Table of Contents

Data Mining Project - Team 1

1. Project description

Describe the context and goals of the project.

- The purpose of the project is to analyze YouTube data to identify key factors influencing video popularity in the US, using predictive models to help companies optimize advertising strategies.
- The scope of the project is to analyze a dataset of trending YouTube videos using predictive modeling techniques to identify key factors influencing video popularity and provide actionable insights for optimizing advertising strategies.
- The dataset contains several months of data on daily trending YouTube videos in the United States. This data was collected by YouTube to analyze the top trending videos for the year. The prominent independent variables are Published month, Tags, Video category, Title length, Published time., etc. The dependent variables are the view count and popularity of a video.

Description of the dataset:

The dataset contains several months of data on daily trending YouTube videos in the United States. This data was collected by YouTube, which was used to analyze the top trending videos for the year. Variables that were used to analyze this are listed below:

- **Trending date:** Date on which the video is trending.
- **Title:** Title of the video.
- **Channel title:** Channel that uploaded the video.
- **Category id:** Category (genre) video fits in.
- **Publish time:** Date and time video was uploaded.
- **Tags:** Specialized words that the Youtube algorithm uses to suggest content to individual users.
- **Views:** Total number of views the video got.
- **Likes:** Total number of likes the video got.
- **Dislikes:** Total number of dislikes the video got.
- **Comment count:** Total number of comments the video got.
- **Thumbnail link:** Sharable link to the thumbnail of the video.
- **Comments disabled:** To check if the comment section is disabled or not.
- **Ratings disabled:** To check if the ratings of the video are disabled.
- **Video removed:** To check if the video is still available on the channel.

Business questions

- What factors play a role in the popularity of YouTube videos in the USA?
- What video categories have the most view counts?
- Based on the analysis, what strategy can an advertiser use to pick a YouTube video for advertising their product?

2. Data Preprocessing

Random sampling

- As for random sampling, the original data set had around 40950 rows, from this a random dataset of 10000 rows was chosen using the RAND() function in excel. This is done to make sure there is a proper mix of high and low view count videos.

Handling missing data

- Missing data was found only in the video description column, this column is completely eliminated because the description of the video predominantly does not affect the view count.

Identify numerical variables vs. categorical variables

Before identifying the variables, the raw data needed to be transformed for variables to be usable. Variables like video title, channel name, and video tags were in the form of text in the original dataset. Using excel formulas, these variables were transformed into numerical variables like Title words (number of words in the title), Channel characters (number of characters in the channel name), and Tag words (number of words in the video tags). Other than that, in the raw data, it can be seen that the published time variable was depicted in a single column with published date (yy-mm-dd) and time bundled together. This column was transformed into two categories: Published month (January – December) and published time (AM/PM). After all the transformations, the following are the numerical and categorical variables in the dataset:

Numerical:

- Views – Number of views on a video
- Title_wrods – Number of words in the video title
- Channel_char – Number of characters in the channel name
- Tag_words – Number of words in the video tags
- Likes – Number of likes on the video
- Dislikes – Number of dislikes on the video
- Comment_count – Number of comments on the video

Categorical:

- Publish_month – Month of the year the video is published in
- Publish_time – Time of the data the video is published (AM/PM)
- Comments_disabled – If the comments are disabled on the video (true/false)
- Ratings_disabled – If the ratings are disabled on the video (true/false)
- Category_id – This number pertains to different categories that the video falls in

Category ID	Category
1	Film & Animation
2	Autos & Vehicles
10	Music
15	Pets & Animals
17	Sports
19	Travel & Events
20	Gaming
22	People & Blogs
23	Comedy
24	Entertainment
25	News & Politics
26	Howto & Style
27	Education
28	Science & Technology
29	Nonprofits & Activism
43	Shows

The table shows the different categories that pertain to their respective category IDs

Summary characterizes:

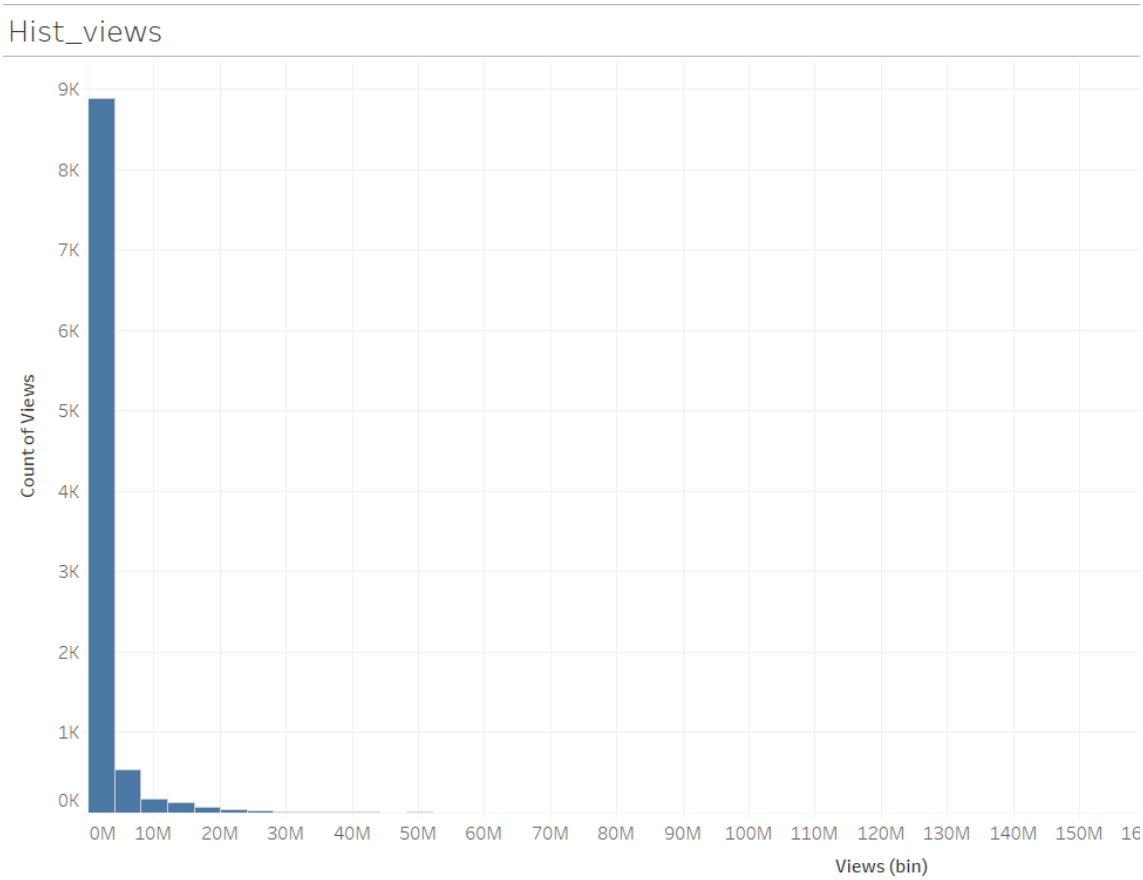
	Mean	**Median**	**Std**	**Outliers**
Log Views	5.784	5.828	0.745	141
Likes	73568.43	18472.5	220947.5	150
Dislikes	3437.995	638	20342.52	70
Comment Count	8190.24	1884	33739.68	70
Title Words	8.4502	8	3.470637	49

Correlation table:

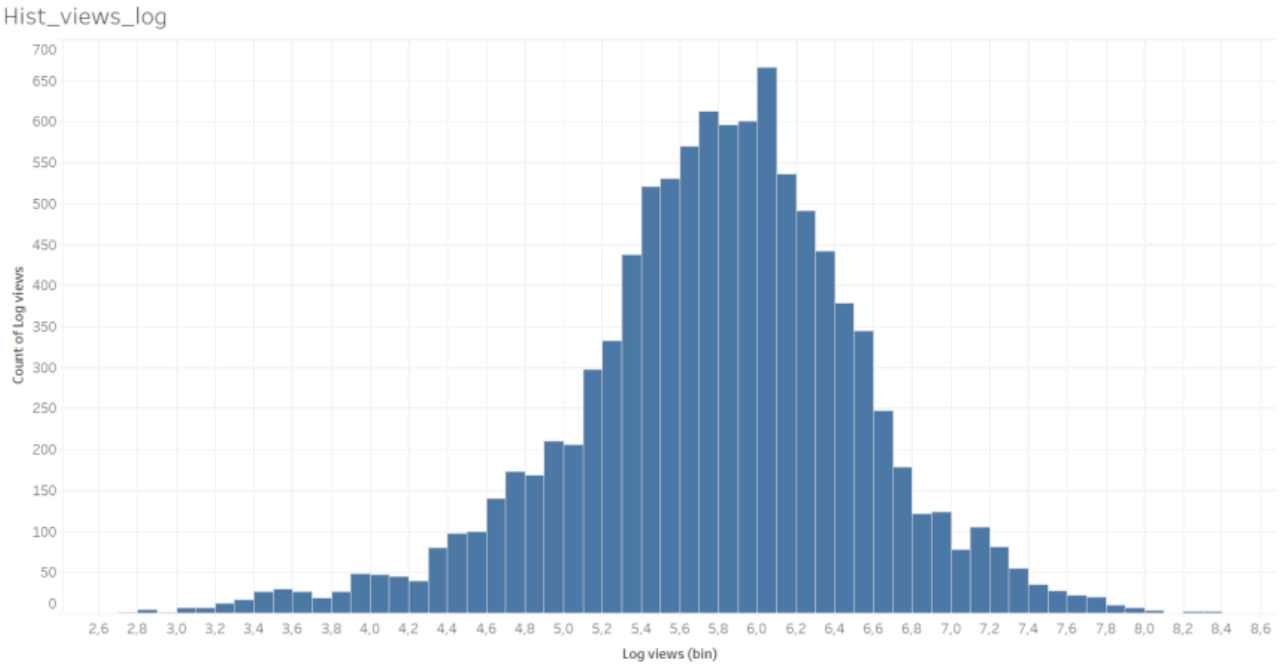
	views	*Title_words*	*Channel_Char*	*Tag_words*
views	1			
Title_words	-0.0349	1		
Channel_Char	0.0362	-0.0067402	1	
Tag_words	0.045664	0.2550787	-0.01019663	1

- From the correlation table, it can be seen that the numerical variables have small influence on the dependent variable. Within that influence, the Title_wrods has a negative correlation, which means that a greater number of words in the title lower the view count. And the other variables have a positive correlation which means higher channel name characters and words in the video tag would relate to higher view count. It needs to be noted that this is not very significant as seen in the table.

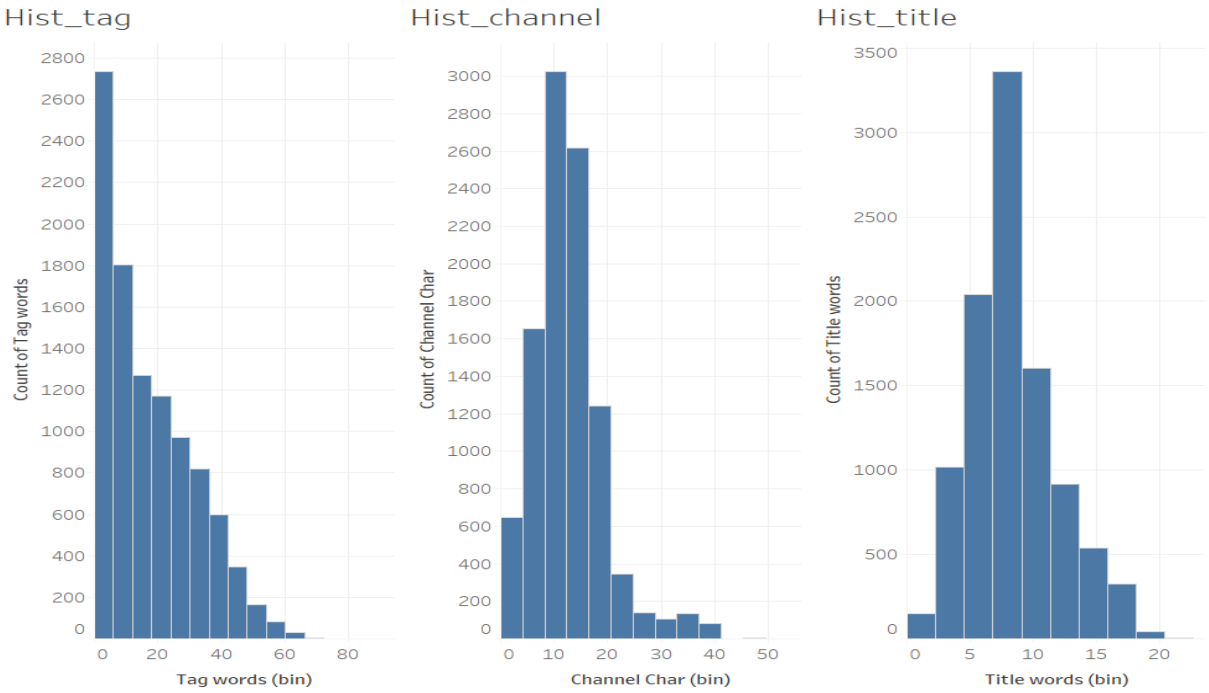
Histogram, scatterplot, boxplot:

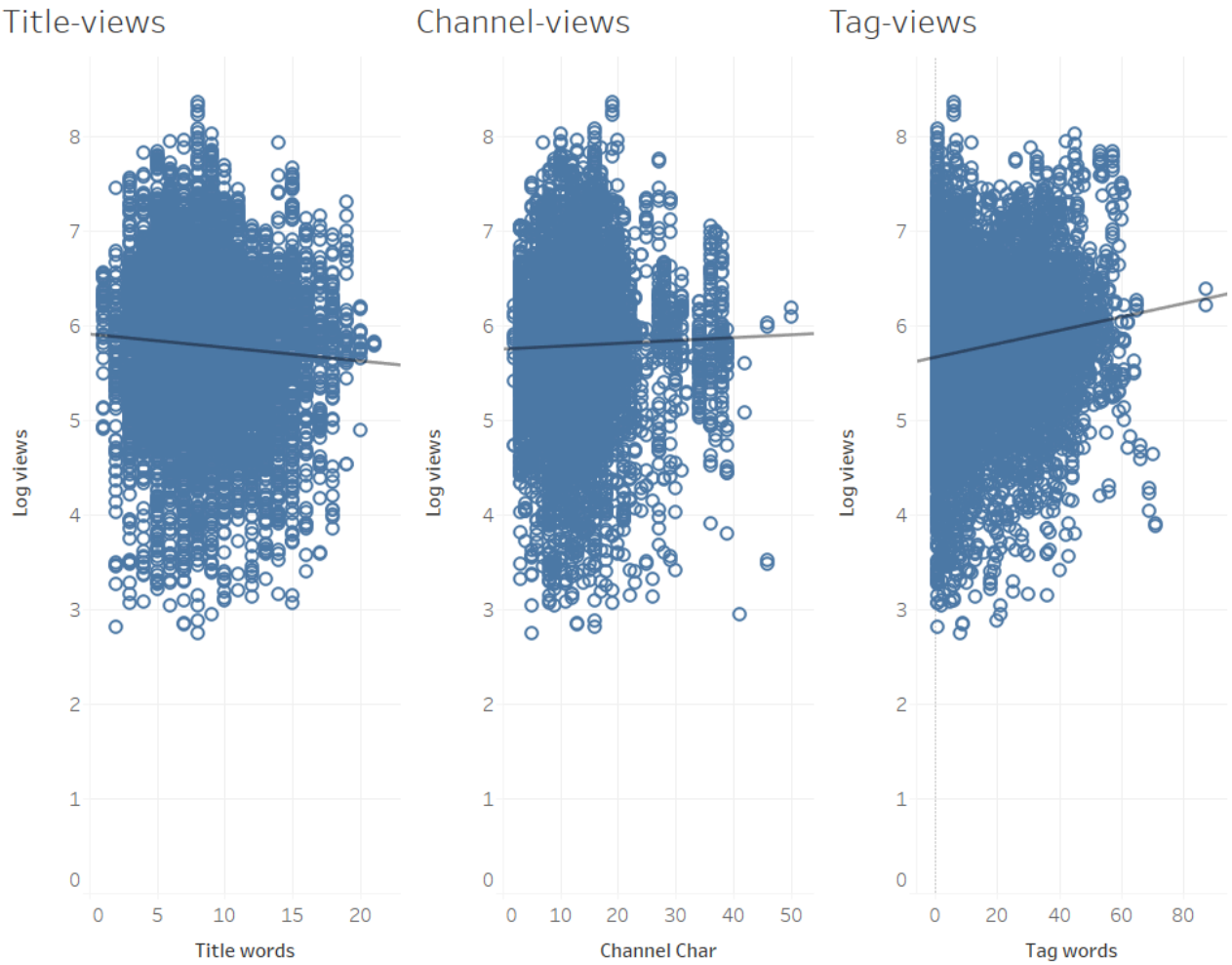


- As it can be clearly seen, the view count column needs to be normalized in order to have a better representation because the numerical differences between the view counts of different videos is very high which in turn is resulting in the skewed representation.
  The normalized view count with logarithmic transformation to the base 10 would give a histogram as follows:



- This shows a normal distribution with a better representation of the data. The other histograms of the numerical variables can be seen in the following picture. Each of the variables has few outliers which are essentially not errors but just extreme values which are important for analysis.

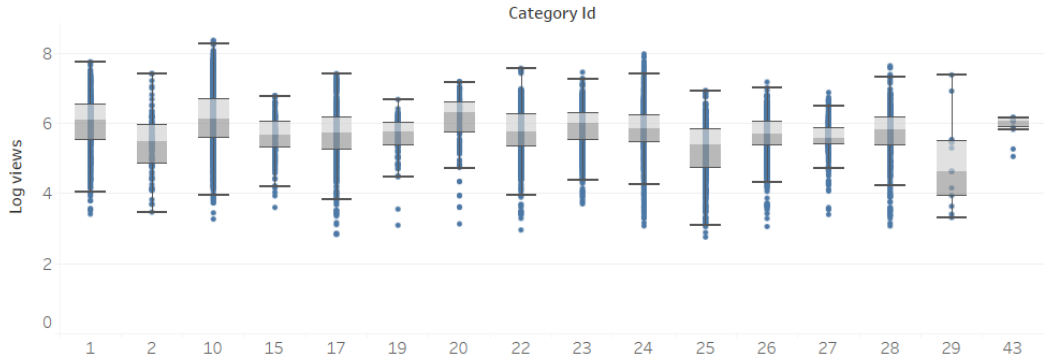




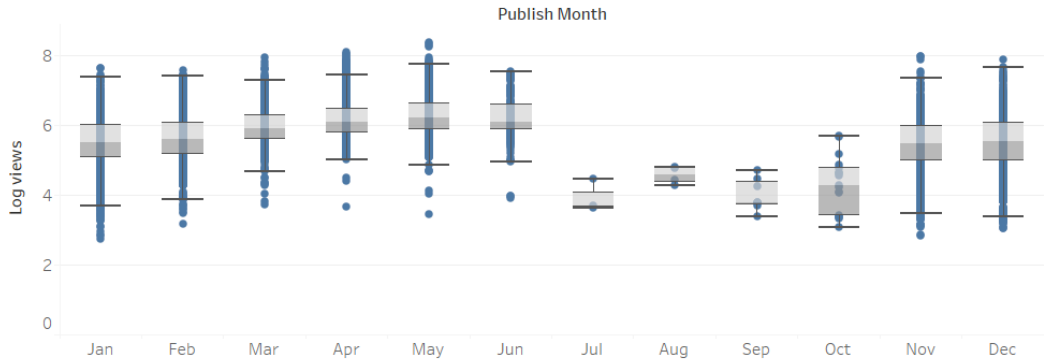
- As seen in the correlation table, the scatter plots shown above, between the numerical variables and the dependent variable show a similar relationship. Where an increase in the title words slightly decreases the view count, and an increase in the Channel name characters and tag words slightly increases the view count.

- The box plots of the view count in relation to the category ID and the month published is shown below. It can be seen that category 10 which is the music category has the highest view count among the others, and the summer months of April, May and June have the higher view counts as compared to the other months.

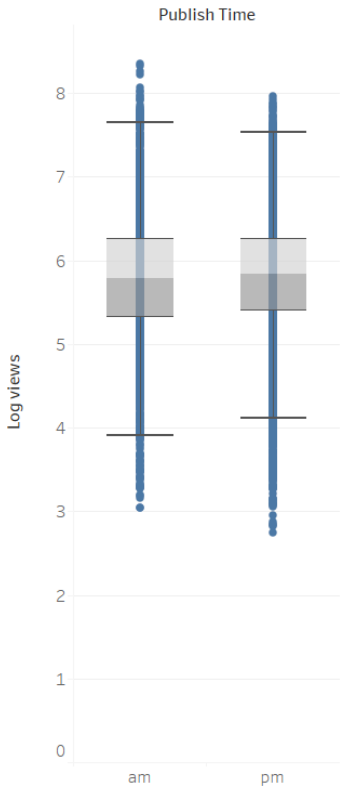
Category-views



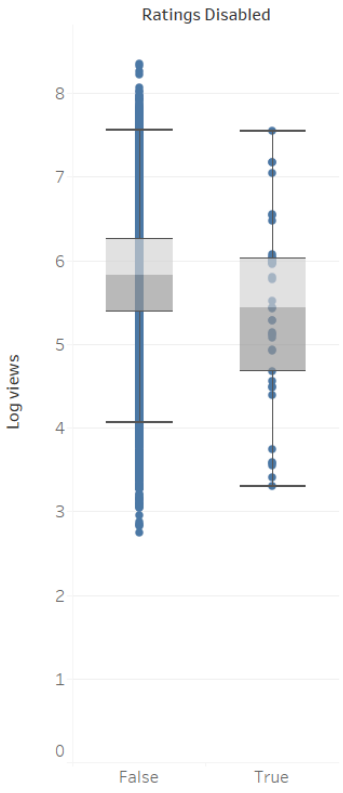
month-views



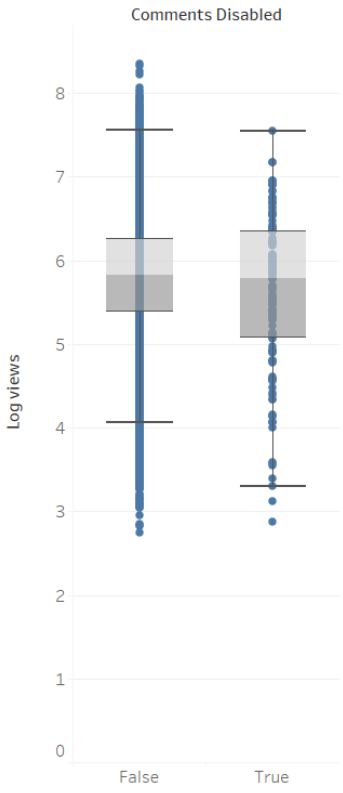
time-views



Ratings-views



Comments-views



- The above box plots show the relationship between the view count and the other categories like time published, comments disabled and ratings disabled. It can be seen that the median view counts for each of these categories are pretty similar, with the videos published in the AM having a slightly higher
- view count. And as expected, the videos with comments and ratings not disabled have a higher median view count.
- From the previous plots, detect outliers and find out whether they are errors or extreme values
- As seen in the histograms and the box plots, there are a few outliers, but they are not errors, but just extreme values in each category that have an effect and are important for the analysis.

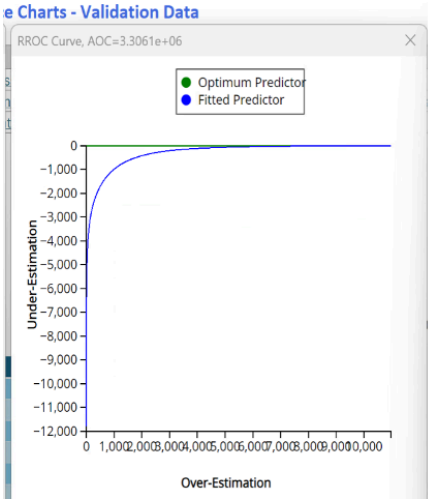
3. Models

A. Linear Regression:

- First model used would be a linear regression model that draws the relationship between all the independent variables and the dependent variable which is the view count.
- For the classification model, the view count variable is transformed into a categorical variable (Popular: 1/0). This is done by first normalizing the view count data through log transformation and classifying view counts above the mean as 1 and below mean as 0.
- These variables were selected based on their relevance to the business questions, availability in the dataset, and statistical significance. They capture a mix of content features (e.g., title words, tags), ensuring comprehensive coverage of factors influencing video popularity.
- The feature selection process ensured that only statistically significant and meaningful variables were included in the model. The output reveals key factors influencing video performance, such as title length, tags, categories, upload months, and ratings, providing clear guidance for content creators and advertisers.

Model Output:

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	-0.4587285	-0.621618379	-0.295838715	0.632086804	0.083108584	30.46630358	3.39716E-08
Title_words	-0.0551948	-0.07200559	-0.038384054	0.94630077	0.00857708	41.41119434	1.23349E-10
Tag_words	0.01974789	0.01554198	0.023953802	1.019944171	0.002145912	84.68713574	3.49517E-20
category_id_1	0.70015264	0.461101663	0.939203615	2.014060108	0.121967025	32.95342808	9.43932E-09
category_id_10	0.55932897	0.400973635	0.717684301	1.749498137	0.080795022	47.92536392	4.42757E-12
category_id_20	0.88861796	0.419388293	1.357847633	2.431766537	0.239407292	13.77702441	0.000205838
category_id_26	-0.5192537	-0.71221241	-0.326294966	0.594964411	0.098450137	27.81804041	1.33278E-07
category_id_27	-0.9437607	-1.242470731	-0.645050669	0.38916156	0.152405878	38.34602447	5.92487E-10
publish_month_Apr	1.58356939	1.400549072	1.766589705	4.872316	0.093379428	287.5886712	1.66713E-64
publish_month_Jun	1.69938493	1.192558937	2.206210917	5.47058155	0.25858944	43.18788652	4.97276E-11
publish_month_Mar	0.92510429	0.767679673	1.082528909	2.522131282	0.080320159	132.6576427	1.07428E-30
publish_month_May	1.74515595	1.574234991	1.91607691	5.726794502	0.087206174	400.4737203	4.34322E-89



Validation: Prediction Summary

Metric	Value
SSE	1653.60553
MSE	0.41340138
RMSE	0.64296297
MAD	0.49619973
R2	0.26617618

- **Model equation**: Y(log_views) = -0.4587-0.055*title_words+0.019*tag_words+0.7*Cat_1+0.559*cat_10+0.88*cat_20-0.519*cat_26-0.943*cat_27+1.58*april+1.69*June+0.92*Mar+1.74*may
- The model explains about **26.6% (R^2)** of the variation in video views, indicating a moderate fit.
- **Title words:** If the words in the title increases by one, the log of view count decreases by 0.055, holding other variables constant.
- **Tag words**: If the number of tags on the video increases by one, the log of view count increases by 0.019, holding other variables constant.
- **Category 1**: If a video is from category 1, the log of view count increases by 0.7 as compared to the categories not present in the model equation, holding other variables constant.
- **Category 10:** If a video is from category 10, the log of view count increases by 0.559 as compared to the categories not present in the model equation, holding other variables constant.
- **Category 20**: If a video is from category 20, the log of view count increases by 0.88 as compared to the categories not present in the model equation, holding other variables constant.
- **Category 26**: If a video is from category 26, the log of view count decreases by 0.52 as compared to the categories not present in the model equation, holding other variables constant.
- **Category 27**: If a video is from category 27, the log of view count decreases by 0.943 as compared to the categories not present in the model equation, holding other variables constant.
- **April**: If a video is published in the month of April, the log of view count increases by 0.92 as compared to the other months, holding other variables constant.
- **June:** If a video is published in the month of June, the log of view count increases by 1.69 as compared to the other months, holding other variables constant.
- **March**: If a video is published in the month of March, the log of view count increases by 1.58 as compared to the other months, holding other variables constant.
- **May**: If a video is published in the month of May, the log of view count increases by 1.7 as compared to the other months, holding other variables constant.

The results provide actionable insights for both content creators and advertisers. By focusing on high-performing categories, leveraging peak months, and optimizing video features, advertisers can effectively target audiences and maximize campaign ROI.

A Hypothetical Example:

Suppose we have a new video with the following characteristics:

- Title_words: 6 (Concise title with six words)
- Tag_words: 15 (Moderate number of relevant tags)
- Category_id: 10 (Music category)
- Publish_month: May (Peak month for video views)
- Ratings_disabled: 0 (Ratings are enabled)
- Log(Views)=5.1492+(−0.1122)+(0.0885)+(0.7493)+(0) =5.8748
- Views≈356,242 views (Based on the model, this video is predicted to achieve approximately **356,242 views**.)
- This example demonstrates how the model predicts the number of views and classifies the video based on predefined criteria. The strong influence of tags, category, and publish month highlights the importance of optimizing these features for video success.

B. Logistic regression

- Logistic regression was chosen to address the need for classifying YouTube videos into "Popular" or "Not Popular" categories, which aligns with one of the key business questions:
- What strategy can an advertiser use to pick a YouTube video for advertising their product?"
- Logistic regression is a widely-used statistical method for classification problems where the dependent variable is categorical (e.g., `1 = Popular`, `0 = Not Popular`).
- These variables were selected because they directly influence video popularity, aligning with the business goal of identifying factors affecting YouTube trends.

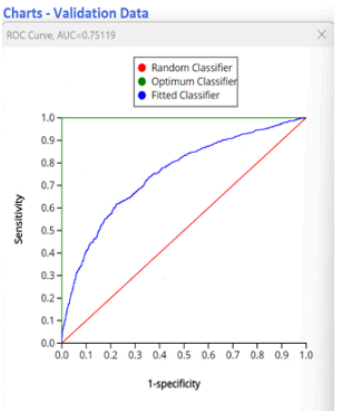
Intercept
Title_words
Channel_Char
Tag_words
category_id_1
category_id_10

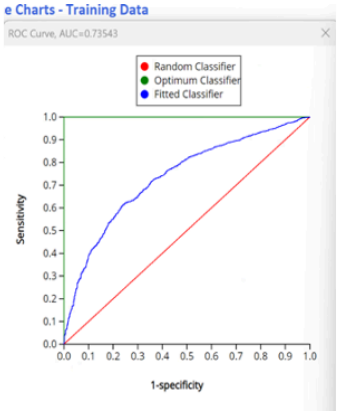
category_id_15
category_id_17
category_id_19
category_id_2
category_id_20
category_id_26
category_id_27
publish_month_Apr
publish_month_Jun
publish_month_Mar
publish_month_May
publish_time_pm
comments_disabled_-1
ratings_disabled_-1

- The feature selection process ensured that only statistically significant and meaningful variables were included in the model. The output reveals key factors influencing video performance to determine the popularity of videos.

Model Output:

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	-0.4587285	-0.621618379	-0.295838715	0.632086804	0.083108584	30.46630358	3.39716E-08
Title_words	-0.0551948	-0.07200559	-0.038384054	0.94630077	0.00857708	41.41119434	1.23349E-10
Tag_words	0.01974789	0.01554198	0.023953802	1.019944171	0.002145912	84.68713574	3.49517E-20
category_id_1	0.70015264	0.461101663	0.939203615	2.014060108	0.121967025	32.95342808	9.43932E-09
category_id_10	0.55932897	0.400973635	0.717684301	1.749498137	0.080795022	47.92536392	4.42757E-12
category_id_20	0.88861796	0.419388293	1.357847633	2.431766537	0.239407292	13.77702441	0.000205838
category_id_26	-0.5192537	-0.71221241	-0.326294966	0.594964411	0.098450137	27.81804041	1.33278E-07
category_id_27	-0.9437607	-1.242470731	-0.645050669	0.38916156	0.152405878	38.34602447	5.92487E-10
publish_month_Apr	1.58356939	1.400549072	1.766589705	4.872316	0.093379428	287.5886712	1.66713E-64
publish_month_Jun	1.69938493	1.192558937	2.206210917	5.47058155	0.25858944	43.18788652	4.97276E-11
publish_month_Mar	0.92510429	0.767679673	1.082528909	2.522131282	0.080320159	132.6576427	1.07428E-30
publish_month_May	1.74515595	1.574234991	1.91607691	5.726794502	0.087206174	400.4737203	4.34322E-89





Validation

Metrics	
Metric	Value
Accuracy (#correct)	2721
Accuracy (%correct)	68.19549
Specificity	0.712553
Sensitivity (Recall)	0.65425
Precision	0.715405
F1 score	0.683462
Success Class	1
Success Probability	0.5

Training

Metrics	
Metric	Value
Accuracy (#correct)	4034
Accuracy (%correct)	67.4131
Specificity	0.704042
Sensitivity (Recall)	0.647021
Precision	0.706927
F1 score	0.675649
Success Class	1
Success Probability	0.5

The logistic regression equation for predicting the probability of a video being "Popular" (PPP) can be written as:

Logit(y=1): -0.4587 - 0.0552 * Title_words + 0.0197 * Tag_words + 0.7002 * Category_id_1 + 0.5593 * Category_id_10 + 0.8886 * Category_id_20 - 0.5193 * Category_id_26 - 0.9438 * Category_id_27 + 1.5836 * Publish_month_Apr + 1.6993 * Publish_month_Jun + 0.9251 * Publish_month_Mar + 1.7452 * Publish_month_May

Odds = EXP(Logit:y=1)

Probability (popular = 1): Odds/(1+Odds)

Coefficient Interpretations

- Title_words: If the title increases in length by one word, the odds of the video being popular gets multiplied by 0.94 (decreases), holding other variables constant.
- Tag_words: If the number of tags on the video increases by one, the odds of the video being popular gets multiplied by 1.02 (increases), holding other variables constant.
- Category_1: The odds of a video in category 1 being popular is 2.01 times the odds of other categories, holding other variables constant.
- Category_10: The odds of a video in category 10 being popular is 1.75 times the odds of video from other categories, holding other variables constant.
- Category_20: The odds of a video in category 20 being popular is 2.43 times the odds of video from other categories, holding other variables constant.
- Category_26: The odds of a video in category 26 being popular is 0.59 times the odds of video from other categories, holding other variables constant.
- Category_27: The odds of a video in category 27 being popular is 0.38 times the odds of video from other categories, holding other variables constant.
- April : The odds of a video published in April being popular is 4.87 times the odds of a video published in other months, holding other variables constant.
- June: The odds of a video published in June being popular is 5.47 times the odds of a video published in other months, holding other variables constant.
- March: The odds of a video published in March being popular is 2.52 times the odds of a video published in other months, holding other variables constant.
- May: The odds of a video published in May being popular is 5.72 times the odds of a video published in other months, holding other variables constant.

- The model performs consistently on training and validation datasets, with good precision and recall.
- AUC values indicate strong classification capability, suggesting that the model is reliable for predicting video popularity.
- The R^2 value (from prior linear regression analysis) indicates room for improvement in explaining variability.
- Some predictors (e.g., Title_words) have a small effect, suggesting the need for additional features.
- By focusing on high-performing categories and peak months, advertisers can optimize their campaigns for maximum reach.
- Simple adjustments, such as using concise titles and relevant tags, can significantly boost video performance.
- Advertisers should align their strategy with data-driven factors to effectively target their audience and maximize ROI.

A Hypothetical Example:

Here's the hypothetical data record:

- Title_words = 7
- Tag_words = 20
- Category_id_10 = 1 (Music category; all other categories = 0)
- Publish_month_May = 1 (Month = May; all other months = 0)

Other categories (e.g., Category_id_1, Category_id_26, Category_id_27) and months (e.g., Publish_month_Apr, Publish_month_Jun) are set to 0 because they don't apply to this record.

- Logit(popular=1): $-0.4587-0.3864+0.3940+0.5593+1.7452 = 1.8534$
- Odds(Popular = 1): EXP(1.8534) = 6.3814
- P(Popular=1): Odds/(1+odds) = 6.3814/7.3814 = 0.8645

Probability: 86.45%

Classification: **Popular** (since P = 0.8645 > 0.5).

C. Decision tree

- For the classification model, the view count variable is transformed into a categorical variable (Popular: 1/0). This is done by first normalizing the view count data through log transformation and classifying view counts above the mean as 1 and below mean as 0.
- For the classification problem, tree models will be used to draw the relationships and rules that predict the popularity of a video.
- All the numerical and categorical variables are selected with the dependent variable being popular (1/0).
- The variables Log views, Log Likes, dislikes and comment count are left out because they are the outcomes of a popular video.
- The category ID variable has 16 categories which crosses the limit (15) for XL Miner to analyze categorical variables. So, two categories 29 and 43 have been removed as they have negligible number of data rows. The updated category table is as follows:

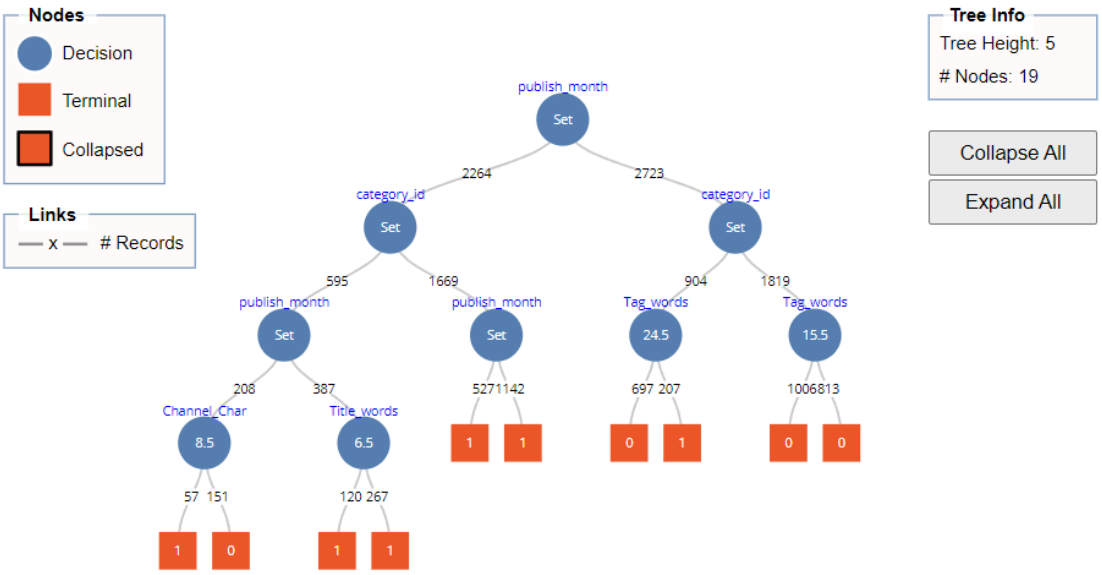
  Category ID-Category
  1-Film & Animation
  2-Autos & Vehicles
  10-Music1
  5-Pets & Animals
  17-Sports
  19-Travel & Events
  20-Gaming
  22-People & Blogs
  23-Comedy
  24-Entertainment
  25-News & Politics
  26-Howto & Style
  27-Education
  28-Science & Technology

- Since it is a classification tree model, the variable selection process is automatic, and all the variables are fed into the model differentiated by numerical and categorical.

Model Output:

- Different iterations have been run for the classification tree by tweaking the threshold for minimum number of records in leaves to 500, 250, 100 and 50.
- Out of these models, both the limit 100 and 50 models performed equally well, while the model with the limit set to 50 presented a better tree that clearly differentiates both the classes at terminal ends.
- The full tree and the best tree chosen is as follows:

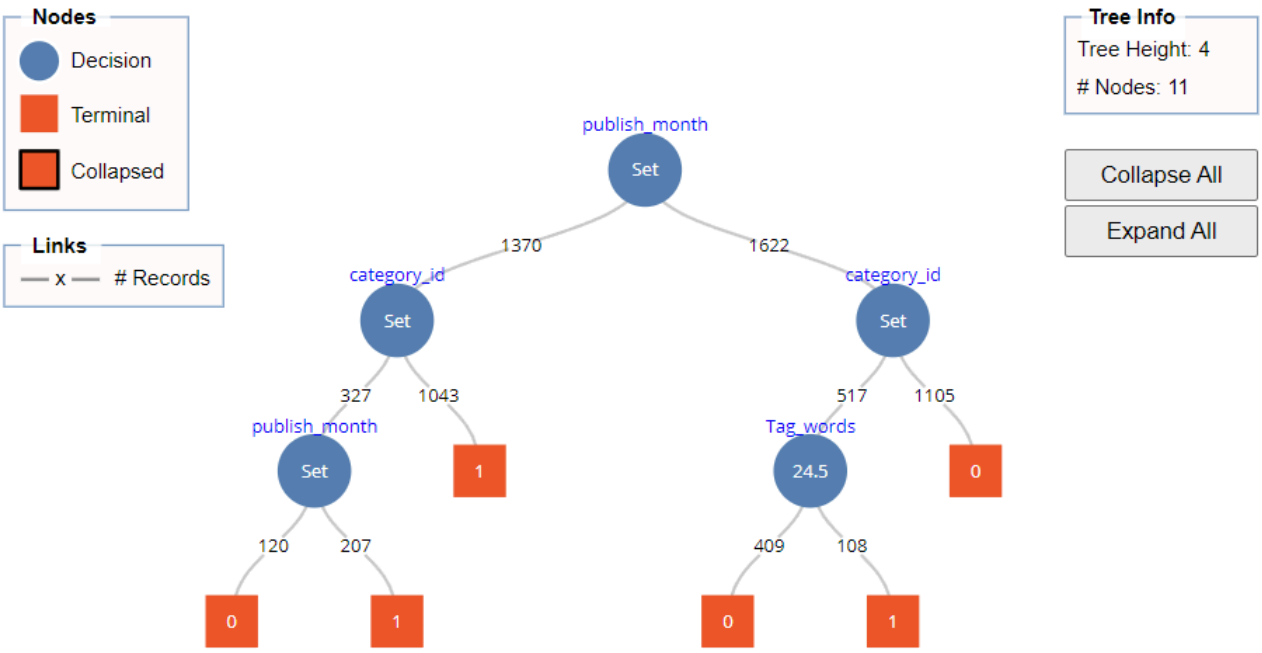
Full tree:



Best tree:

Classification rules based on the best pruned tree are as follows:

- If a video is published in the months {March, April, May, June} and is not from the categories {2, 15, 17, 25, 26, 27}, it is predicted to be popular.



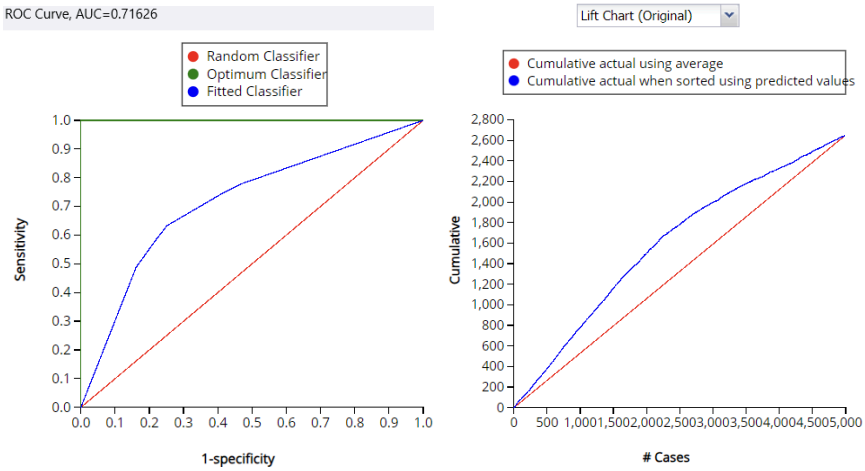
- A video published in the months {March, April, may, June} and is from the categories {2, 15, 17, 25, 26, 27} is predicted to be popular if not published in the month of March and not popular if published in the month of March.
- A video that is not published in the months {March, April, May, June} and is not from the categories {1, 10, 19, 20, 23} will not be popular.

- A video that is not published in the months {March, April, May, June} and is from the categories {1, 10, 19, 20, 23} is predicted to be popular if the number of tags on the video is more than or equal to 24.5 and not popular if the number of tags is less than 24.5.

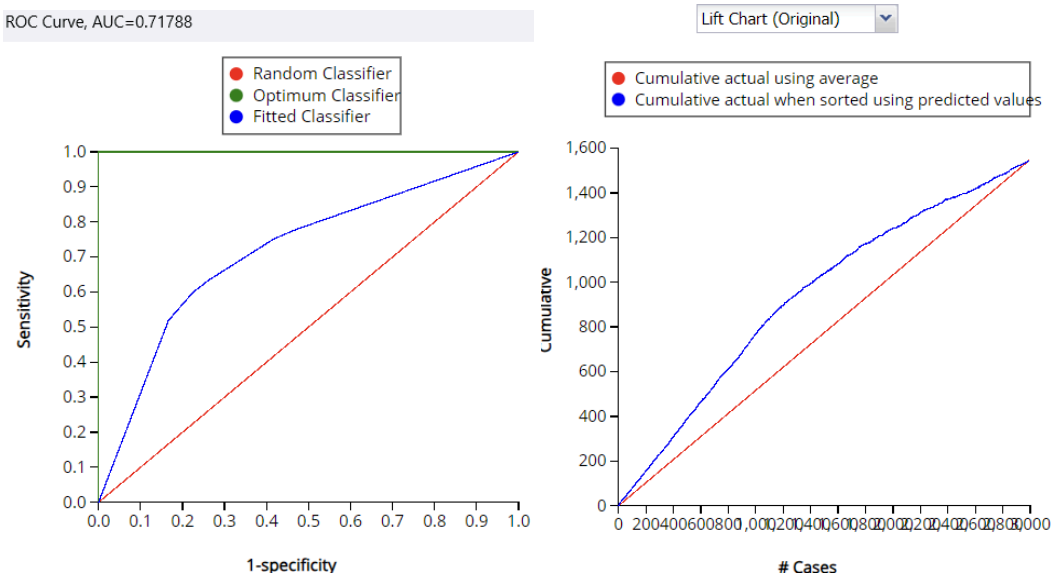
Summary report for training, validation, and test data

Metrics - Validation data		Metrics - Training data		Metrics - Test data	
Metric	Value	Metric	Value	Metric	Value
Accuracy (#correct)	2047	Accuracy (#correct)	3427	Accuracy (#correct)	1394
Accuracy (%correct)	68.41578	Accuracy (%correct)	68.71866854	Accuracy (%correct)	69.87469
Specificity	0.738079	Specificity	0.748505551	Specificity	0.762605
Sensitivity (Recall)	0.633657	Sensitivity (Recall)	0.63289225	Sensitivity (Recall)	0.64046
Precision	0.720913	Precision	0.739726027	Precision	0.747204
F1 score	0.674475	F1 score	0.682151589	F1 score	0.689726
Success Class	1	Success Class	1	Success Class	1
Success Probability	0.5	Success Probability	0.5	Success Probability	0.5

Lift charts – Training data

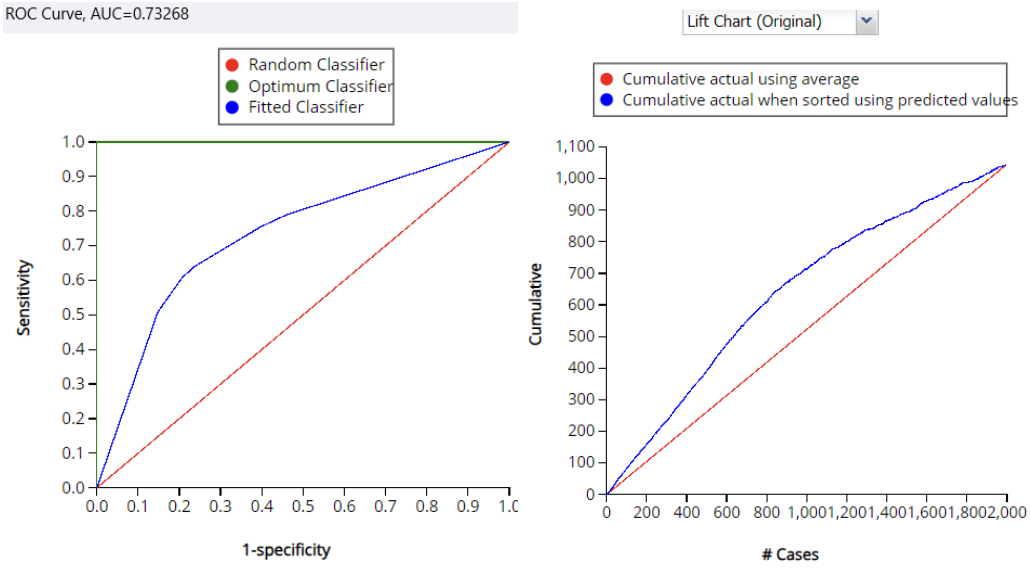


Lift charts – Validation data



As indicated in the summary tables, the model shows good results equally on all training, validation and test data with an average overall error rate of 31%, and accuracy of 69%. The model presents precision and recall values of 74% and 64% respectively, and an F1 score of 68%. As can be seen in the lift charts, the model performs well on all datasets with a good AUC score of 72%.

Lift charts – Test data



1. **For classification models, determine an appropriate cutoff value based on your results. Run the model with alternative cutoff values and compare performance.**

- The results produced above have a standard cutoff value of 0.5. Based on the decile lift chart from the validation data, it was observed that the first 40% of data had a decile/ global mean greater than 1. Based on this, the new cutoff would be around 0.65 which is an increase from the standard value. Changing this cutoff would in turn increase the error rate in class 1 and reduce precision and recall, with not much difference in the F1 score and the AUC performance. So, the cutoff remains at 0.5 for better model performance.

2. **Explain how the results address your business questions.**

- Based on this classification tree, the results address the business questions effectively. It shows that the two main variables that have a significant effect on the popularity of a video on YouTube are published month and the video category.
- It also shows that the videos from categories {1, 10, 19, 20, 23} have the higher chance of being popular. These categories are Film, music, travel & events, gaming and comedy.
- It also clearly shows that the engagement disabled, and the title length does not have a significant effect on the popularity of a video, but the number of tags on the video have some effect on the popularity.

3. **Offer a hypothetical example of a new data record and demonstrate prediction or classification.**
    - A hypothetical example of a new data record would be: A video published in the month of May, with a view count of 1655033, in category 28, with 8 characters in the title, 15 tags on the video, published in the evening and comments not disabled:
    - Based on the best tree, the classification for this data record would be 1, and the video is predicted to be popular.

D. Neural Network

1. **Identify the models used and provide a rationale for each selection.**

- The classification and prediction neural network model was chosen for its ability to capture complex, non-linear relationships between features such as tags, categories, and publish months, which are essential for accurately predicting video popularity. Its adaptability and potential for higher accuracy compared to simpler models like logistic regression make it well-suited for this dataset and business objectives.

2. **Specify the variables included and justify their choice.**

- Tag_words
- Category_id_1
- Category_id_10
- Category_id_19
- Category_id_20
- Publish_month_Apr
- Publish_month_Jun
- Publish_month_May
- Publish_time_pm
- Popular (Output Variable)
- These variables directly relate to key business questions, such as the impact of category, timing, and video optimization on popularity.
- The chosen variables—spanning tags, categories, publish timing, and audience behavior—are critical to understanding and predicting video popularity. This selection is backed by their relevance to the problem, and feature selection based on the previous logistic regression and decision tree models.

3. **Describe any variable selection techniques applied.**

- The variables selected for the neural network models are based on all the shortlisted variables that were filtered through feature selection in the previous logistic regression and decision tree models.

4. **Present the model output, including equations (if applicable) and coefficient interpretations (if relevant). For tree-based or neural network models, explore various parameters and select the best-performing model.**

Classification Neural Network:

The best performing neural network is selected by first applying an automatic neural network and selecting the best performing one based on the Validation F-1 score, non zero sensitivity and specificity. Multiple neural networks with these conditions are selected and run to select the best one based on a balance between the error rates, and high F-1 scores.

NetID	# Hidden Layers	# Neurons (Layer 1)	# Neurons (Layer 2)	Training # Errors	Training % Error	Training % Sensitivity	Training % Specificity	Training % Precision	Training % F1-Score	Validation # Errors	Validation % Error	Validation % Sensitivity	Validation % Specificity	Validation % Precision	Validation % F1-Score
Net 37	2	5	6	2849	47.6102942	99.87257	0	52.424475	68.757539 2	1897	47.543866	99.90449	0.052743	52.47053	68.804447
Net 48	2	7	5	2844	47.5267379 7	99.07614	1.054482	52.489455	68.623124 45	1888	47.3183	99.28367	1.21308	52.60628	68.777274
Net 19	2	2	6	2848	47.5935828 9	95.12584	5.272408	52.561117	67.709750 57	1904	47.7193	95.31996	4.746835	52.498688	67.706922
Net 13	2	1	6	2908	48.5962566 8	95.73112	2.495606	51.999862	67.391791 88	1930	48.3709 3	95.84527	2.795359	52.12987	67.530 28
Net 5	1	5	0	2946	49.2312834 2	93.2144	3.936731	51.70 52	66.51 5117 07	1948	48.8220	93.69628	4.219409	51.93 22	66.8256

								5			6			4	1
Net 1	1	1	0	2863	47.8442513 4	89.45524	11.00176	52.584427	66.234225 73	1909	47.844461	90.11461	10.23207	52.57732	66.408589
Net 12	2	1	5	2978	49.7660427 8	93.2144	2.811951	51.414451	66.274065 69	1983	49.699225	93.12321	3.006329	51.464477	66.292771
Net 7	1	7	0	3023	50.5180481 3	91.11182	3.550088	51.034498	65.423767 59	1998	50.075199	92.07259	3.375527	51.2766	65.869499
Net 49	2	7	6	2972	49.6657754	93.08697	3.163445	51.470885	66.288566 24	2016	50.526322	91.78606	2.742616	51.035558	65.597277
Net 6	1	6	0	3327	55.5982620 3	77.50876	7.873462	48.140009	59.392164 04	2267	56.817044	75.69245	7.278481	47.4125	58.304221
Net 21	2	3	2	3389	56.6343582 9	71.39216	12.44288	47.358441	56.943209 25	2276	57.042611	71.06017	11.91983	47.11843	56.666413
Net 27	2	4	2	3367	56.2667112 3	69.70373	15.07909	47.523889	56.515562 44	2255	56.516299	69.48424	14.76793	47.3787	56.340076
Net 9	2	1	2	2403	40.1570855 6	39.98089	81.75747	70.744088	51.088947 69	1600	40.100225	39.87584	82.01477	71.00034	51.070334

Net 33	2	5	2	3328	55.6149732 6	49.69736	38.52373	47.14415	48.387096 77	2258	56.59148	47.99427	38.34388	46.22815	47.094666
Net 2	1	2	0	2923	48.8469251 3	35.20229	68.7522	55.416255	43.054743 81	1947	48.79699	35.95989	68.03797	55.40839	43.614255
Net 39	2	6	2	3029	50.6183155 1	37.24116	62.7768	52.46858	43.562511 65	2060	51.62907	35.67335	62.39451	51.16438	42.037144
Net 45	2	7	2	3310	55.3141711 2	38.9296	51.03691	46.73004	42.474800 14	2256	56.54135	36.81948	50.79114	45.24648	40.600302

The best performing Neural network is as shown below:

Neuron Weights

Neuron Weights: Input Layer - Hidden Layer 1

Neuron	Tag_word	category_id_	category_id_1	category_id_1	category_id_2	publish_month_A	publish_month_J	publish_month_M	publish_time_pr	Bias
Neuron 1	-0.06978271	0.207776444	-0.775438423	0.304779035	0.311404492	0.281747379	-0.176521741	0.347007077	-0.1095147	0.001841
Neuron 2	0.49831376	-0.289485749	-0.201336799	0.058326086	-0.051382976	-0.324224028	-0.106210997	-0.370810106	0.377853758	-0.00311
Neuron 3	-0.01161574	-0.193782242	0.077581145	0.099786164	-0.305047357	0.139720278	0.060187314	-0.225130836	-0.107288039	0.012209
Neuron 4	0.21407605	0.576324608	0.213670058	-0.443507697	0.030745655	0.682521994	0.211985852	-0.046318792	0.063723758	-0.00556
Neuron 5	-0.02379714	-0.335120986	-0.366548141	-0.273050911	0.182598834	-0.498412833	-0.298505166	0.267395137	-0.139834222	0.001713
Neuron 6	-0.17460901	-0.671491303	-0.047108532	-0.136592087	-0.645216776	0.595825279	0.000151403	0.641712815	0.164747104	0.005852

Neuron Weights: Hidden Layer 1 - Output Layer

Neuron	Neuron 1	Neuron 2	Neuron 3	Neuron 4	Neuron 5	Neuron 6	Bias
0	-0.15256469	0.246332632	-0.955861284	0.273346285	0.354749607	0.250095962	-0.082724377
1	-0.17944464	0.398660089	-0.152270461	0.656701176	-0.348519138	-0.181843606	0.019925602

Confusion Matrix

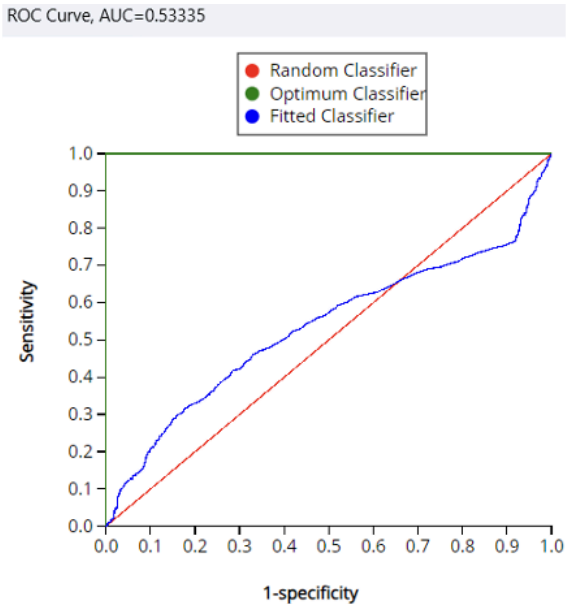
Actual\Predicted	0	1
0	1018	878
1	978	1116

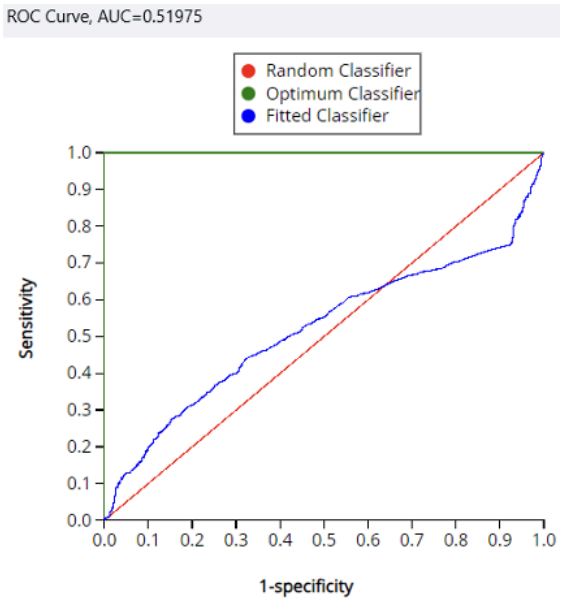
Metrics

Metric	Value
Accuracy (#correct)	2134
Accuracy (%correct)	53.48371
Specificity	0.53692
Sensitivity (Recall)	0.532951
Precision	0.559679
F1 score	0.545988
Success Class	1
Success Probability	0.519067

Error Report

Class	# Cases	# Errors	% Error
0	1896	878	46.30801688
1	2094	978	46.70487106
Overall	3990	1856	46.51629073





- The AUC values are slightly above 0.5, indicating the model performs marginally better than random classification. The ROC curves show limited ability to distinguish between popular and non-popular videos.
- The neural network captures some non-linear relationships but performs moderately, as reflected in the accuracy (53.48%) and AUC values (~0.52). Overall, the model performs poorly, and cannot work well for classifying videos as popular or not as it suffers from overfitting due to complexity. The cutoff value is adjusted multiple times through Decile iterations, to achieve the best possible error rate balance, and the resulting model is not satisfactory for the classification problem.

Prediction Neural Network

Similar to the classification neural network, an automatic prediction model is first run to get the architecture search error log, which is used to filter the best performing model based on the least validation RMSE values. The log is as shown:

NetID	# Hidden Layers	# Neurons (Layer 1)	# Neurons (Layer 2)	Training SSE	Training RMSE	Training MSE	Validation SSE	Validation RMSE	Validation MSE	Testing SSE	Testing RMSE	Testing MSE
Net 39	2	7	2	2667.68	0.73	0.53	1748.53	0.76	0.58	1093.26	0.74	0.55

Net 9	2	1	2	2714.18	0.74	0.54	1780.91	0.77	0.59	1111.94	0.75	0.56
Net 14	2	2	2	2717.59	0.74	0.54	1760.62	0.77	0.59	1109.34	0.74	0.55
Net 19	2	3	2	2742.25	0.74	0.55	1798.88	0.77	0.6	1124.85	0.75	0.56
Net 24	2	4	2	2729.61	0.74	0.55	1789.84	0.77	0.6	1118.53	0.75	0.56
Net 27	2	4	5	2760.17	0.74	0.55	1799.35	0.77	0.6	1127.02	0.75	0.56
Net 29	2	5	2	2711.04	0.74	0.54	1778.83	0.77	0.59	1111.32	0.75	0.56
Net 34	2	6	2	2706.57	0.74	0.54	1767.52	0.77	0.59	1107.17	0.74	0.55
Net 42	2	7	5	2700.01	0.73	0.54	1779.8	0.77	0.59	1107.09	0.74	0.55
Net 2	1	2	0	2822.38	0.75	0.56	1827.58	0.78	0.61	1150.99	0.76	0.58
Net 22	2	3	5	2780.03	0.75	0.56	1825.82	0.78	0.61	1134.85	0.75	0.57
Net 32	2	5	5	2748.7	0.74	0.55	1807.11	0.78	0.6	1123.8	0.75	0.56
Net 37	2	6	5	2794.94	0.75	0.56	1825.07	0.78	0.61	1139.92	0.75	0.57
Net 17	2	2	5	3047.36	0.78	0.61	1933.78	0.8	0.64	1235.02	0.79	0.62
Net	2	3	1	2998	0.77	0.6	1935.77	0.8	0.65	121	0.78	0.61

18				.88						8.86		
Net 31	2	5	4	2959 .37	0.77	0.59	1928.72	0.8	0.64	120 3.18	0.78	0.6

The best performing neural network is selected as follows:

Neuron Weights: Input Layer - Hidden Layer 1

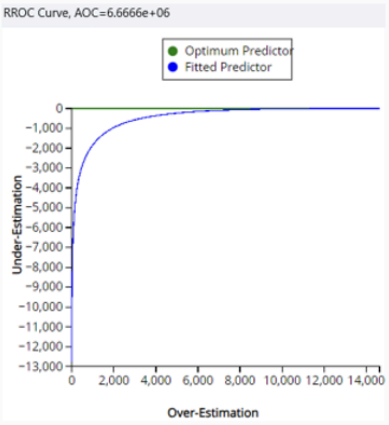
Neurons	Tag_words	category_id_1	category_id_10	category_id_20	category_id_23	publish_month_Apr	publish_month_Jun	publish_month_Mar	publish_month_May	ratings_disabled_0	Bias
Neuron 1	-0.06657718	0.19636509	-0.733657638	0.289676319	0.29579777	0.269381879	-0.167066402	0.33480221	-0.106264046	0.470931123	-3.1E-06
Neuron 2	-0.27337543	-0.193555146	0.052559488	-0.050767969	-0.312289629	-0.102382163	-0.364901259	0.362505735	-0.053417047	-0.202382212	7.29E-07
Neuron 3	0.04941605	0.102614106	-0.296333977	0.089792491	0.042315263	-0.281340615	-0.117670392	0.201648807	0.546667224	0.196540369	-5.6E-06
Neuron 4	-0.42148206	0.024629441	0.640808737	0.198967168	-0.070234403	0.067930456	0.003116947	-0.306047319	-0.330283298	-0.259743899	2.35E-06
Neuron 5	0.18333761	-0.448961522	-0.277654955	0.310689064	-0.129087875	-0.146104684	-0.628002767	-0.032687716	-0.130856547	-0.609979828	1.48E-06
Neuron 6	0.58344522	0.008490085	0.641797033	0.157439051	0.207880023	-0.067507299	-0.854190668	-0.782469294	0.492295996	0.133805656	3.67E-07
Neuron 7	0.35734508	-0.309565143	0.234513786	0.063932748	-0.290823775	-0.178168476	-0.406165442	0.12402075	-0.068273691	-0.394863775	-2.4E-08

Neuron Weights: Hidden Layer 1 - Hidden Layer 2

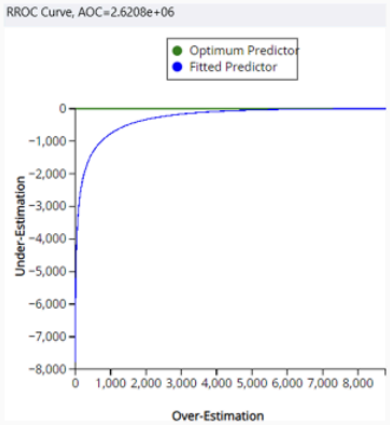
Neurons	Neuron 1	Neuron 2	Neuron 3	Neuron 4	Neuron 5	Neuron 6	Neuron 7	Bias
Neuron 1	-0.07957622	0.234702406	-0.876895693	0.346228876	0.353550022	0.321960166	-0.199683558	2.91624E-06
Neuron 2	0.40015108	-0.127030487	0.562870964	-0.326748398	-0.231358294	0.062839579	-0.060681734	-8.78475E-06

Neuron Weights: Hidden Layer 2 - Output Layer

Neurons	Neuron 1	Neuron 2	Bias
Response	-0.14892506	0.439055748	-6.96629E-05



Training data



Validation data

Training: Prediction Summary

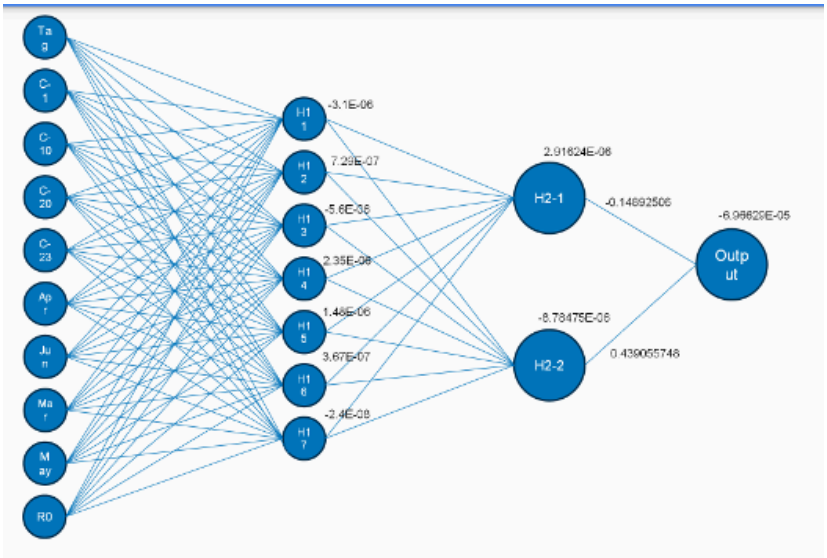
Metric	Value
SSE	2667.6842
MSE	0.53353684
RMSE	0.73043606
MAD	0.55551214
R2	0.00953532

Validation: Prediction Summary

Metric	Value
SSE	1748.52553
MSE	0.58284184
RMSE	0.76344079
MAD	0.57823726
R2	0.00725863

- Both **R^2 values** (0.0095 for training, 0.0073 for validation) indicate that the model explains **very little variance** in the output variable.
- **RMSE** and **MAD** values show that the model predictions deviate moderately from the observed values.
- The area under the curve (AUC) values are **6.66e+06** (training) and **2.62e+06** (validation), suggesting minimal predictive power.
- The neural network prediction model processes inputs through two hidden layers to estimate the target variable. However, its performance is poor, as indicated by low R^2 values. The model struggles to capture meaningful relationships in the data, likely due to feature complexity.

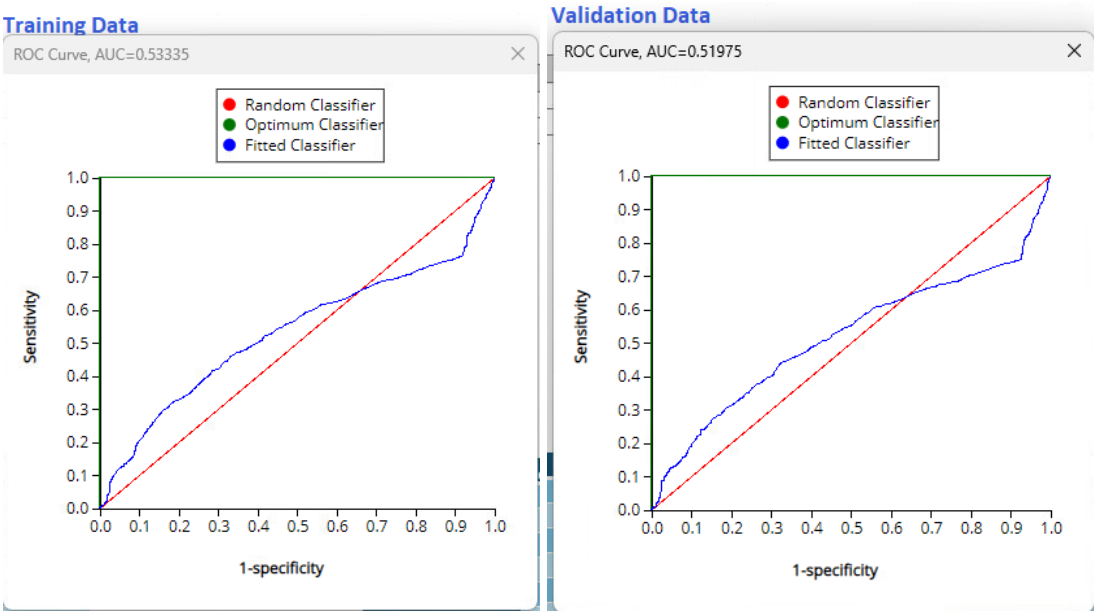
A snapshot of the neural network built is as shown below:



5. **For classification models, determine an appropriate cutoff value based on your results. Run the model with alternative cutoff values and compare performance.**

As mentioned earlier, for logistic regression and classification tree models, there was no necessity for changing the cutoff value, as there was substantial balance between error rates, and changing the cutoff based on the decile would increase the error rate in class 1. For the classification neural network, an appropriate cutoff value was selected through different iterations to achieve the best possible model.

The selected best model had the following performance:



- Final Mode (1) : Cutoff value = 0.519067
- Comparison Model (2) : Cutoff value = 0.529263

Metric	Model 1 (Cutoff: 0.519067)	Model 2 (Cutoff: 0.529263)
Training Accuracy (%)	**54.56**	**55.81**
Validation Accuracy (%)	**53.48**	**54.31**
Specificity	**0.5369**	**0.6427**
Sensitivity (Recall)	**0.5329**	**0.4703**
Precision	**0.5596**	**0.5834**
F1 Score	**0.5459**	**0.5193**
Validation AUC	**0.5198**	**0.5232**

- Model 1 has almost similar validation accuracy (53.48%) compared to Model 2 (54.31%).
- Training accuracy remains almost identical between the two models.
- Model 2 shows lower recall (0.4703) as compared to Model 1 (0.5329), which means model 1 predicted more positives from the actual positive classifications.
- Model 2 performs slightly better at specificity, and this is likely due to high error rate in predicting class 1.
- Model 1 has slightly lower precision (0.5596) compared to Model 2 (0.5834), meaning slightly more false positives when predicting "Popular" videos.
- But model 1 has way better F1 score (0.5459) as compared to model 2 (0.5193) which is the sole measure of a better performance, which is why model 1 was selected.
- The AUC scores for both the models are very similar and will not have a significant effect.

6. Explain how the results address your business questions.

- The analysis from Model 1 identifies the key factors influencing video popularity (tags, categories, timing), highlights top-performing categories (Music and Gaming), and provides a clear strategy for advertisers to maximize reach and engagement through targeted campaigns.

- **Best performing models for prediction and classification:**
    - **Predicting video view count**
        - Linear Regression
    - **Classifying video as popular**
        - Logistic regression
        - Decision Tree

4. Recommendations

- The factors that affect how popular a YouTube video can be in the USA are the month the video is published, the category of video it is, the number of tags used in the description, and the number of words in the title.

- The categories that have the most views which provide a larger audience for advertisers are music, film, gaming, travel & events, and comedy.

- Companies should advertise on videos that are published between March and June and avoid videos published between September and October. These videos should be in the categories music, film, gaming, travel & events, and comedy while staying away from videos in the categories pets & animals and news & politics. Of these videos, the ones that will find a larger audience have more tags in the description and fewer words in the title.

5. Project Summary

- **5.1 Lessons learned from the project.**

  This project on YouTube video data revealed critical insights using multiple analytical models (linear regression, logistic regression, decision trees, and neural networks). By analyzing a dataset of trending videos, the research uncovered that video popularity is significantly influenced by publication month, video category, and subtle factors like tag count and title length. The most successful categories for advertising were identified as Music, Film & Animation, Gaming, Travel & Events, and Comedy, with videos published between March and June showing the highest performance. Across different modeling approaches, the project demonstrated that strategic content creation—such as using around 20-25 tags, maintaining concise titles, and choosing the right category—can substantially impact video views. While no single model captured all nuances, the combined insights provided a comprehensive understanding of YouTube video trends, offering valuable strategic guidance for content creators and advertisers seeking to maximize their digital content's reach and engagement.

- **5.2 Dataset limitations and suggestions for future data extensions.**

    - Advertiser's target audience may be in a category that is less popular and not in more popular categories. A model predicting video view counts and popularity within specific categories would be helpful for these advertisers.

    - Creators on YouTube can use this model as it would guide them to create videos that tend to trend best, creating the best environment to grow their accounts.

- Current dataset only covers US YouTube trends, limiting global applicability

6. Task Allocation

Austin Serody	Data preprocessing, Model construction and analysis, report generation.
Balram Iyengar	Data preprocessing, Model construction and analysis, report generation.
Praneeth Reddy	Data preprocessing, Model construction and analysis, report generation.
Shiva Praneeth Kodali	Data preprocessing, Model construction and analysis, report generation.