Regression Tree Role for Interpret Monetizing of Game Live Streaming

Erlin Windia Ambarsari¹, Reza Avrizal¹, Erlando Doni Sirait¹, Saruni Dwiasnati² and Robbi Rahim³,*

¹Informatics Department, Universitas Indraprasta PGRI
²Informatics Department, Universitas Mercu Buana
³Management Department, Sekolah Tinggi Ilmu Manajemen Sukma

*usurobbi85@zoho.com

Abstract. Game Live Streaming is streamers activity who play games with a purpose to monetizing. The running game recorded into a streaming statistic. Therefore, streamers can monitor it. The company of Game Live Streaming evaluated whether they are eligible for compensation to make it easier to interpret monetizing, a decision tree is needed. However, the dataset obtained from statistical streaming is numeral; the calculation uses Standard Deviation Reduction for construct Regression Tree which the concept is similar to ID3 Algorithm. The Regression stop constructs Tree when the CV is zero, or the number of instances is the least. Also, the mean reached dataset homogeneously.

1. Introduction
The live broadcast is broadcasting in real time, and it can reach an audience at the same time without delay. One of the Live shows has been carried out by Television stations since the advent of the cathode-ray tube, which the functions to broadcast sports tournament, entertainment, and news. Also, significant internet network deployment has influenced viewers to watch live broadcasts extensively through websites and Android platform, as known as live streaming.

Live streaming become children and young adults favorite. Therefore, they have the opportunity to become a creator, a presenter, and noticed by viewers. However, live streaming is uncensored and unedited. Thus, each of streamers [1] has responsibility for their content, although their actions caused controversy for the community. Nevertheless, On the feasible side from live streaming is that brands endorse increased for marketing [2].

One of the most prominent endorse brands is electronic sports (eSports) live streaming. eSports is a competitive digital game such as real-time strategy (RTS), first-person shooters, multiplayer online battle arena (MOBA) and arcade-style fighting games; which one plays is professional gamers (pro-gamers) [3]. The tournament is similar to a traditional sport, which players have intense training to improve their skill. Also, they got the champion prize from sponsorships and advertising on their live streams as their income. Although pro-gamers does a live stream on their tournament, in general, they still monetize by attracting viewers a daily using game live streaming platform. In Indonesia, a popular
platform using by gamers is Cube TV which not only pro-gamer, the amateur plays the game to get paid from monetizing which follow protocol. The protocol for monetizing has considered such as duration playing games, several viewers, reward, and violation; which the policy determined by the company.

The data from game live streaming has recorded in real time every day; streamers could monitoring their record on Streaming Statistics. Their streaming statistics had evaluated by the company whether they deserve compensation every month — their overall Live stream report sent to the admin or agent who is in charge.

In live streaming case, various conditions must fulfill to monetizing by streamers, for the example: duration games which they play are three hours for a day, they have 500 followers who give reward to them, and no verbal violence. Therefore, the company gave a compensation category as a target. If they fulfill a goal more from limit specified, they get paid the same. However, when they could not achieve it, they still get paid base on condition determined. Unfortunately, the company did not explain how the flow of compensation is, especially for streamer who does not achieve the requirement.

According to the problem case about streamers compensation flow, we interpret monetizing of game live streaming which we obtain the dataset using Decision Tree. The dataset that we examine as the example is Cube TV’s dataset. Several methods of Decision Tree to solve the problem, such as ID3, CART, and C4.5. ID3 is a classification algorithm using Information Gain to select the best attribute, which proposed by J. Ross in 1979 [4]. CART handle a classification and regression tree; if the variable has a category class, CART constructs classification tree based on the binary attributes splitting. Also, CART builds regression tree when the variable was continuous or numeric. C4.5 improved of ID3 algorithm which used for the statistical classifier. Similar to ID3, C4.5 uses information gain for splitting criteria [5].

Live stream dataset is numerical data obtained from the streaming statistic. The method that we use is Regression Tree which considered the streamer’s compensation from the company. The concept which applies to this case is ID3. However, ID3 is a classification method that handles category data which use information gain. Therefore, the solve problem to construct a Decision Tree for numeric data on the streaming statistic is to use the coefficient of variation [6]. It is the ratio of standard deviation and the mean (average); the standard deviation expressed units as the variable under study that enable to interpret [7] monetizing game live streaming.

2. Methodology

Decision Tree is the method to classify homogeneous instances dataset. The way the Decision Tree work has split a dataset into the smallest subsets for construct decision nodes and leaf nodes. Also, at the same time, an associated Decision Tree is incrementally developed [8]. As mention before, the Decision Tree method for this case use is ID3 algorithm which the information gain has to replace Standard Deviation Reduction, by a decrease standard deviation after a dataset split on an attribute. Constructing a Decision Tree is to discover an attribute that returns the highest standard deviation reduction.

Standard deviation is useful to compute a numerical sample. If the example is entirely homogeneous, the standard deviation has zero value. The formula of standard deviation ($\sigma$) as follows [9]:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (u_i - \mu)^2}$$  \hspace{1cm} (1)

Standard deviation is functionated to measure spread data while the mean ($\mu$) used to determined central tendency. Standard deviation is profoundly affected when the mean gives a poor measuring.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} u_i$$  \hspace{1cm} (2)

Therefore, the order to gain regression tree for the decision does as follows:
a. Calculate the standard deviation of the target as a class (equation 1). The target obtains from streamer’s compensation or honorarium while they do live stream from the streaming statistic.

b. Split the streaming statistic dataset into the different attributes. Separation is using a scatter plot to examine the relationship two attributes variable, namely target and predictor. The predictor predicts the class value for the new pattern. It able to determine the root node and sub-node (decision node and leaf node).

c. Calculate the standard deviation for each branch. Also, the standard deviation reduction (SDR) calculated from subtracting the standard deviation for two attributes (target and predictor) and the standard deviation before the split (step number 1).

\[ SDR = \sigma - \sigma(T, P) \]  

where

\[ \sigma(T, P) = \left( \frac{n_1}{n} \times \sigma(P_1) \right) + \left( \frac{n_2}{n} \times \sigma(P_2) \right) \]

The standard deviation for two attributes to be done with separate instances based on binary attribute splitting which total number of split \((n_1\) and \(n_2\)) divided the full amount instances before the split \((n)\), and it multiple by standard deviation after the split \((\sigma(P_1)\) and \(\sigma(P_2)\)).

d. Select the attribute with the most substantial of standard deviation reduction for the decision node.

e. Calculate the Coefficient of Variant (CV) to stop the tree process based on a branch become smaller than the threshold (10% for example), or the number of instances \((n)\) is the least.

\[ CV = \frac{\sigma(P)}{\mu} \times 100\% \]

3. Result and Discussion

Cube TV’s streaming statistic which obtains 114 streamers instances \((n)\) of live game streaming for one month. Attributes considered such as ID HOST, APCU Awal, Rate, APCU Akhir, Followers, Jam Siaran, Hari Siaran, Energy, Pelanggaran, and Komisi. The explanation is as follows:

a. ID HOST: streamers who register on game live streaming.

b. APCU (Average Peak Concurrent Users) Awal: the highest average viewers value which not audit yet.

c. Rate: The rate of the live stream which determined by Cube TV.

d. APCU Akhir: APCU Awal multiply by Rate which determined by Cube TV.

e. Followers: viewers who follow the streamers.

f. Jam Siaran: duration of games live streaming per second.

g. Hari Siaran: streamers play games for 30 days.

h. Energy: coin which gives by followers.

i. Pelanggaran: offense, such as verbal violation, pornography, illegal ads, and game cheating.

j. Komisi: compensation obtained by streamers.
Figure 1. Cube TV Dataset

Calculate standard deviation ($\sigma$) for specifying the root node. The result is $\sigma = 6955.229709$, $\mu = 6138.815789$, and $CV = 113.2992086\%$. The CV result indicates greater value than threshold = 10%; it is necessary to split instances on the attribute. Separate instances using a scatter plot as Figure 2 with obtained two attributes (Komisi and APCU Akhir for example). The separation can divide by taking midpoint value or the value that approaches. Also, it uses the highest SDR by iterations.

Table 1. The Example Dataset of APCU Akhir to Komisi for 17 instances

| Komisi   | ID HOST | APCU AKHIR | Rate  | APCU AKHIR | Followers | Jam Sizarang(Detik) | Hari Sizarang | Energy | Pelangganan |
|----------|---------|------------|-------|------------|-----------|---------------------|---------------|--------|-------------|
| 1        | 2188.0  | 24522443.0 | 910.3767273 | 988673 | 807.694786 | 812.0 | 341179.0 | 20.0 | 1530.0 | 0.0 |
| 2        | 2184.0  | 24522445.0 | 951.3718243 | 952557 | 820.465431 | 861.0 | 275710.0 | 22.0 | 245.0 | 0.0 |
| 3        | 2182.0  | 24522446.0 | 514.3600396 | 948749 | 617.576224 | 648.0 | 289491.0 | 21.0 | 636.0 | 1.0 |
| 4        | 2183.0  | 24522447.0 | 627.1739126 | 954760 | 630.592394 | 820.0 | 256972.0 | 20.0 | 610.0 | 1.0 |
| 5        | 2180.0  | 24522448.0 | 944.7160290 | 925402 | 590.616520 | 927.0 | 252842.0 | 20.0 | 1017.0 | 1.0 |
| 6        | 1802.0  | 2425981.0  | 802.8250000 | 963967 | 580.322360 | 1013.0 | 305703.0 | 22.0 | 460.0 | 0.0 |
| 7        | 6.0     | 23744965.0 | 839.4287514 | 776459 | 496.490183 | 835.0 | 288154.0 | 21.0 | 2101.0 | 1.0 |
| 8        | 1803.0  | 24019856.0 | 463.7671904 | 988604 | 454.785516 | 773.0 | 247810.0 | 21.0 | 470.0 | 0.0 |
| 9        | 2102.0  | 24789271.0 | 444.9647619 | 992324 | 444.957830 | 767.0 | 261464.0 | 20.0 | 1917.0 | 0.0 |
| 10       | 1002.0  | 17.1932576 | 433.3500000 | 992342 | 430.923540 | 273.0 | 269483.0 | 20.0 | 510.0 | 0.0 |

Based on Figure 2, value that obtained for split is 496.4901837. Cluster the dataset by separating values below 496.4901837, and values above or equal to 496.4901837 (shown on Table 1). For the example, APCU Akhir > 496.4901837 then Komisi dataset is 31500, 28350, 28350, 28350, 21000, and 18900 which is marked by bold. Similiary to APCU Akhir <= 496.4901837. The number of dataset when calculate: APCU Akhir <= 496.4901837 is 108 instances (n), $\sigma = 5281.24096$, $\mu = 5060.41667$, and $CV = 104.3637571\%$; APCU > 496.4901837 is 6 instances (n), $\sigma = 4399.431781$, $\mu = 25550$, and $CV = 17.21891108\%$. 
Determination of root node is done by using SDR; which \( \sigma \) (Komisi, APCU Akhir) = \( ((108/114) \times 5281.24096) + ((6/114) \times 4399.431781) = 5234.82995 \) and SDR = 6955.229709 - 5234.82995 = 1720.4. The selection of attributes observed shown on the table below:

| Attributes          | Split  | \( n \) | \( \Sigma \mu \) | \( \sigma (T,P) \) | SDR     | CV   |
|--------------------|--------|--------|----------------|----------------|---------|------|
| APCU Awal          | <=500  | 105    | 5285.438       | 5205           |         |      |
|                    | >500   | 9      | 12568.63       | 17033          | 5860.43 | 1094.8 |
| Rate               | <=0.9  | 26     | 4872.5         | 4038           |         |      |
|                    | >0.9   | 88     | 7346.084       | 6759           | 6781.93 | 173.296 |
| APCU Akhir         | <=496.4901837 | 108 | 5281.241       | 5060           |         |      |
|                    | >496.4901837 | 6    | 4399.432       | 25550          | 5234.83 | 1720.4 | 104.36% |
| Followers          | <=755  | 94     | 4902.146       | 4759           |         |      |
|                    | >755   | 20     | 10571.26       | 12626          | 5896.73 | 1058.5 |
| Jam Siaran         | <=344286 | 108 | 7079.492       | 6072           |         |      |
|                    | >344286 | 6     | 3928.74        | 7350           | 6913.66 | 41.5668 |
| Hari Siaran        | <=20   | 80     | 6672.017       | 4692           |         |      |
|                    | >20    | 34     | 6398.91        | 9543           | 6590.56 | 364.666 |
| Energy             | <=1388 | 101    | 6531.506       | 5993           |         |      |
|                    | >1388  | 13     | 9556.744       | 7269           | 6876.49 | 78.7406 |
| Pelanggaran        | <=0.5  | 99     | 6924.145       | 6639           |         |      |
|                    | >0.5   | 15     | 6217.206       | 2835           | 6831.13 | 124.103 |

Based on Table 2, the highest SDR is 1720.4. Therefore, the attribute used for the root node is APCU Akhir. However, CV for APCU Akhir is not the least than 10%. It need split for decision nodes and leaf nodes which the similar steps as the root node.

Figure 3. Construct Regression Tree

Figure 4. Compensation Decision Tree

The Regression has to be done to construct Tree which the step has stopped when CV has 0, or the number instances is the least (as shown in Figure 3). The mean evidence by reached homogeneous instances, and the longest path on a tree is the best decision (Figure 4). Therefore, streamers have compensation by monetizing in normally which the average is 10500 with the condition is Hari Saran
> 18, Energy > 337, 222.2413793 < APCU Awal <= 440.5, Pelanggaran <= 0, Rate > 0.859628, and APCU Akhir <= 496.4901837.

4. Conclusion
The activity has done by the streamer as a gamer in live stream platform was recorded on the streaming statistic. Dataset of streaming statistics are numeral has challenging to examine by Decision tree Classification. Therefore, this is the Regression Role to calculate the numeral dataset by using Standard Deviation Reduction. The Regression Tree stops construction when the CV is zero, or the number of instances is the least. Also, the mean reached homogeneous instances. Therefore, streamers have compensation by monetizing in normally which the average is 10500 with the condition is Hari Saran > 18, Energy > 337, 222.2413793 < APCU Awal <= 440.5, Pelanggaran <= 0, Rate > 0.859628, and APCU Akhir <= 496.4901837.

References

[1] A. N. Wendt, “The empirical potential of Live Streaming beyond cognitive psychology,” *J. Dyn. Decis. Mak.*, vol. 3, no. 1, pp. 1–9, 2017.
[2] P. R. Todd and J. Melancon, “Gender and live-streaming: source credibility and motivation,” *J. Res. Interact. Mark.*, vol. 12, no. 1, pp. 79–93, Mar. 2018.
[3] N. Edge, “Evolution of the Gaming Experience: Live Video Streaming and the Emergence of a New Web Community,” *Elon J. Undergrad. Res. Commun.*, vol. 4, no. 2, pp. 33–39, 2013.
[4] S. Rai, P. Saini, and A. K. Jain, “Model for Prediction of Dropout Student Using ID3 Decision Tree Algorithm,” *Int. J. Adv. Res. Comput. Sci. Technol.*, vol. 2, no. 1, pp. 142–149, 2014.
[5] H. Sharma and S. Kumar, “A Survey on Decision Tree Algorithms of Classification in Data Mining,” *Int. J. Sci. Res.*, vol. 5, no. 4, pp. 2094–2097, 2016.
[6] H. Bindu K, S. Rani K, and R. Rao C, “Coefficient of Variation based Decision Tree (CvDT),” *Int. J. Innov. Technol. Creat. Eng.*, vol. 1, no. 6, pp. 1–6, 2011.
[7] J. D. Curto and J. C. Pinto, “The coefficient of variation asymptotic distribution in the case of non-iid random variables,” *J. Appl. Stat.*, vol. 36, no. 1, pp. 21–32, 2009.
[8] M. Ouahilal, M. El Mohajir, M. Chahhou, and B. E. El Mohajir, “A novel hybrid model based on Hodrick–Prescott filter and support vector regression algorithm for optimizing stock market price prediction,” *J. Big Data*, vol. 4, no. 31, pp. 1–22, 2017.
[9] M. F. Al-Saleh and A. E. Yousif, “Properties of the Standard Deviation that are Rarely Mentioned in Classrooms,” *AUSTRIAN J. Stat.*, vol. 38, pp. 193–202, 2009.