hcarceno@up.edu.ph, jamosong@up.edu.ph

Contributing Factors to Twitch Viewership

Hannah Bella C. Arceño, Jan Kristine A. Mosong, and Ara Abigail E. Ambita

University of the Philippines Visayas, Miagao, Iloilo, Philippines
ipo.upvisayas@up.edu.ph
https://www.upv.edu.ph

Abstract. Streaming platforms provided career or earning opportunities for streamers. Twitch streamers with high stream views in particular, get discovered for brand advertisements and some are recruited for professional career. However, there are only few studies that explore the factors that affect streaming views. Understanding the fundamental factors that affect stream viewership is helpful for strategies in increasing stream views, which may be of help to aspiring streamers as well. This paper focused on finding the significant contributors to Twitch stream viewership. Five (5) models were built with Linear regression, using different combinations of features from the dataset. The best model earned an accuracy=0.829 or 82.9%. Based on the model, the factors that contribute the most to the average number of viewers per stream are watch time, (total) stream time (in a year), number of peak viewers, number of followers, and whether partnered with Twitch and feature mature content.

Keywords: Twitch · streaming · views

1 Introduction

With the rising popularity of the gaming industry, streaming platforms such as Twitch began to rise which allowed gamers to share their content online with other game enthusiasts. Gaining more than two million concurrent viewers monthly [1], the platform provides career-earning opportunities for its streamers through advertisement revenue.

Gaining high stream views is one of many ways to establish a career in Twitch. Streamers with high Twitch stream views have higher probabilities of earning more while streaming on the platform, such as Ninja, Shroud, and TimTheTatman. However, there have only been a few research on factors that affect Twitch views.

A study by Malm and Friberg explored the prediction of the average number of viewers of a sponsored stream using past available data. The study concluded that the past viewership numbers of a stream and other features such as the type of game and similar genres accurately predict the number of viewers of a

recent stream [2].

Another study by Lê et al. aimed to determine what factors play a part in the success of streamers. They discovered that the common factors among successful streamers were the duration of a stream, the number of streams per week, and the type of content. Live streams of successful streamers did not exceed five (5) hours and had a combination of gaming and non-gaming content posted at least five times a week [3].

Understanding the fundamental factors that affect stream viewership is advantageous for designing strategies to increasing stream views. This paper focused on finding the significant contributors to Twitch viewership. Particularly, the study aimed to explore whether the number of followers and total stream time correlates with viewership.

2 Methodology

This section delineates the implementation of the study. Actual implementation can be found in the Jupyter Notebook[1].

2.1 Data Preparation

Dataset The dataset[2] was obtained from an online platform called Kaggle, a website that allows experts and learners of the data science and machine learning community to distribute their own dataset and use GPU-integrated notebooks such as Jupyter for others to view and work on [4].

The downloaded dataset consisted of 1000 records that takes a look at the applicable data related to Twitch users namely: Channel, Watch time (minutes), Stream time (minutes), Peak viewers, Average viewers, Followers, Followers gained, Views gained, Partnered, Mature, and Language.

Channel - refers to the Twitch channel name.
Watch time (minutes) - refers to the total watch time in the past year[3] in minutes.
Stream time (minutes) - refers to the total stream time in the past year in minutes.
Peak viewers - refers to the highest amount of viewers in all streams in the past year.
Average viewers - refers to the average number of viewers per stream in the past year.

[1] Jupyter Notebook: https://github.com/yourstrulyhb/197-ML-mini-paper/blob/main/Contributing%20Factors%20to%20Twitch%20Viewership.ipynb
[2] Dataset found in https://www.kaggle.com/datasets/aayushmishra1512/twitchdata
[3] Past year pertains to the year 2019.

Followers - refers to the total number of followers of the streamer (by the time of data collection).
Followers gained - refers to the total number of followers gained in the past year.
Views gained - refers to the total number of views gained in the past year.
Partnered - refers to whether the Twitch channel is partnered with Twitch or not.
Mature - refers to whether the Twitch channel contains adult or explicit content.
Language - refers to the language spoken by the Twitch channel.

The independent variable in this dataset is the Channel column while the ten remaining columns are the dependent variables. The feature of interest is 'Average viewers' while other variables are potential predictors except 'Channel'.

Cleaning the dataset Highly variable and non-quantifiable features such as 'Channel' and 'Language' were removed from the experimentation to avoid issues with quantitative data. Quantifiable features however, such as 'Partnered' and 'Mature', which have two options only, were transformed to numerical equivalents.

The raw dataset contained no missing values. But several outliers were identified which can significantly affect the development of the models. To improve accuracy of each model, rows or observations with outlier value for any of the numerical columns or features were removed. This reduced the size of the dataset from 1000 observations to 613 observations.

Standardizing the dataset The dataset was further standardized for efficient computation. Values under initially numerical features - Watch time(Minutes), Stream time(minutes), Peak viewers, Average viewers, Followers, Followers gained, and Views gained - were standardized. For non-numerical columns such as Partnered and Mature, values were maintained (i.e. 0 and 1).

After standardization, the dataset was divided into training and testing datasets.

2.2 Experimental Setups

To effectively determine the factors or features that predict the average number of viewers per stream, the study was divided into three different setups:

Setup 1 All variables or features in the dataset were used in building the model.

Setup 2 Only the hypothesized features, number of followers and total stream time in a year, are used in building the model.

Setup 3 Employ feature selection with Pearson correlation to select the best features in building the model. This setup is further divided into three correlation thresholds=[0.3, 0.4, 0.5]. The upper limit of correlation thresholds were bounded to 0.5 as higher thresholds resulted to all features being included (which is similar to Setup 1).

All in all, five (5) models were generated in the study.

2.3 Linear Regression

This study employed Linear Regression in building models for each experimental setup. Linear Regression is a supervised learning algorithm used in machine learning. It is one of the algorithms capable of predicting or visualizing the relationship between two (or more) different features or variables [5]. It takes the form:

$$y = a_0 + a_1 x + \epsilon \tag{1}$$

where Y is a dependent variable or target variable, X is an independent variable or predictor variable, a_0 is the intercept of the line, a_1 is the coefficient of the linear regression and ϵ is a random error.

Specifically, the study employed Multiple Linear Regression to predict the average number of views per stream, as several variables were used in prediction or building the models.

2.4 Tools and Packages

Data processing, model building, and other computational processes were done using Jupyter Notebook, an interactive computing platform on the web. In pre-processing the data, and creation and evaluation of the model, Python packages of NumPy, pandas, and scikit-learn were used. To visualize data, Python packages of Matplotlib and Seaborn were utilized.

2.5 Evaluation Metrics

The evaluation metrics used in the study were: R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), Accuracy (Regression Score), Cross-Validation Mean.

R^2 score or the coefficient of determination is used to evaluate the performance of a linear regression model. It refers to the amount of variation in the output predicted using predictor(s). An R^2 score value lies between 0 and 1, where a value closer to 1 means a better regression fit.

On the other hand, Mean Absolute Error (MAE) is an evaluation metric that takes the average of all absolute errors between actual and predicted values. And Mean Squared Error (MSE) is the average of the squares of the errors. The lower the value of MAE and MSE, the better the model fits the data.

$$\mathbf{MAE} = \sum_{i=1}^{D} |x_i - y_i| \tag{2}$$

$$\mathbf{MSE} = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{3}$$

Accuracy is described as the portion of correct predictions made among the total predictions. The values in this category lie between the values 0 and 1. A higher score shows that the model was able to properly determine the right prediction.

Cross-Validation Mean is a method used to determine whether the model generalizes itself to the test dataset. It partitions the dataset into sections of training and testing datasets then averages the final result to ensure model optimization. Cross-Validation Mean was used as secondary or verification metric to evaluate the models.

3 Results and Discussion

3.1 Model 1: Model with All Features

Table 1 show the intercept and coefficients of the model with all features: watch time (minutes), stream time (minutes), peak viewers, followers, followers gained, views gained, partnered and mature.

Table 1. Model 1 intercept and coefficients from linear regression with all features.

	Int	**Watch time (minutes)**	**Stream time (minutes)**	**Peak viewers**	**Followers**
Coef	0.2933015	0.726076	-0.680808	0.05976	0.021213
	Followers gained	**Views gained**	**Partnered**	**Mature**	
Coef	0.018277	0.006903	-0.312851	0.047701	

Figure 1 shows the scatterplot of the actual values against the predicted values of Model 1. The values closely fit the regression line, indicating that the model worked well in predicting the average number of viewers per stream.

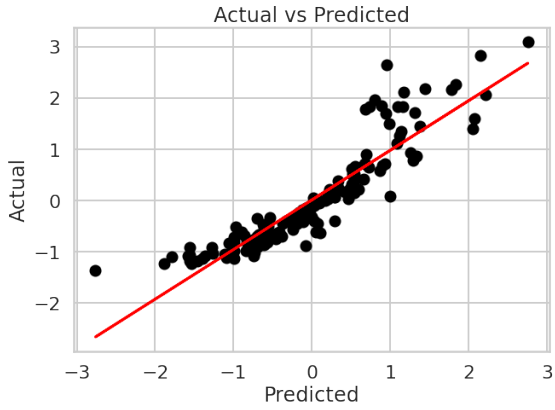

Fig. 1. Actual vs. Model 1 Predicted values

3.2 Model 2: Model with Selected Features (Followers and Stream Time)

Table 2 shows the intercept and coefficients of the model with the hypothesized features, number of followers and total stream time in a year (minutes).

Table 2. Model 2 intercept and coefficients from linear regression with hypothesized features (Followers and Stream Time).

	Int	**Followers**	**Stream time (minutes)**
Coef	-0.014194	0.292029	-0.491101

Figure 2 shows the scatterplot of actual data label against predicted values of Model 2. As observed, the predicted values are dispersed across the graph and does not fit closely with the regression line. This indicates that the model performed poorly in predicting average viewers per stream.

3.3 Model 3: Feature Selection with Pearson Correlation

Model 3.1 Selected features (threshold=0.3) Table 3 shows the intercept and coefficients of the model with selected features: watch time(minutes), stream time (minutes), partnered and mature.

Figure 3 shows the scatterplot of the actual 'average stream viewers' values against the predicted values of Model 3.1. It can be seen that the values are closely fitted to the regression line with the exception of some outliers, which

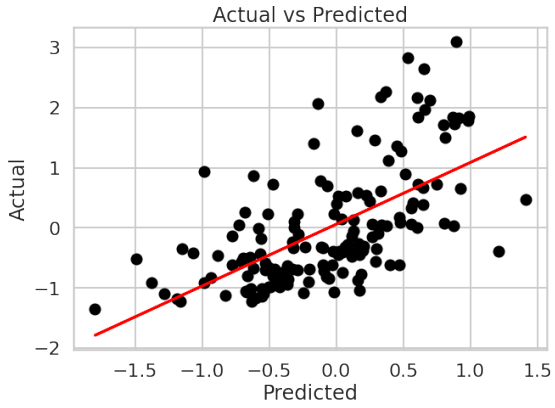

Fig. 2. Actual vs. Model 2 Predicted values

Table 3. Model 3.1 coefficients from linear regression with selected features: watch time (minutes), stream time (minutes), partnered, mature.

	Int	**Watch time (minutes)**	**Stream time (minutes)**	**Partnered**	**Mature**
Coef	0.260330	0.762565	-0.720557	-0.273532	0.025943

indicates good performance from the model with correlation threshold=0.3.

Model 3.2 Selected features (threshold=0.4) Table 4 shows the intercept and coefficients of the model with selected features: watch time(minutes), stream time (minutes), peak viewers, partnered and mature.

Table 4. Model 3.2 coefficients from linear regression with selected features: Watch time(minutes), Stream time(minutes), Peak viewers, Followers, Partnered, Mature.

	Int	**Watch time (minutes)**	**Stream time (minutes)**	**Peak viewers**	**Followers**	**Partnered**	**Mature**
Coef	0.297777	0.72955	-0.684541	0.064437	0.028368	-0.316929	0.043872

Figure 4 shows the scatterplot of the actual values against the predicted values of Model 3.2. The results are seen to have a few outliers and fit the regression line, suggesting that the model behaved well with a correlation threshold=0.4.

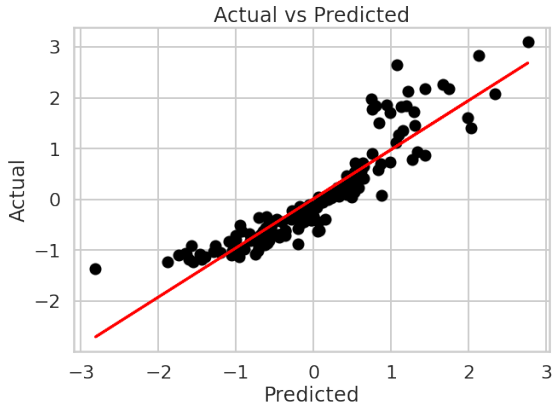

Fig. 3. Actual vs. Model 3.1 Predicted values

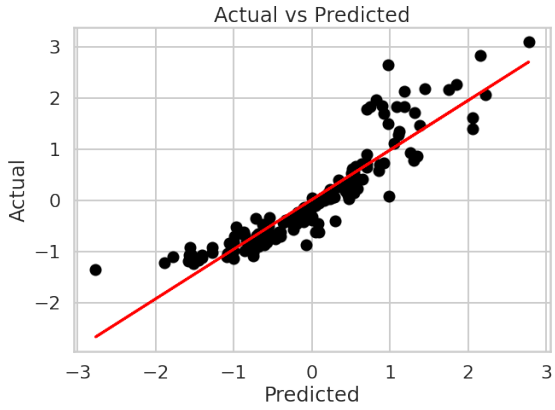

Fig. 4. Actual vs. Model 3.2 Predicted values

Model 3.3 Selected features (threshold=0.5) Table 5 shows the intercept and coefficients of the model with all the features, except views gained, with correlation threshold=0.5.

Figure 5 shows the scatterplot of the actual values against the predicted values of Model 3.3. The results are also found to fit the regression line with a few outliers, showing that the model behaved well with a correlation threshold=0.5.

3.4 Evaluation of Models

Table 6 shows evaluation scores of all models for the following metrics: R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), Accuracy (Regres-

Table 5. Model 3.3 intercept and coefficients from linear regression with all the features, except views gained.

	Int	**Watch time (minutes)**	**Stream time (minutes)**	**Peak viewers**
Coef	0.295097	0.72971	-0.680838	0.060772
	Followers	**Followers gained**	**Partnered**	**Mature**
Coef	0.021738	0.018391	-0.314334	0.045911

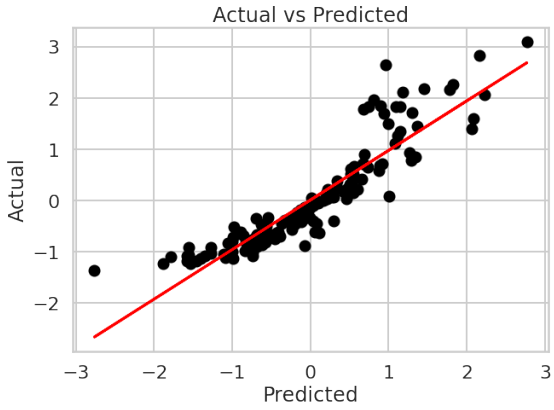

Fig. 5. Actual vs. Model 3.3 Predicted values

sion Score), and Cross-Validation Mean.

All models, except Model 2, performed well in predicting the average viewers per stream in unseen data. Model 2 got the lowest R^2 (=-0.713) and accuracy (=0.379) scores, and the highest MAE and MSE scores. This rejects the hypothesis that the number of followers and total stream time (in a year) alone significantly affect the average viewers per stream of a Twitch streamer.

Meanwhile, Model 3.2 performed the best with the highest R^2 (=0.807) and accuracy (=0.829) scores, and the lowest MAE and MSE scores. This tells that the selected features with Pearson correlation namely, Watch time (minutes), Stream time (minutes), Peak viewers, Followers, Partnered, and Mature, contribute the most to the average viewers per stream of a Twitch streamer.

Thus, the best model for predicting the average viewers per stream or viewership for a Twitch stream is:

Table 6. Summary of evaluation metrics scores of all models.

Model	R^2	**MAE**	**MSE**	**Accuracy**	**CV Mean**
Model 1	0.805	0.287	0.163	0.827	70.890
Model 2	-0.713	0.602	0.586	0.379	-6.129
Model 3.1	0.806	0.297	0.163	0.827	70.514
Model 3.2	0.807	0.285	0.160	0.829	71.129
Model 3.3	0.805	0.287	0.163	0.827	71.051

Average Viewers Per Stream = 0.298 + 0.730 * *Watch time (minutes)* - 0.685 * *Stream time (minutes)* + 0.064 * *Peak viewers* + 0.028 * *Followers* - 0.317 * *Partnered* + 0.044 *Mature* (4)

4 Conclusion

The objective of this paper was to evaluate the factors that influence Twitch stream viewership. In doing so, three experimental setups were devised in order to understand the effects of the features against the average number of viewers. Setup 1 involved all of the features in the dataset, Setup 2 only involved the proposed features of higher viewership while Setup 3 involved feature selection using Pearson correlation and is subdivided into three thresholds=[0.3, 0.4, 0.5]. Linear regression was used to build models for each setup which were evaluated with the metrics: R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), Accuracy (Regression Score), and Cross-Validation Mean.

Except for Setup 2, all of the models scored well, with Model 3.2 showing the best evaluation results. The number of followers and total stream time in a year does not correlate with the average number of views per stream. However, watch time, stream time, number of peak viewers, number of followers, and whether partnered and features mature content significantly contribute to or predict the average number of viewers per stream. These factors may be taken into consideration by aspiring and experienced Twitch streamers to maintain current viewership ratings or increase viewership in future streams.

Further improvements of the models may be done by using a larger dataset, employing other feature selection methods, or by using other machine learning algorithms to build the models such as Huber regression.

5 Bibliography

References

1. "Twitch statistics & charts · twitchtracker." [Online]. Available: https://twitchtracker.com/statistics
2. J. Malm and M. Friberg, "Viewership forecast on a twitch broadcast: Using machine learning to predict viewers on sponsored twitch streams," 2022.
3. H. Le, J. Wu, L. Yu, and M. Lynn, "A study on channel popularity in twitch," *arXiv preprint arXiv:2111.05939*, 2021.
4. C. Uslu, "What is kaggle?" Mar 2022. [Online]. Available: https://www.datacamp.com/blog/what-is-kaggle
5. D. Nelson, "What is linear regression?" Jun 2021. [Online]. Available: https://www.unite.ai/what-is-linear-regression/

A Appendix

A.1 Dataset source

The dataset used in this study is available online through Kaggle:
`https://www.kaggle.com/datasets/aayushmishra1512/twitchdata`

A.2 Jupyter notebook

Processes and computations done in this study can be viewed in this Jupyter notebook:
`https://github.com/yourstrulyhb/197-ML-mini-paper`