BiliBili Data Analysis

Yingjia Liu(柳莺佳)
12311452

Dawei Huang(黄大为)
12312405

Huipeng Huang(黄辉鹏)
12312107

Abstract

With the growing popularity of BiliBili, an increasing number of potential content creators schedule to become uploaders. To maximize their chances of success, this report proposes a comprehensive guide on how to create contents that aligns with users' preferences and gain more followers. Through analysis of video categories, we provide recommendations for choosing promising content categories. Furthermore, we investigate factors that matter most to number of followers. Finally, we build multiple models to predict uploaders' video categories.

1 Data Processing

Before using data, we assess the dataset to identify missing or duplicated entries and introduce Table 1, which guarantees the data is already clean. In Figure 1, we use histplot to show the distribution of followers and observe that over 80% of followers have less than 5×10^5 followers, implying a **Long-Tail Distribution**. Therefore, we will apply a **logarithmic transformation** when analyzing the relationship between followers and other variables.

Name	Missing Data	Duplicated Data
BiliBili	0	0

表 1: Missing or Duplicated Data

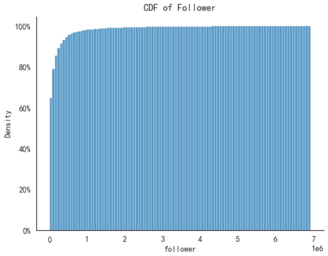


图 1: Cumulative Distribution of Followers

2 Analysis on different Video Categories

To gain insight into BiliBili users' preferences, we introduce Figure 2, a pie chart of the distribution of total followers across various video categories. It shows that 'Game' and 'Life' are two dominant video categories in BiliBili. The other categories each account for less than 10% of the total user base. Therefore, it is an imbalanced dataset.

By counting the frequency of self-tags, we introduce Figure 3, a word cloud diagram where the size of words represents its' frequency. Top frequent self-tags include 'Game', 'Life' and 'Delicious Food', supporting that 'Game' and 'Life' are two major categories. Additionally, the abnormally high frequency of the 'Guichu' tag suggests a particular trend among content creators in this category. It appears that creators are more likely to self-identify with this tag, possibly as a strategic choice to stand out in a niche yet highly engaging content category.

Considering 'Game' and 'Life' categories' large user base, they appear to be the most prospective video categories. However, the number of uploaders across different categories is varied according to Figure 4. 'Game' and 'Life' categories also have the largest amount of uploaders, revealing that they are extremely competitive. It motivates us to analyze the intensity of competition in different categories and explore which category is less competitive and easier to succeed for new uploaders.

Intensity of Competition in different Video Categories We draw boxplot with **video_tag_combine** as hue, to quantify the difficulty of reaching certain amount of followers in different fields. Through Figure 5, we know Dance, Ghost

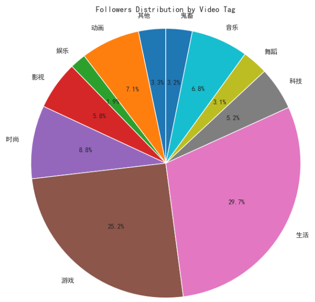


图 2: Followers Distribution

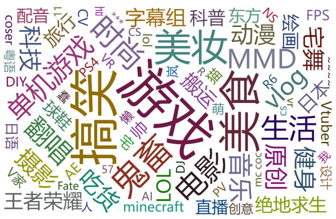


图 3: Self-Tags Word Cloud

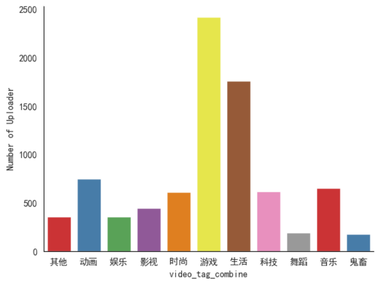


图 4: Number of Uploaders in different categories

& Animal Video, and film have good box height. The majority of uploaders in this categories have good followers base. Besides, we want to analyze the risk of becoming outliers. Although Boxplot provides a clear visualization of outliers, it can not fully capture the risk of becoming an outlier as it only shows the number of outliers and neglects the variation in the total number of uploaders across categories.

To address this problem, we introduce Figure 6, a bar chart that shows the probability of becoming outliers in each fields. 'Game', 'Entertainment', and 'Life' have high risk of becoming outliers. Although 'Game' and 'Life' have large user bases, they are extremely competitive and have high risk of failure. 'Dance', 'Guichu' and 'Animation' balance box height and risk of becoming outliers well, marking promising growth potential for new uploaders. New uploaders must carefully estimate the risk they could accept and the target number of followers they aim to achieve to choose suitable video categories. If they could accept high risk and have ambitious target, ’Game’ and ’Life’ are the best choices. If they could not undertake high risk of failure, 'Guichu', 'Animation' and 'Dance' are better choices.

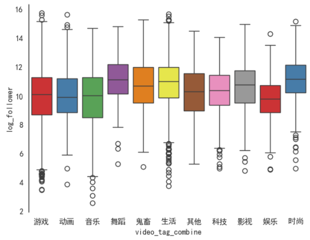
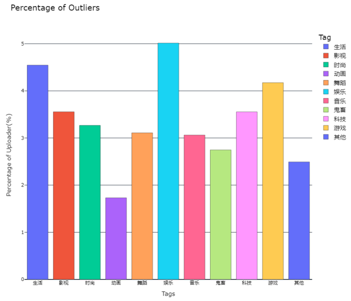


图 5: Boxplot of Followers in different categories



Tag
🟦 生活
🟥 影视
🟩 时尚
🟪 动画
🟧 舞蹈
🟦 娱乐
🟥 音乐
🟩 鬼畜
🟪 科技
🟨 游戏
🟦 其他

图 6: Outliers Percentage in different categories

Gender Preference We use treemap to visualize gender preferences for different categories, where the area represents total number of uploaders and the depth of color represents the average number of fans. From the area of different category, we find that women prefer creating contents regarding 'Fashion', 'Dance' and 'Life'. From the color of each categories, we find that in 'Dance', 'Fashion' and 'Music' categories, women have a higher average number of followers than men. 'Secret' gender has an extraordinarliy high average number of followers in 'Guichu'. It appears that most uploaders creating content of 'Guichu' prefer to keep their gender and personal information private.

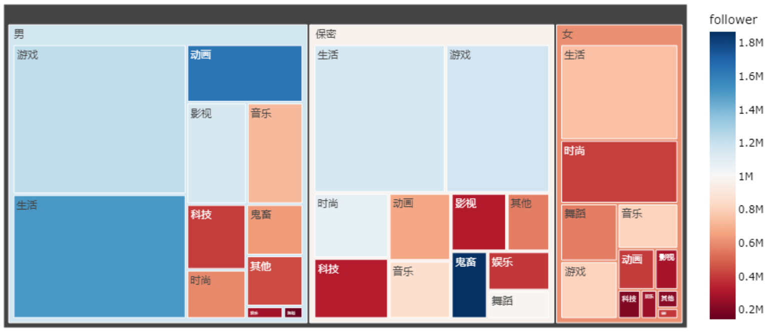


图 7: Treemap of Followers by Sex and Video Tag

3 Analyze Factors affecting Number of Follower

3.1 Impact of Video Length on Follower Count

To ascertain the predominant factor influencing the number of followers, we constructed a random forest model utilizing follower data and its correlates. The result is showed in Figure 8. The bar plot illustrates a significant correlation between the number of followers and various factors, including video view, video length, and the number of videos. Consequently, we intend to undertake a further investigation into the influence of these factors.

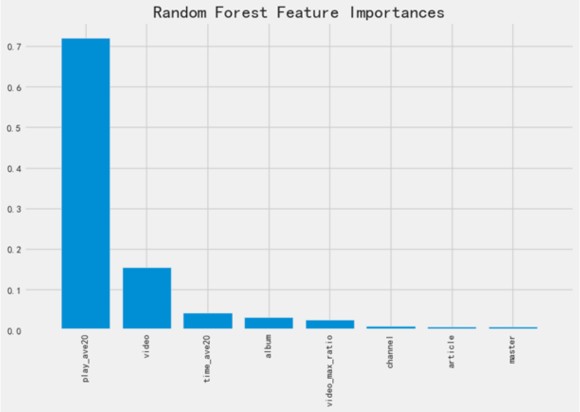
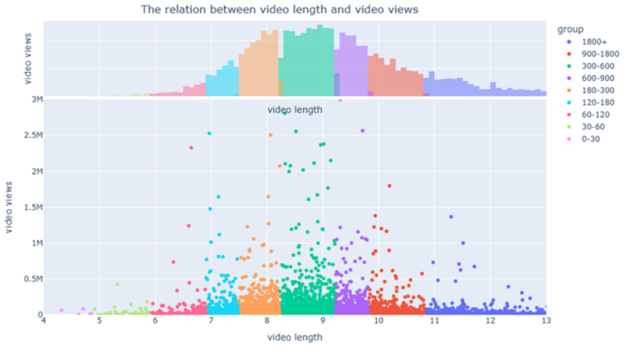


图 8: Rank of factors matter to number of followers

Initially, we focused our research on the impact of video length on follower. To provide an overview of this relationship, we utilize a combination of scatter plot and histogram (Figure 9), which reveals a normally distributed pattern between video length and video views. Our analysis indicates that videos ranging from 3 to 15 minutes are more popular. This trend can be attributed to the platform-specific preferences of viewers: short videos are favored on TikTok, while longer videos are more common on Tencent Video. On Bilibili, users exhibit a preference for medium-length videos. Moreover,the heatmap (Figure 10) indicates that the correlation coefficient between video category and video length is the most significant. In light of this, we generate scatter plots based on video categories, where the color represents the bloggers with different follower number.



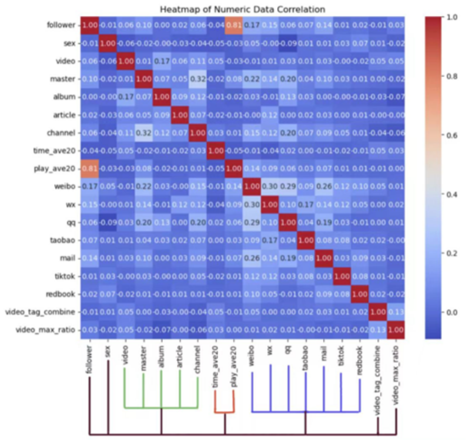


图 9: The scatter plot of video length and video views

图 10: Heatmap of all data

We take the ‘Entertainment’ and ‘Life’ category for example. Within the domains of ‘Entertainment’, the scatter points are concentrated at the bottom and evenly distributed, indicating video length has little impact on video views. When differentiating by color, it is noticeable that bloggers with varying numbers of followers are intermingled. This indicates that in the ‘Entertainment’ category, there is a relatively low level of fan loyalty, which creates a more accessible environment for new bloggers. Nonetheless, within the domains of ‘Life’, the differentiation in coloration is evident. Videos created by bloggers with millions of followers tend to have a consistent length of three minutes, and garner a greater viewership than those produced by other bloggers. Within the ‘Life’ category, there is a marked degree of fan loyalty, which serves to erect a substantial impediment to the entry of new bloggers into this domain.

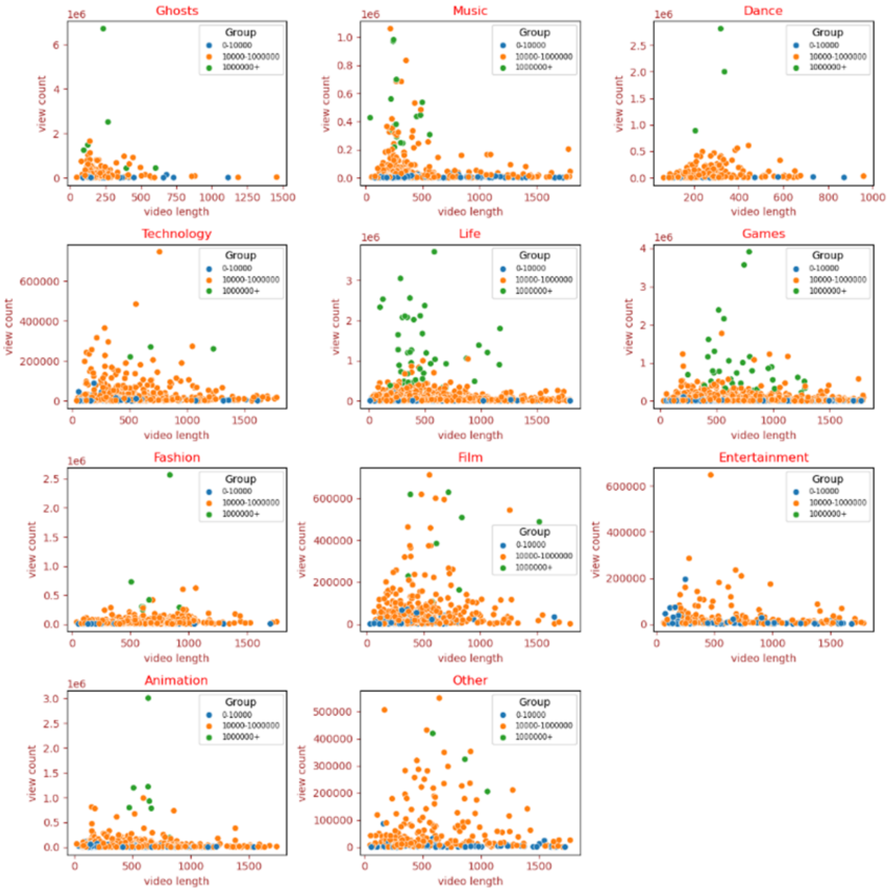


图 11: The bar plot based on video categorifies

3.2 Analysis of Relationship betweeen Video Number, Master Number and Follower Count

In contrast to the mature video production employed by bloggers with million followers, new bloggers possess limited resources in energy and experience. Consequently, they must navigate a strategic dilemma between increasing video number and improving video quality. Therefore, we investigate the the relationship between the number of videos, the number of master and followers. Due to the involvement of three variables, we construct a three-dimensional scatter plot (Figure 12) to visualize their relationships. Furthermore, we quantify the promotional impact using precise statistical measures such as correlation coefficients and p-values, which are summarized in the following table (Figure 13).

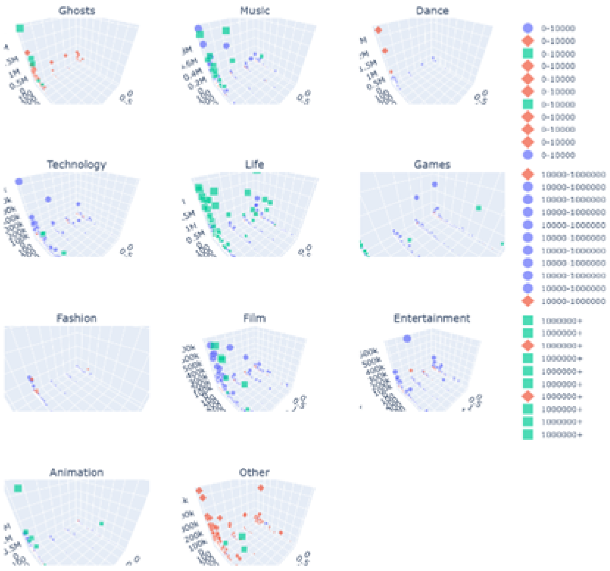
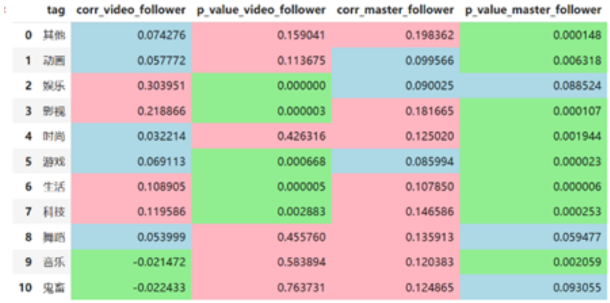


图 12: The three-dimensional scatter plot



	tag	corr_video_follower	p_value_video_follower	corr_master_follower	p_value_master_follower
0	其他	0.074276	0.159041	0.198362	0.000148
1	动画	0.057772	0.113675	0.099566	0.006318
2	娱乐	0.303951	0.000000	0.090025	0.088524
3	影视	0.218866	0.000003	0.181665	0.000107
4	时尚	0.032214	0.426316	0.125020	0.001944
5	游戏	0.069113	0.000668	0.085994	0.000023
6	生活	0.108905	0.000005	0.107850	0.000006
7	科技	0.119586	0.002883	0.146586	0.000253
8	舞蹈	0.053999	0.455760	0.135913	0.059477
9	音乐	-0.021472	0.583894	0.120383	0.002059
10	鬼畜	-0.022433	0.763731	0.124865	0.093055

图 13: The correlation coefficient and p value

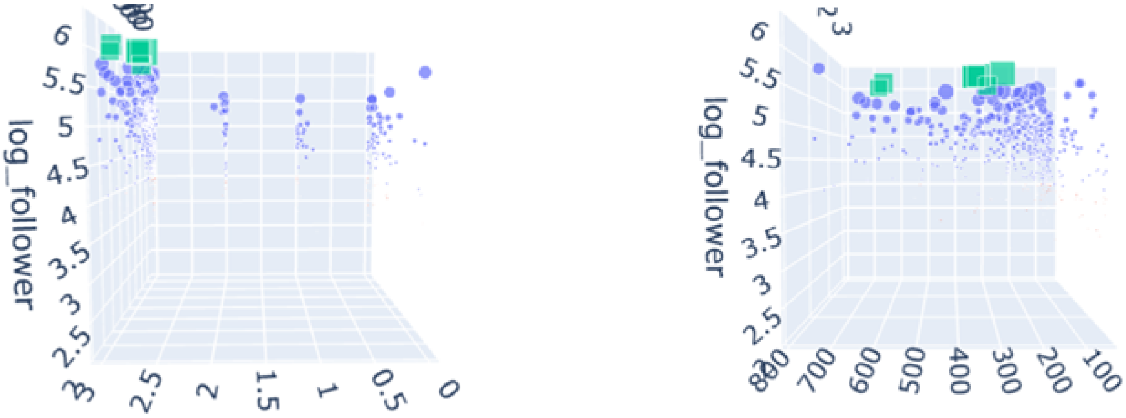


图 14: Scatter for master-follower and video-follower

The statistics data indicates that the promotional impact of the number of videos and the number of masters on follower varies across different categories. Consequently, we select the 'Film' category for a further research. Figure 14 depicts two-dimensional scatter plots illustrating the fluctuations in follower in relation to the number of videos and the number of masters respectively. We apply linear regression models to these scatter plots (Figure 15), with the correlation coefficients of 0.22 and 0.18. This suggests that in the 'Film' category, the quantity of videos exerts a more pronounced promotional effect on the followers. Hence, for new bloggers in the 'Film' category, it is recommended to increase the frequency of video updates and improve exposure on the homepage to gain more followers.

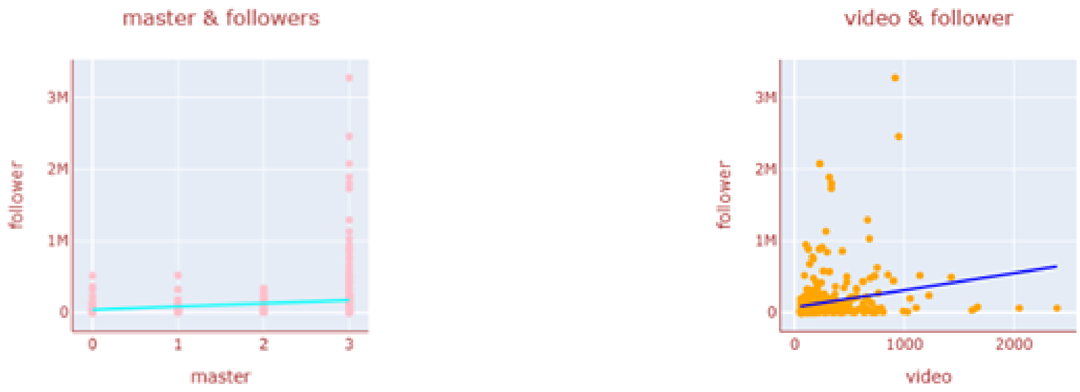


图 15: The linear regression for master-follower and video-follower

4 Analyze factors affecting the number of Fans from other perspectives

BiliBili is widely recognized not only as China's largest bullet screen video website but also as a diversified content sharing platform, offering a range of features to enhance blogger's exposure and increase follower count. Beyond its core function of video publishing, BiliBili provides additional content formats such as albums, articles, and channels. In the following discussion, we will first analyze the impact of these supplementary features on follower numbers. Secondly, we will compare these diversified content formats with strategies that rely solely on video publishing, examining their differing effects on attracting new followers and retaining existing ones.

4.1 The Impact of Integrated Factors on Follower Count

4.1.1 Definition of Integrated Factors

Use *Principal Component Analysis (PCA)* to define integrated factors. PCA is a powerful dimensionality reduction technique widely used to transform high-dimensional data into low-dimensional representations while preserving as much of the data's main variance as possible. This method helps simplify complex datasets by identifying the directions of the most significant variation in the data.

The main reason for choosing PCA is to address potential multicollinearity and redundant information issues in the dataset. PCA helps by reducing the dimensionality of features and extracting the most representative components, highlighting the main patterns of variation in the data. It not only simplifies analysis but also enhances the interpretability of the results.

Steps to apply PCA to the dataset:

- Standardize data: Use StandardScaler to standardize features (video, album, article, channel) to ensure each feature has the same scale.
- PCA transformation: Apply PCA to reduce the four features to one principal component.

Results Interpretation: The calculated principal component loadings are shown in Table 2.

Content Formats	Video	Album	Article	Channel
Component Loadings	0.4120	0.6137	0.5167	0.4321

表 2: Principal Component Loadings of different Content Formats

The loadings represent the contribution of each original feature to the composite factors. Higher loadings indicate that the feature contributes more to the principal component.

4.1.2 Overall Data Analysis

Plot a scatter diagram of the number of videos and integrated factors against the number of followers, and add a regression line. In Figure 16, **the left image** shows a positive correlation between the number of videos and the number of followers.

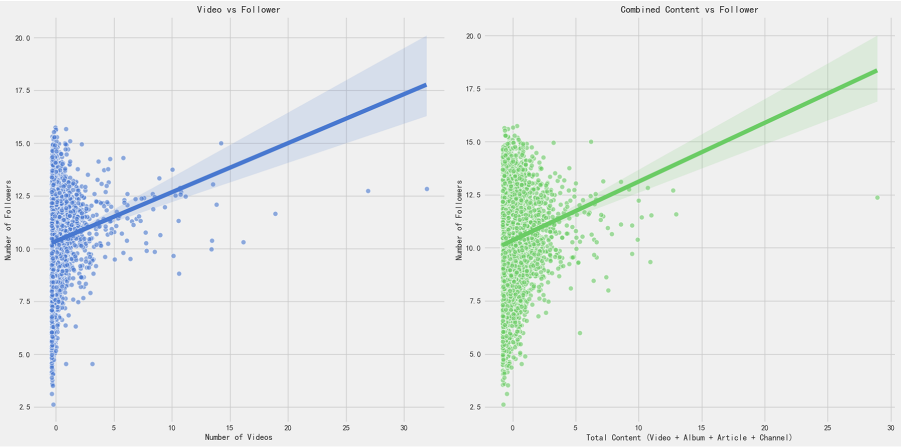


图 16: Number of Videos and Integrated Factors against Number of Followers

This suggests that posting more videos may help increase the number of followers, but the relationship is not very strong and may be influenced by other factors.

The right image shows that there is also a positive correlation between integrated content and the number of followers, and the correlation appears to be stronger than relying solely on videos.

Overall, to better demonstrate the correlation between the two, we used the *Pearson correlation coefficient* to calculate their correlation relative to the number of followers.

Video Correlation	0.0666
Integrated Content Correlation	0.0693

表 3: Pearson Correlation of Number of Videos and Integrated Factors against Number of Followers

Although both correlations are relatively weak, overall, the combined factors have a greater positive impact on the increase in followers. Below, we will analyze the correlation in different regions by video category.

4.1.3 Data Analysis on Video Category

Calculate the correlation coefficients under different video categories and display them using a bar chart , as shown in the Figure 17. Figure 18 is the relevant heat map , which can better compare the differences in the correlation of influence on followers between the two aspects. First, the following analysis focuses on the video categories where the influence of integrated factors on the increase in the number of followers is greater than that of the number of videos.

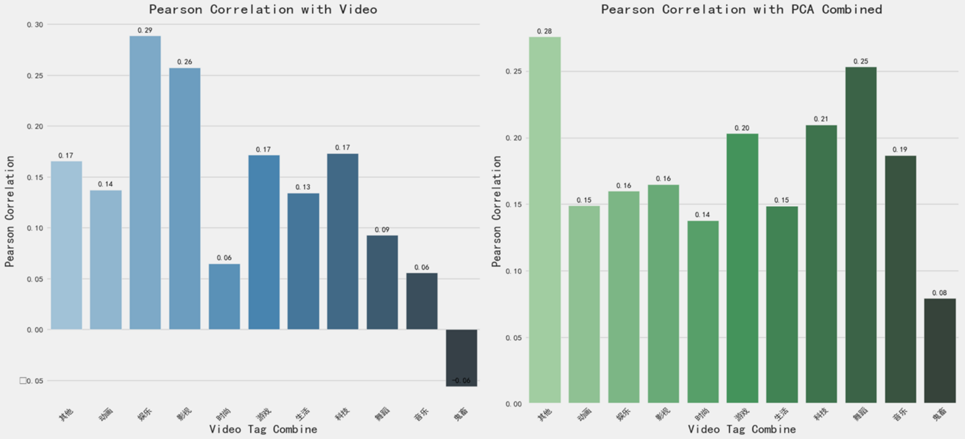


图 17: Correlation Coefficient

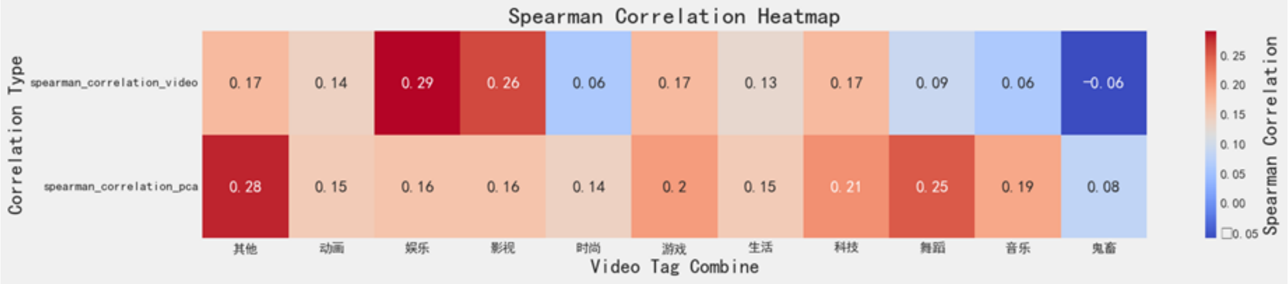


图 18: Relevant Heatmap

- **Dance**

In diverse forms of content, dance can more fully showcase its unique charm. For example, content creators in the dance field can not only share photos of stage performances but also upload pictures of their daily lives in albums. This approach not only makes creators appear more approachable but also deepens fans' understanding of them from multiple angles, thereby enhancing fans' loyalty and engagement.

Secondly, the following analysis focuses on the video categories where the influence of integrated factors on the increase in the number of followers is worse than that of the number of videos.

- **Entertainment**

Video format entertainment content is more appealing because its immediacy and interactivity better engage fans, while static photos and long articles struggle to hold users' attention in the rapidly changing social media environment.

4.2 Analysis of the Impact of Social Media Connections on Follower Count

Currently, various social media and video platforms emerging continuously, leading to intense competition. Against this background, we will explore the interactions between BiliBili and the current mainstream platforms—Weibo, WeChat, QQ, mail, Taobao, TikTok, and Redbook. The core focus of the upcoming analysis is how this cross-platform linkage will impact Bilibili's fan growth and user engagement.

4.2.1 Relationship between Number of Connected Platforms and Number of Followers

To explore the relationship between the number of linked platforms and the number of followers, we created the following box plot (Figure 19). We can draw the following conclusion：

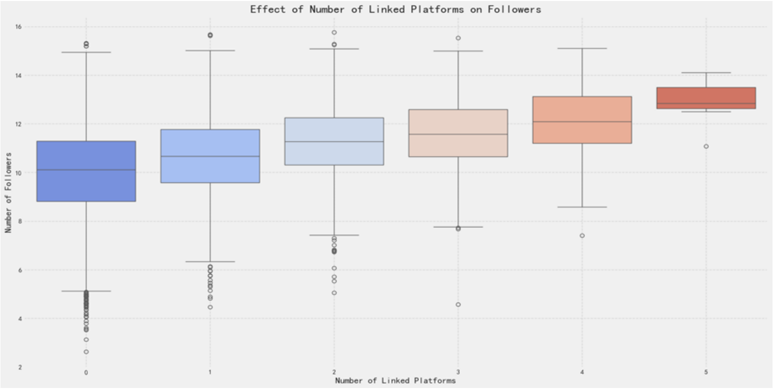


图 19: Boxplot of the Number of Associated Platforms and the Number of followers

- **The number of associated platforms is positively correlated with the number of followers.**
  As the number of associated platforms increases, the median number of followers (represented by the black line in the box plot) gradually rises. This indicates that associating with more platforms has a positive effect on increasing the number of followers.
- **The range of fluctuations gradually decreases as the number of platforms increases.**
  When the number of associated platforms is small (e.g., 0 or 1 platform), the distribution range of follower numbers is wider, with a larger interquartile range and more outliers. As the number of associated platforms increases (e.g., 4 or 5 platforms), the distribution range of follower numbers gradually narrows, and volatility decreases, indicating that associating with more platforms may lead to more stable follower growth.
- **The effect is most significant when the number of linked platforms is 4 or 5.**
  When the number of linked platforms reaches 4 or 5, the median number of followers is significantly higher than in other groups, and the distribution is more concentrated. This suggests that linking to 4 or 5 platforms may be a more optimal strategy to significantly increase the number of followers.

Based on the results provided by this dataset, we believe that cross-platform integration can reach a larger potential audience. While there may be overlap among user groups on different platforms, there are also independent segments. Cross-platform integration can cover more users, thereby increasing the number of followers.

4.2.2 Select Optimal Platform Combination

From the above analysis, we conclude that “new bloggers should link as many different platforms as possible.” However, considering the reality that new bloggers may not have enough time and energy to manage so many accounts, we will next analyze how to achieve the highest follower gains with the smallest combination of linked platforms.

First, we presented a box plot of platform combinations and fan counts where all data points are greater than 10 (in the 01 string on the horizontal axis, 1 represents an associated platform at the corresponding position, while 0 represents no association), as shown in the Figure 20.
To select the optimal platform combination, we weigh the options from the following two aspects：

- Analyze the platform combination under high follower data. (Figure 21&22)
  Define ”high follower count data” from two dimensions—mean and median. Select the platform combinations where the median follower count is above the 75th percentile and the mean follower count is above the 75th percentile.
- Analyze platform combinations under high follower count stability data. (Figure 21&22)
  Define a platform combination with high stability in fan numbers using three metrics—interquartile range, outliers, and coefficient of variation. Calculate these three parameters for each group of platform combinations, and define a composite score to select the top six platform combinations with a small interquartile range, few outliers, and a low coefficient of variation.

Take the intersection of all platforms that meet the above conditions, and the optimal platform combination is *Weibo, WeChat, QQ, and mail*.

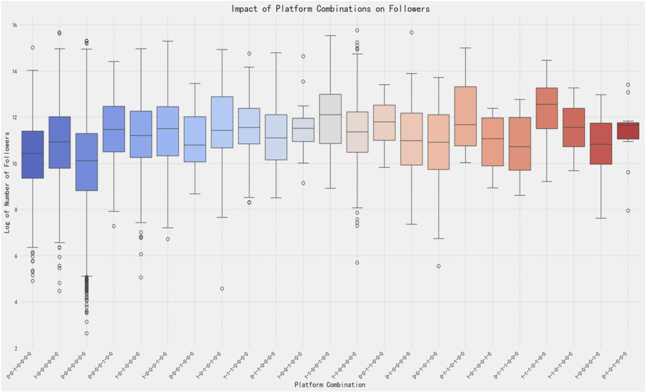


图 20: Boxplot of different Platform Combinations and Number of Followers

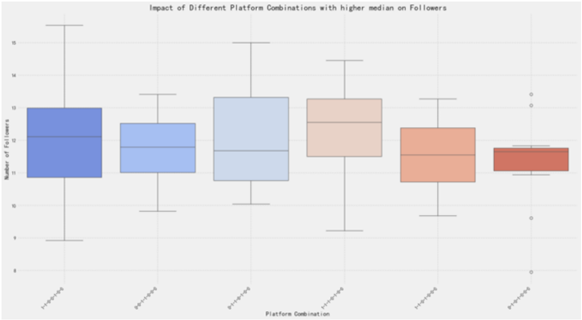


图 21: Box plot of the median follower count for the top 25% platform combinations

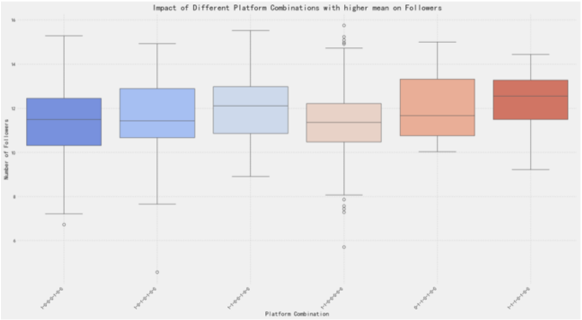


图 22: Box plot of the mean follower count for the top 25% platform combinations

We suggest that new bloggers prioritize linking the above four platforms. This not only makes it easier to recover passwords but also allows them to achieve the highest follower gains with the minimum number of linked platforms.

5 Model prediction

Finally,we are wondering whether we can build a accurate model to predict uploaders' video categories based on their other information. Specifically, we use columns '**follower**', '**sex**','**video**', '**time_ave20**', '**play_ave20**', and '**video_max_ratio**' to predict '**video_tag_combine**' column. We build different models to do the classification task. Results are shown in Figure 23.

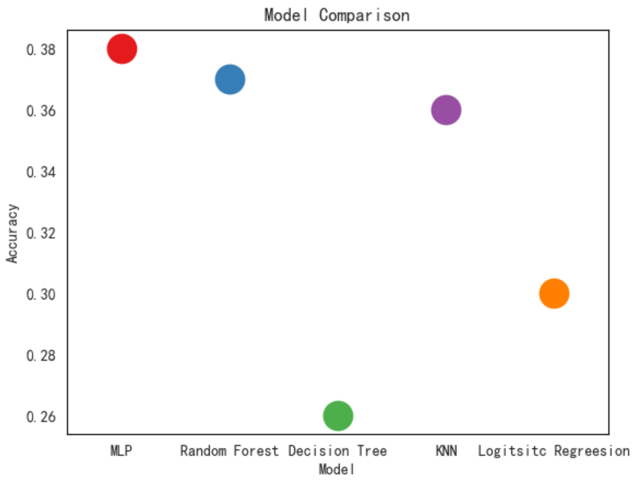


图 23: Performance of different models

To gain slightly better results, we do the alternative processing to data. To use 'sex' column, we encode it using the formula.

$$\text{Encoded}(x) = \begin{cases} \begin{bmatrix}1,0,0\end{bmatrix} & \text{if } x = \text{'男'} \\ \begin{bmatrix}0,1,0\end{bmatrix} & \text{if } x = \text{'女'} \\ \begin{bmatrix}0,0,1\end{bmatrix} & \text{if } x = \text{'保密'} \end{cases}$$

The reason why we encode it in this way rather than encode it into 0, 1 and 2 is that in algorithm like KNN, we are computing the distance between two points. If using 0, 1, 2 as label, the distacne between '男' and '保密' is larger than the distance between '男' and '女' while there is no clear evidence that the difference between '男' and '保密' is larger than '男' and '女'. By encoding it in this way, the distance between every two genders is the same. For 'follower' column, if we directly use it, then in algorithm like KNN, the difference in followers will be highly emphasised. For example, suppose numbers of followers for two different uploaders differ by 100 and the two uploaders' gender are different. In L2 distance,

$$d_{L2} = \sqrt{100^2 + |[1,0,0] - [0,1,0]|^2} \approx 100.01$$

The difference of gender will almost be neglected. To address this problem, we use normalized log followers in models.

$$normalized\ log\_follower = \frac{(log\_follower - \mu)}{\sigma}$$

Through normalization, we can make sure that each variable's contribution to L2 distance are similar.

Assume that Y_1, Y_2 are two normalized variables drawn from the same distribution. Y_1, Y_2 are i.i.d, then

$$E((Y_1 - Y_2)^2) = E(Y_1^2) - 2E(Y_1 * Y_2) + E(Y_2^2) = E(Y_1^2) - 2E(Y_1)E(Y_2) + E(Y_2^2) = 2$$

Similarly, we use normalized time_ave20 and play_ave20 to fit models. Shown in Figure 24, this change helps us gain an 10% increase in accuracy for KNN model. More surprisingly, before processing data, our MLP model can not learn anything through backpropogation. After processing the data, our MLP model gets an accuracy of 39.2%.

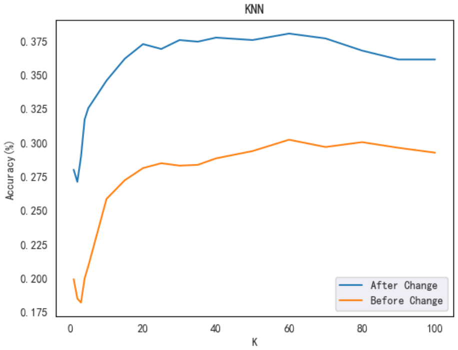


图 24: KNN Improvement

Accuracy(%)	Unprocessed Data	Processed Data
CE Loss	4.6	39.2
Focal Loss(γ=3)	4.6	38.6
Focal Loss(γ=5)	4.6	38.9
Balanced CE Loss	4.6	37.9

表 4: Accuracy of MLP Model

BiliBili data is an imbalanced dataset. We try to use **Balanced Cross Entropy Loss** and **Focal Loss** to mitigate the imbalance and get better results. However, results in Table 4 shows that they do not bring any improvement. Therefore, a limitation for why we can not get better prediction result is that the BiliBili dataset is small and imbalanced.

6 Conclusion

"Game" and "Life" categories dominate BiliBili but are highly competitive. Dance and 'Guichu' offer growth potential, with gender trends highlighting opportunities within these fields. New bloggers should prioritize enhancing the quality of their videos to strive for popularity and in turn, gain a higher number of followers. For different video categories, bloggers should consider the advantages and characteristics of each section to decide whether to increase the diversity of content categories to boost their follower count. As for related platforms, bloggers should aim to connect with as many platforms as possible to achieve high exposure, primarily focusing on platforms like Weibo, WeChat, QQ, and mail as the main associated platforms. We normalize data and attain increase in model's accuracy. We hope our analysis on BiliBili data could offer actionable advice to bloggers.