Biscaro graduated in Statistics by the Federal University of São Carlos (UFSCar), São Carlos – SP. She works as a Data Scientist at Globo.com

Received in 2019-11-12 | Approved in 2019-11-28

http://www.set.org.br/ijbe/ doi: 10.18580/setijbe.2019.4 Web Link: http://dx.doi.org/10.18580/setijbe.2019.4