Group 1 – Milestone 3

For this milestone, you are tasked with completing **the first model analysis**.

The final output should include at least **three** distinct analytical models, with **one being a classification** model. Follow the guidelines below:

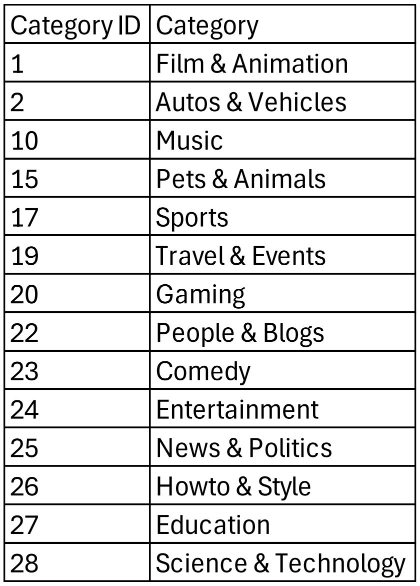
1. Define the business questions (note any changes to the questions from the first two milestones).
   - What factors play a role in the popularity of YouTube videos in USA?
   - What video categories have the most view counts?
   - Does disabled engagement affect the popularity of a video?
   - How significant is the title length to the view count or popularity?
   - Based on the analysis, what strategy can an advertiser use to pick a YouTube video for advertising their product?
2. Identify the models used and provide a rationale for each selection.
   - 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.
   - For the classification problem, both logistic regression and classification tree models will be used to draw the relationships and rules that predict the popularity of a video.
   - For this milestone, a classification tree model is analyzed and presented in the following sections.
   - For the final presentation, all the models mentioned above will be presented along with a prediction and classification neural network that would predict the view count and classify the popularity of a video. All

the models will be compared, and the best model will then be chosen that is appropriate for the given business questions.

- An additional dependent variable will be added to the final analysis which is the categorical Popular-liked variable, that provides us classification of videos that are popular and most liked.

3. Specify the variables included and justify their choice.

   - As mentioned above, a classification tree has been built and analyzed for this milestone. So, 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	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

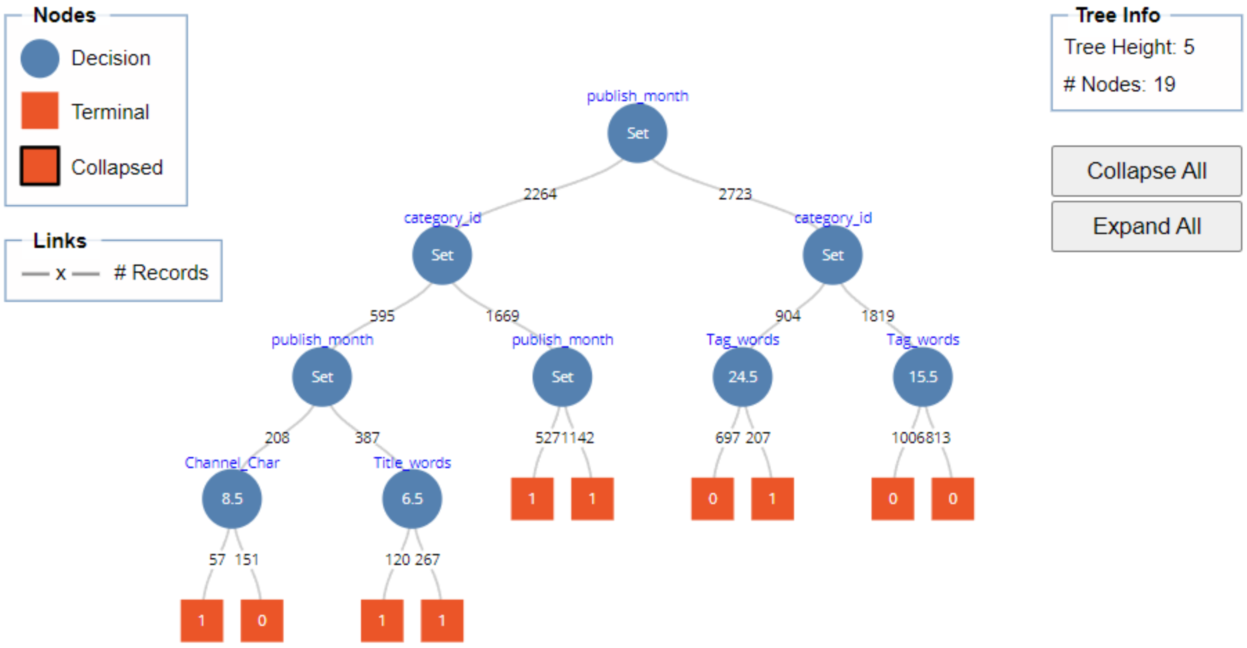
4. Describe any variable selection techniques applied.

   - 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.

5. 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.

   - 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:



Tree Info

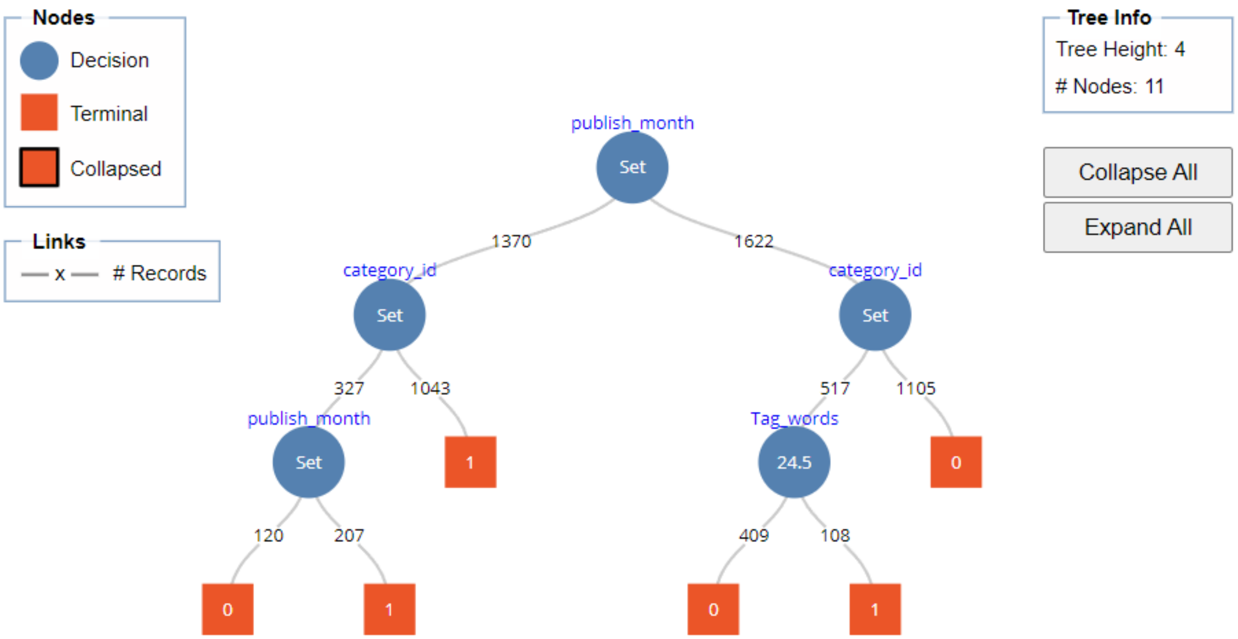
Tree Height: 5

Nodes: 19

   Best tree:

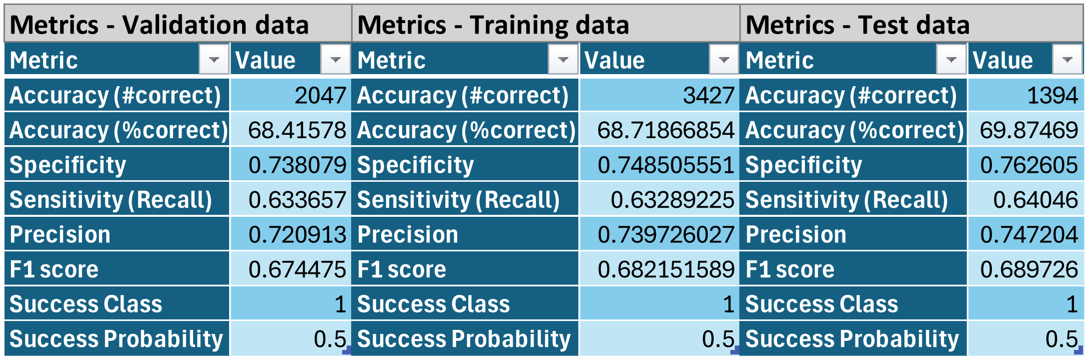
   Classification rules based on the best pruned tree are as follows:

   - 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.



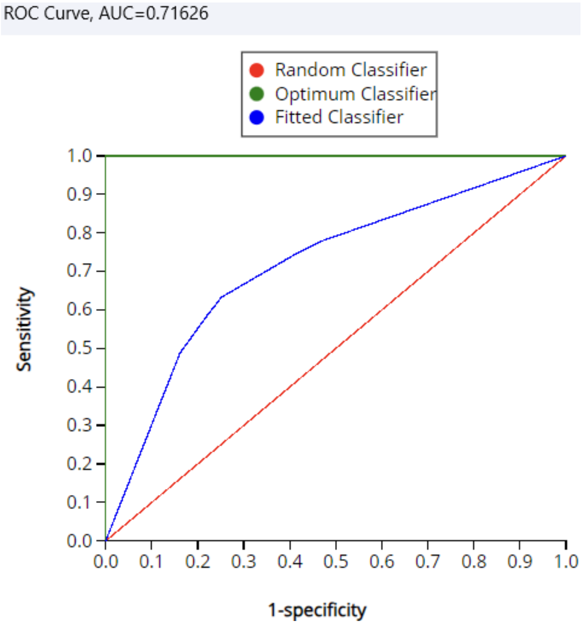
- A video published in the months {March, April, may, June} and is from the categories {2, 15, 17, 25, 26, 27} is predicted to be popular if not published in the month of March and not popular if published in the month of March.
- A video that is not published in the months {March, April, May, June}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

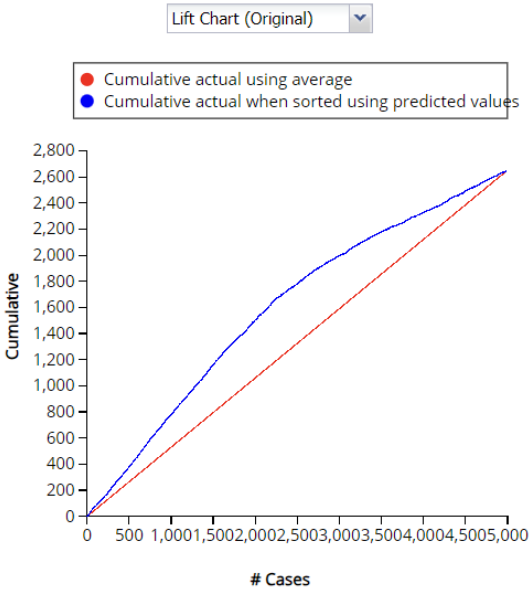
6. Provide a summary report for training, validation, and test data (if applicable), along with lift charts. Assess the model's performance.



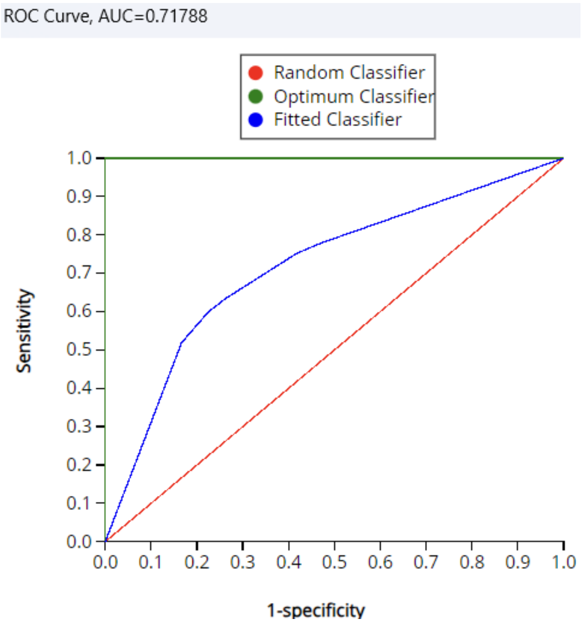
Metrics - Validation data		Metrics - Training data		Metrics - Test data	
Metric	Value	Metric	Value	Metric	Value
Accuracy (#correct)	2047	Accuracy (#correct)	3427	Accuracy (#correct)	1394
Accuracy (%correct)	68.41578	Accuracy (%correct)	68.71866854	Accuracy (%correct)	69.87469
Specificity	0.738079	Specificity	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

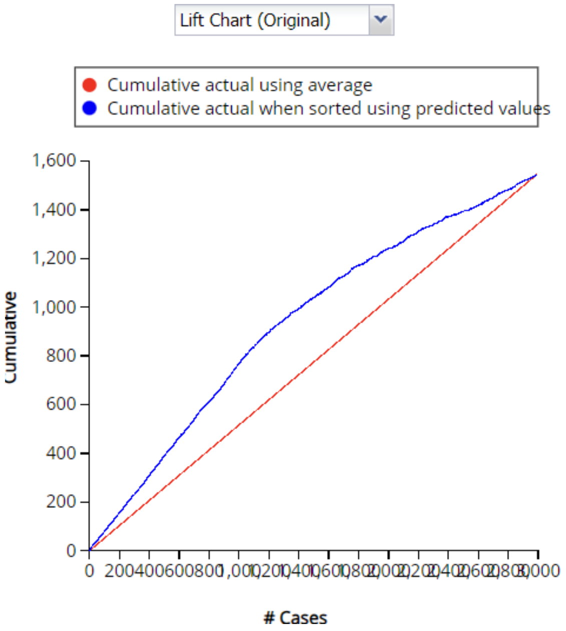
Lift charts – Training data





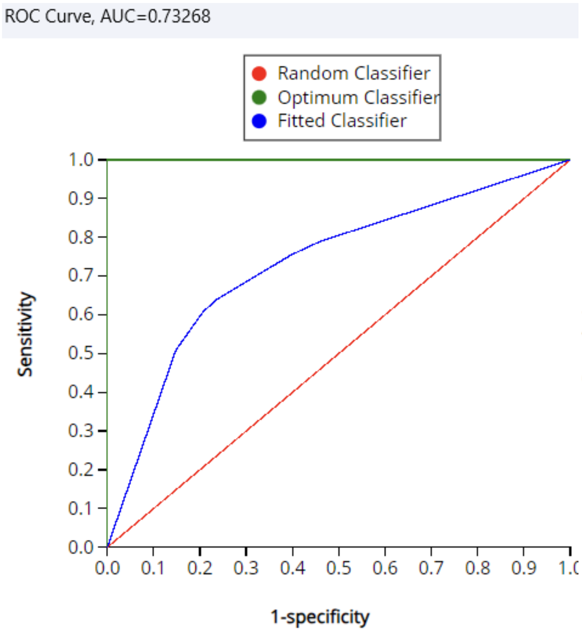
Lift charts – Validation data

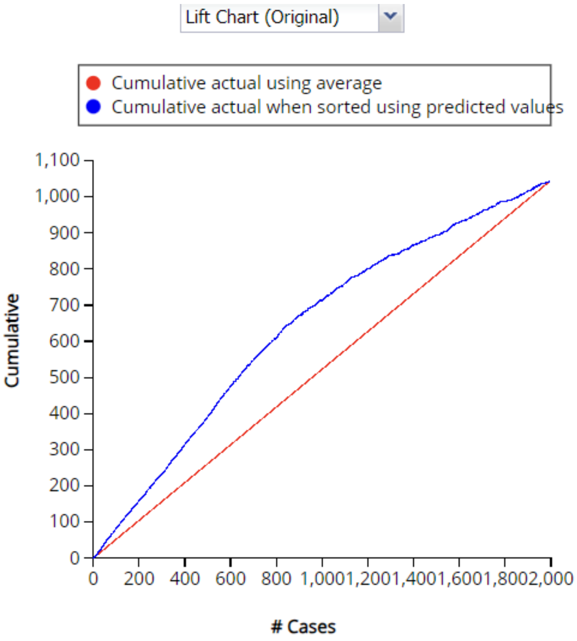




As indicated in the summary tables, the model shows good results equally on all training, validation and test data with an average overall error rate of 31%, and accuracy of 69%. The model presents precision and recall values of 74% and 64% respectively, and an F1 score of 68%. As can be seen in the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

Lift charts – Test data





7. 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.

8. 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 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 do not have a significant effect on the popularity of a video, but the number of a tags on the video have some effect on the popularity.

9. 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.
10. Summarize your conclusions, recommendations, or insights for the organization, clearly detailing potential actions in response to the findings. Write the conclusions in non-technical terms for a general audience.
    - Based on this classification tree, for an advertiser to pick a video for showcasing their product, the best possible outcome would be to pick a video published in the months of March, April, May and June, from the categories of Film, music, travel, gaming and comedy, with the average number of tags on the video greater than or equal to 25. According to the analyzed model, it is probable that this criterion would make sure that a product is advertised on a popular video.

Additionally, include a **task allocation table** to outline each group member's responsibilities.

- Praneeth Anumula – Data transformation, Model build and analysis
- Austin Serody – Model analysis and Reporting
- Bhanu Ram – Model analysis and Reporting
- Shiva Kodali – Model analysis and Reporting