Affiliate Status Identification using Twitch Streaming Data: Network Analysis

Ahmad Kamal Baig

Abstract

This report provides a comprehensive analysis of Twitch streaming data, focusing on the graph analysis, feature engineering, and predictive modeling using Graph Convolutional Networks (GCN). By dissecting user engagement metrics, account features, and network structure, we gain insights into factors influencing content creators' success and the underlying social network dynamics. Additionally, the report highlights the use of predictive modeling in identifying Twitch affiliate status, which is crucial for targeted advertising and facilitating collaborations, thereby enhancing the monetization strategies for creators.

1 Introduction

Twitch, a leading platform in the live streaming realm, especially within the gaming community, offers a rich dataset for analysis. This report aims to uncover the nuances of user interactions, content trends, and the communal framework on Twitch. Our objective is to conduct an exploratory data analysis, fine-tune the dataset for machine learning, and employ a graph neural network model to predict users' affiliate status. By doing so, we aim to provide actionable insights for streamers and marketers to better navigate the platform's dynamic ecosystem.

2 Dataset

The dataset for our analysis, focusing on Twitch's streaming ecosystem, was obtained from the Stanford Network Analysis Project (SNAP) in Spring 2018. It comprises an extensive social network of Twitch users, characterized by nodes representing individual users and edges indicating mutual follower relationships. This dataset not only includes user features but also encapsulates their network interactions, providing a comprehensive view for understanding mutual follower-ship trends and their correlation with various aspects of video streams on Twitch.

3 Graph Analysis

A subset of the original edge data was selected to create a manageable and representative sample of the entire network. The sample includes 1,000 edges, chosen randomly with a fixed seed for reproducibility. The sampled edges were used to construct a graph, using the NetworkX library.

3.1 Degree Centrality and Communities

The visualization suggests a network with well-defined communities, as indicated by clusters of the same color. Larger nodes, representing higher degree centrality, are distributed across various communities, implying that influential nodes are not confined to a single cluster but are spread throughout the network. This pattern of influential nodes suggests multiple points of robust communication or information dissemination. The varying node sizes within each community also indicate a hierarchy of influence or connectivity, with some members being more central than others within their respective groups.

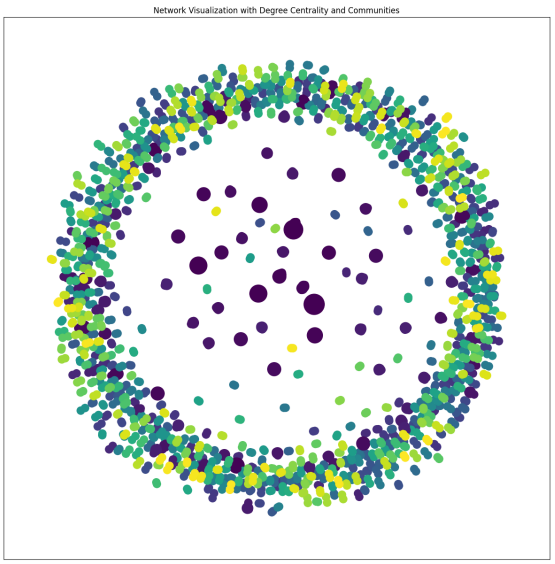

Figure 1: Degree Centrality and Communities

3.2 Node Attribute Analysis

3.2.1 Language Attribute

- **Predominance of English:** Blue nodes symbolize English speakers, showing English as the predominant language in the network. The extensive spread represents English as a primary communication medium.
- **Linguistic Bridge:** English-speaking nodes link various language groups, acting as a linguistic bridge.
- **Localized Language Communities:** Non-English speakers, depicted in other colors, form less numerous but more clustered groups, suggesting localized language communities.
- **Bilingual Communication:** English nodes within non-English clusters may represent bilingual or multilingual communication, underscoring English's central role in the network.
- **English as a Link Language:** The analysis emphasizes the significance of English as a connecting language within the network's multicultural landscape.

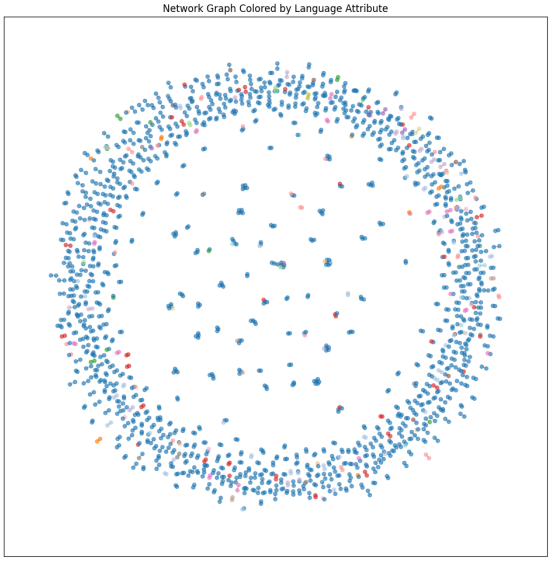

Figure 2: Language Attribute Node Analysis

3.2.2 Affiliate Status Attribute

- **Affiliated vs. Non-affiliated Nodes:** The nodes exhibit a mixture of affiliation statuses, with green representing affiliated and red representing non-affiliated users.
- **Distribution Pattern:** The even distribution of green and red nodes throughout the network indicates that affiliation status does not segregate users into distinct clusters.
- **Connectivity Patterns:** This interspersion suggests that affiliation status is not the primary driver of connectivity patterns among users on the platform.
- **Lack of Segregation:** The absence of clear segregation by affiliation hints at the possibility that other factors might influence network connections more significantly. Such factors could include content genre, streaming schedule, or other user attributes.
- **Connectivity Landscape:** No single group appears to dominate the connectivity landscape, suggesting a balanced network structure.

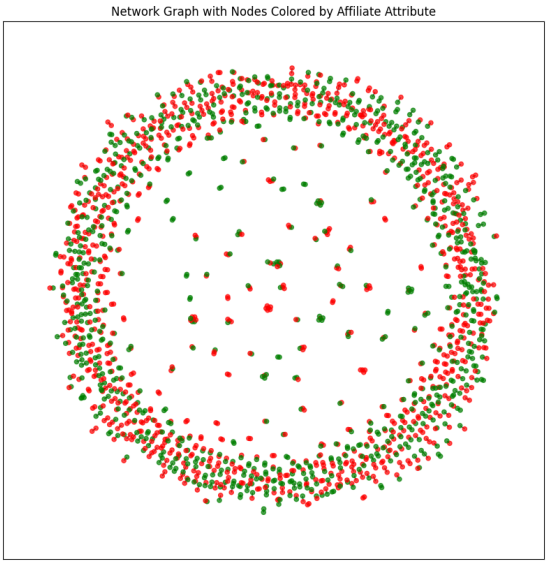

Figure 3: Affiliate Attribute Node Analysis

4 Predictive Modeling

4.1 Data Cleaning and Preprocessing

The raw dataset underwent a series of preprocessing steps to ensure data quality and readiness for analysis. Key steps included:

- Converting 'created_at' and 'updated_at' to the datetime format to facilitate time-based analysis.

- Calculating the 'account_age' to assess the influence of account longevity on user engagement.

4.2 Feature Engineering

Feature engineering aimed to enrich the dataset with informative signals for the GCN model. This process included:

- Log transformation of the 'views' feature to mitigate the effects of extreme values.
- Box-Cox transformation of 'life_time' to stabilize variance and make the data more Gaussian-like.

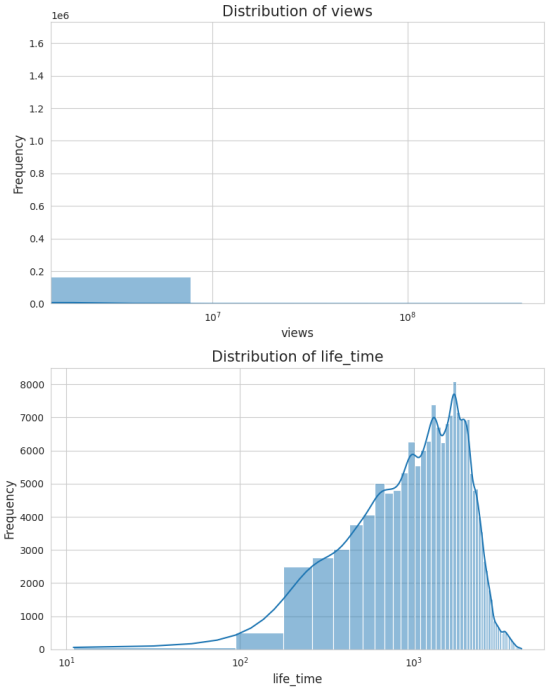

Figure 4: Before Feature Engineering

- Encoding categorical variables such as 'language' using binary encoding to capture the diversity of content creators' languages.

4.3 Graph Convolutional Network Model

A GCN model was chosen for its ability to leverage the network structure inherent in the data. It operates on the graph's nodes and their connections, allowing it to incorporate neighborhood information when making predictions. The model consists of the following layers:

- An input layer that applies graph convolution over the node features and edge indices.

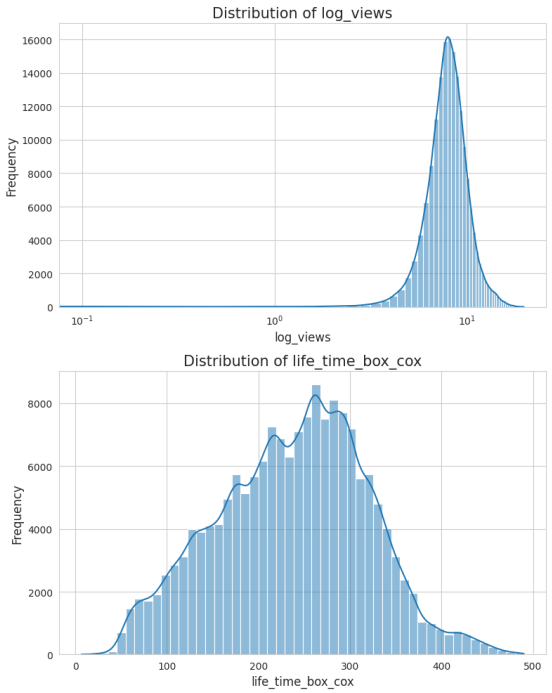

Figure 5: After Feature Engineering

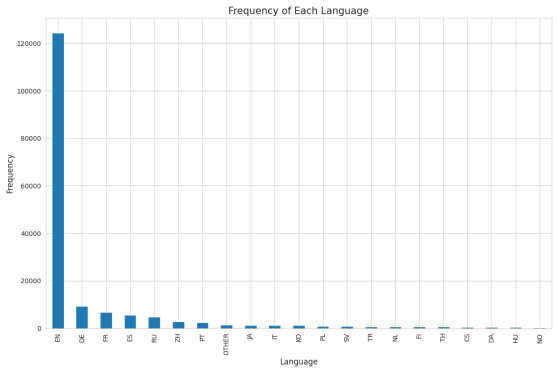

Figure 6: Language Distribution

	account_age	log_views	life_time_box_cox	mature	dead_account	language_bin_0	language_bin_1	language_bin_2	language_bin_3	language_bin_4
0	970	8.972083	175.602666	1	0	0	0	0	1	1
1	2704	6.216606	361.409672	0	0	0	0	0	1	1
2	3150	12.854492	402.815366	1	0	0	0	0	1	1
3	1356	5.958425	221.188442	0	0	0	0	0	1	1
4	1786	7.818832	270.042502	0	0	0	0	0	1	1
...	...	...	...	...	...	...	...	...	...	...
168109	815	8.510370	154.722885	0	0	0	0	0	1	1
168110	2081	8.325791	300.868937	1	0	0	0	0	1	1
168111	1800	8.173575	271.426964	0	0	0	0	0	1	1
168112	2136	13.702047	306.447851	1	0	0	0	0	1	1
168113	2090	6.674561	293.189686	0	0	0	0	0	1	1

168114 rows × 10 columns

Figure 7: Final Features

- Several hidden layers with graph convolution, batch normalization, and ReLU activation functions to capture the non-linear interactions between nodes.
- An output layer with a sigmoid activation function to predict the binary affiliation status.

Layer Type	Input Dimension	Output Dimension
GCNConv	10	128
GCNConv	128	64
BatchNorm	64	64
GCNConv	64	32
BatchNorm	32	32
GCNConv	32	16
BatchNorm	16	16
GCNConv	16	8
BatchNorm	8	8
Linear	8	4
Linear	4	2
Linear	2	1

Table 1: Architecture of the Twitch Affiliate Neural Network

4.4 Model Training and Evaluation

4.4.1 Considerations

- Utilizing the **Adam optimizer** with an initial learning rate of 0.05 and a learning rate scheduler to reduce the rate upon validation loss plateauing.
- Applying **gradient clipping** to stabilize training.
- Evaluating model performance based on **binary cross-entropy loss** and accuracy on a held-out test set.

4.4.2 Results

- **Rapid initial learning** is indicated by swift improvements in accuracy metrics during the early epochs for training, validation, and test sets.
- The **plateauing of accuracy** metrics suggests the model has reached its learning capacity.
- **High and consistent accuracy** across all data sets indicates good generalization and absence of overfitting.
- **Effective loss minimization** is shown by the loss graph during the early stages of training.
- **Stabilization of loss values** implies that the model is converging to a minimum loss.
- Training, validation, and test losses closely converging suggest that the model is **well-generalized** and neither overfitting nor underfitting.

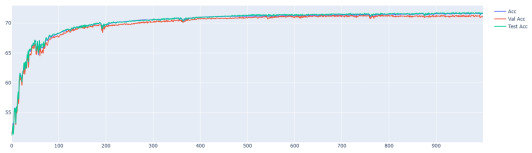

Figure 8: 70% Accuracy

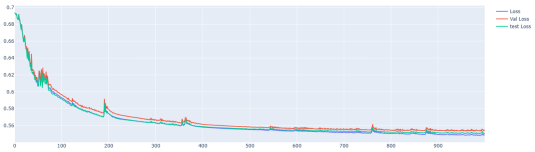

Figure 9: 54% Loss

5 Conclusion

The GCN model demonstrated promising results, with an accuracy that suggests a strong relationship between the features engineered and the likelihood of a Twitch user being an affiliate. The analysis highlighted the potential of machine learning in social network analysis and opened avenues for further exploration, such as incorporating additional user engagement metrics and refining the model architecture.

6 References

1. https://snap.stanford.edu/data/twitch_gamers.html