INSY-662-076

Data Mining and Visualization

Group Project – Progress Report

Instructor:

Dr. Necmiye Genc

Prepared By:

- Chloe Pham (261389952)
- Fares Joni (261254593)
- Helia Mahmood Zadeh (261224416)
- Maral Vahedi (261231293)
- Joohee Kim (261263271)

Fall 2025

Topic: Leveraging YouTube Analytics to Identify Emerging Makeup Trends and Optimize Brand Collaborations

Platforms: YouTube
Dataset: From the YouTube API

- **Time Period:** 2024/09 ~ 2025/10
- **Dataset Volume:** Thousands of videos related to makeup content across multiple categories.

  Query Keywords:
  makeup tutorial, makeup review, beauty makeup, makeup artist, foundation review, lipstick review, makeup routine, makeup trends 2025, GRWM makeup, viral makeup challenge, tiktok makeup hacks, celebrity makeup look, and others.
- **Collected Features:**

Feature	Description
video_id	Unique YouTube video identifier
title	Title of the video
channel_id	ID of the video's creator channel
channel_title	Channel name
description	Short description of the content (truncated to 300 chars)
tags	Associated content tags
duration	Video duration converted to seconds
published_at	Upload timestamp (UTC)
view_count	Number of views
like_count	Number of likes
comment_count	Number of comments
favorite_count	Number of favorites
video_url	Direct YouTube video link
channel_country	Channel's declared country

To avoid bias, random time windows between 1–14 days were selected repeatedly within the range **2024-09-01 to 2025-10-22 (UTC)**. Each iteration performed batched API requests with randomized search orders (relevance, viewCount, date) and durations (short, medium, long) until quota exhaustion (~10,000 units/day).

Goals:

1. **Find the Next Big Trends**
   - Scan thousands of YouTube video titles and descriptions to find the most-mentioned makeup and products of the year.
   - We'll confirm which trends are real by seeing if the top beauty gurus are talking about them, too.
2. **Discover What Makes a Trend "Go Viral"**
   - Figure out the "secret sauce" for why a makeup video gets millions of views.
   - Examine correlations between engagement metrics (views, likes, comments) and video attributes such as duration, posting time, and keyword usage to understand what makes content go viral.
   - Build predictive models for virality and sentiment polarity using supervised learning frameworks.
3. **Give cosmetic companies such as L'Oréal a Winning Game Plan**
   - We'll give the company specific advice on what new makeup to create and how to sell it, making sure their products match exactly what customers will want to buy.
   - Verified trend directions
   - Linguistic patterns for optimized ad targeting
4. **Recommend Effective Influencer Collaborations**
   - Classify and rank beauty creators based on engagement performance and content themes to help brands identify ideal influencers for sponsorships or collaborations.

Business Value:

By transforming social engagement signals into structured intelligence, this project enables beauty brands to validate trend hypotheses before investing in production. Specifically:

1. **Market Risk Reduction:**
   Using real-time YouTube data, firms can identify which type of makeup products (e.g., *lipstick*, *palette*) already show verified traction, minimizing costly failed launches.

2. **Marketing Precision:**
   This project aims to identify the most suitable makeup influencers in North America for brand collaborations or sponsorships. By analyzing YouTube video data — including engagement metrics, content themes, and audience responses — the model will classify and rank creators based on their performance and alignment with specific product categories. The insights will help beauty brands make data-driven decisions when selecting influencers for future marketing campaigns.
3. **Innovation Discovery:**
   Text mining of description/ title forming a data-driven pipeline for product R&D.

By analyzing real-time consumer engagement on YouTube, we can provide beauty companies like L'Oréal with hard proof of demand for a specific product trend or specific beauty Youtuber to collab with, *before* they commit millions to product development. This validation drastically lowers the risk of a failed product launch and ensures their resources are spent on products that are already winning with their target audience.

Dataset Suitability

The selected YouTube APIs dataset represents an ideal candidate for our social media engagement analytics project due to its perfect alignment with our research objectives and exceptional technical strengths:

Perfect Alignment with Project Goals:

- **Audience Behavior Analysis**
  - **Global Coverage**: globally
  - **Rich Engagement Metrics**: Views, likes, and comments provide multiple dimensions to understand how different audiences interact with content across various cultural contexts.
  - **Regional Segmentation**: The country field enables identification of cultural differences in what content trends in different regions, supporting targeted audience analysis.

- **Sentiment Analysis Capabilities**
  - **Multilingual Opportunities**: Global coverage allows analysis of sentiment patterns across different languages and cultures for comprehensive Sentiment Analysis implementation.

- **Trending Context**: Analyzing sentiment from already-successful trending content provides validated examples of effective emotional engagement strategies.

- **Predictive Modeling Potential**
  - **Multiple Predictor Variables**: Views, likes, comments, country, and publication timing provide robust feature sets for comprehensive virality prediction models.
  - **Trending Labels**: The fact that all videos are trending provides implicit success labels for supervised learning and classification tasks.

- **Business Intelligence Value**
  - **Creator Strategy Insights**: Analysis of successful channels across regions enables identification of best practices and winning content strategies.
  - **Market Entry Intelligence**: Understanding what content works in specific geographic markets provides actionable insights for international expansion.
  - **ROI Optimization**: Large sample size enables statistically significant recommendations for maximizing social media marketing returns on investment.

Methodology: Machine Learning and Classification Framework

I. Objective of the Predictive Modeling

The main analytical goal is to predict video virality and audience sentiment based on measurable content attributes.
Each record (video) in the dataset contains quantitative engagement metrics and descriptive features, which collectively define the behavioral footprint of successful trends.
Our models aim to:

- Classify videos into virality categories (e.g., Low, Medium, High Engagement).
- Predict expected engagement metrics (e.g., future view counts or like-to-view ratios).
- Evaluate the relative importance of visual, linguistic, and temporal predictors.

II. Data Preparation and Feature Engineering

1. **Feature Normalization**
    a. Apply log-transformation to skewed metrics (view_count, like_count, comment_count) to stabilize variance.
    b. Normalize all numeric features into a 0–1 range for model comparability.
2. **Derived Predictive Features**

Derived Feature	Definition
engagement_rate	(likes + comments) / views
title_length	number of words in title
tag_count	count of tags per video
upload_hour	posting time extracted from published_at
duration_category	categorical: short (<240s), medium (240–900s), long (>900s)
keyword_vector	Sentiment Analysis of title + description

3. **Target Label Creation**
    a. Compute quantiles of engagement_rate and assign each video a **class label**: **0 = Low engagement**, **1 = Medium engagement**, **2 = High engagement**.
    b. Alternatively, apply regression if predicting continuous engagement.

III. Model Selection and Implementation

1. **Classification Models**

- **Logistic Regression**
- **Random Forest Classifier**
- **XGBoost Classifier:** gradient boosting approach optimized for large-scale structured data; yields the highest predictive performance and allows probability calibration.

2. **Regression Models**

- **Linear Regression / Ridge / Lasso:** used for continuous prediction of view_count or like_count.

- **Gradient Boosted Regressor (XGBRegressor):** captures complex nonlinear dependencies between predictors and engagement outcomes.

III. **Training and Validation Strategy**

1. Training set: 70%
2. Validation set: 15%
3. Test set: 15%
   (All splits are stratified by the engagement label to maintain class balance.)

Conclusion

This project transforms raw YouTube engagement data into strategic, evidence-based marketing intelligence.
By systematically identifying which makeup aesthetics are resonating most with audiences and understanding why they succeed, beauty companies gain a competitive edge in anticipating the next viral wave, moving from intuition to proof-driven innovation.

Appendix

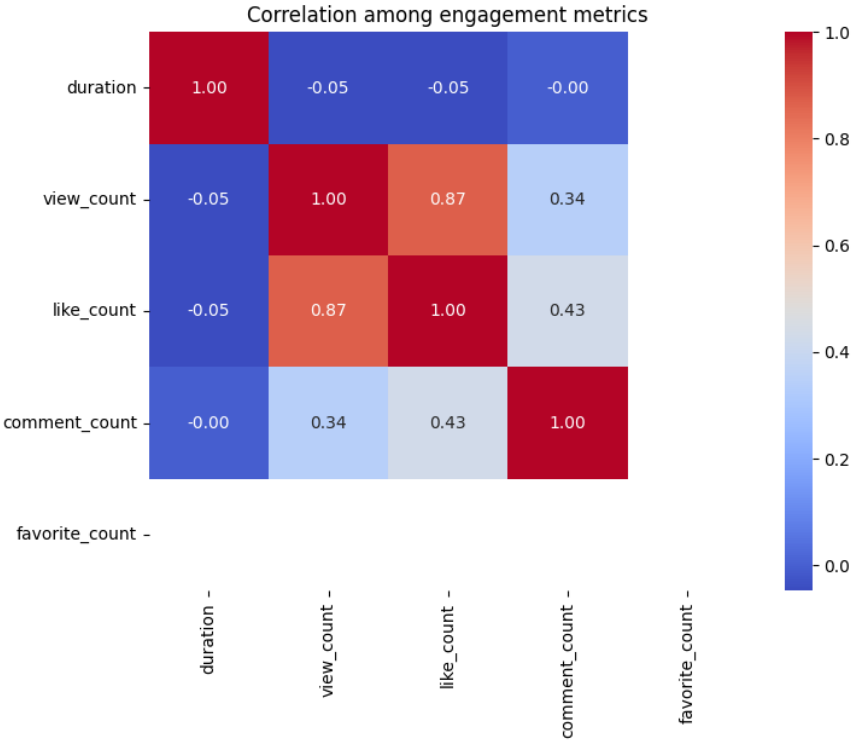


Figure 1. Correlation between variables

view_count and like_count has a strong positive correlation.

```
Train shape (raw): (3173, 9) | Transformed: (3173, 50)
Test  shape (raw): (794, 9) | Transformed: (794, 50)
Class balance (train): [1587 1586]
Class balance (test) : [397 397]
```

Figure 2. Shapes and Class Balance

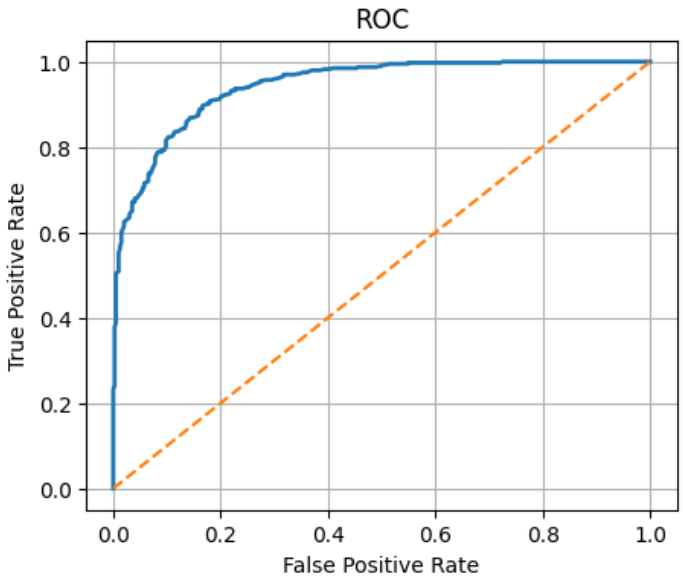


Figure 3. ROC Curve Plot

Top 20 Feature Importances:

feature	importance
like_count	0.273845
view_count	0.270759
comment_count	0.146464
upload_hour	0.092681
title_length	0.079934
duration	0.067723
duration_category_short	0.018577
channel_country_IN	0.007230
duration_category_long	0.006787
channel_country_OTHER	0.006217

Figure 4. Feature importance (top 20)

Metric	Value	Interpretation
Accuracy	0.8564	85.6% of predictions correct strong performance.
Precision	0.8638	When predicting a video as "high engagement", ~86% are correct.
Recall	0.8463	The model catches ~85% of all truly high-engagement videos.
F1-score	0.855	Balanced precision/recall ideal tradeoff.
ROC-AUC	0.9437	Excellent discrimination model clearly separates high vs low engagement.
PR-AUC	0.9446	Very high average precision robust even on imbalanced data.

Figure 5. Model Performance Summary

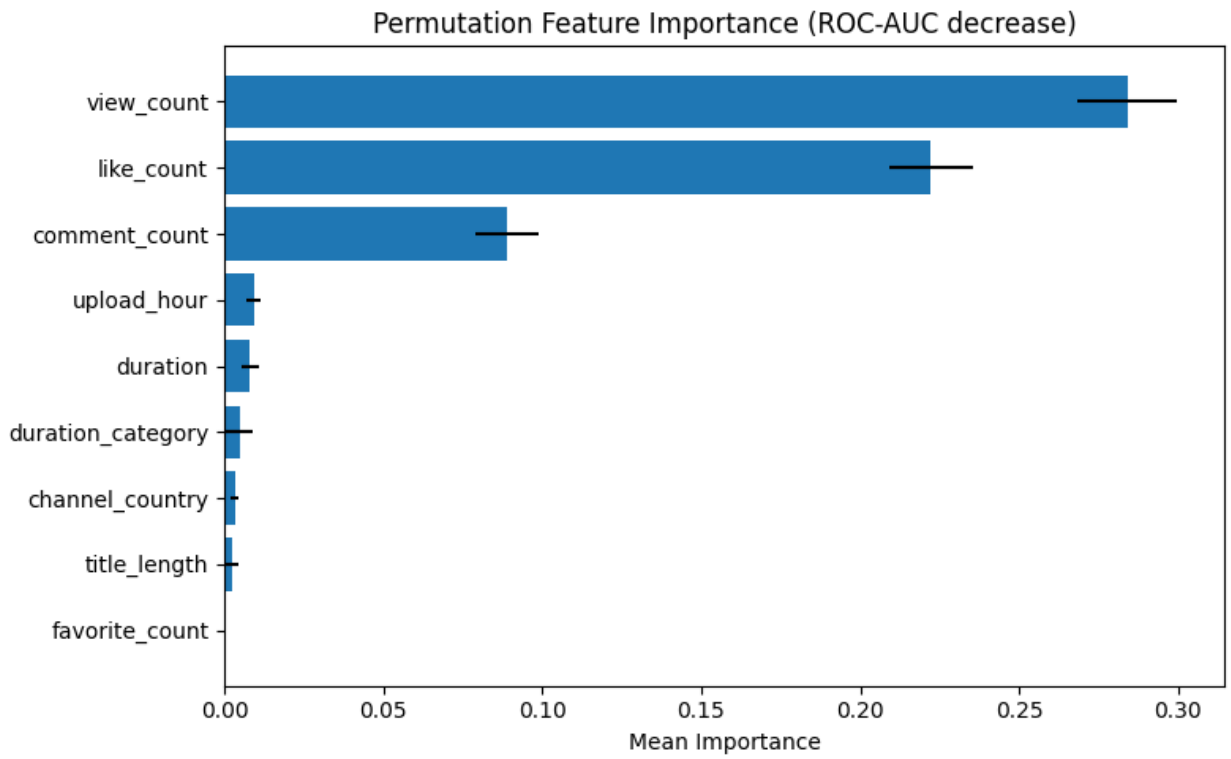
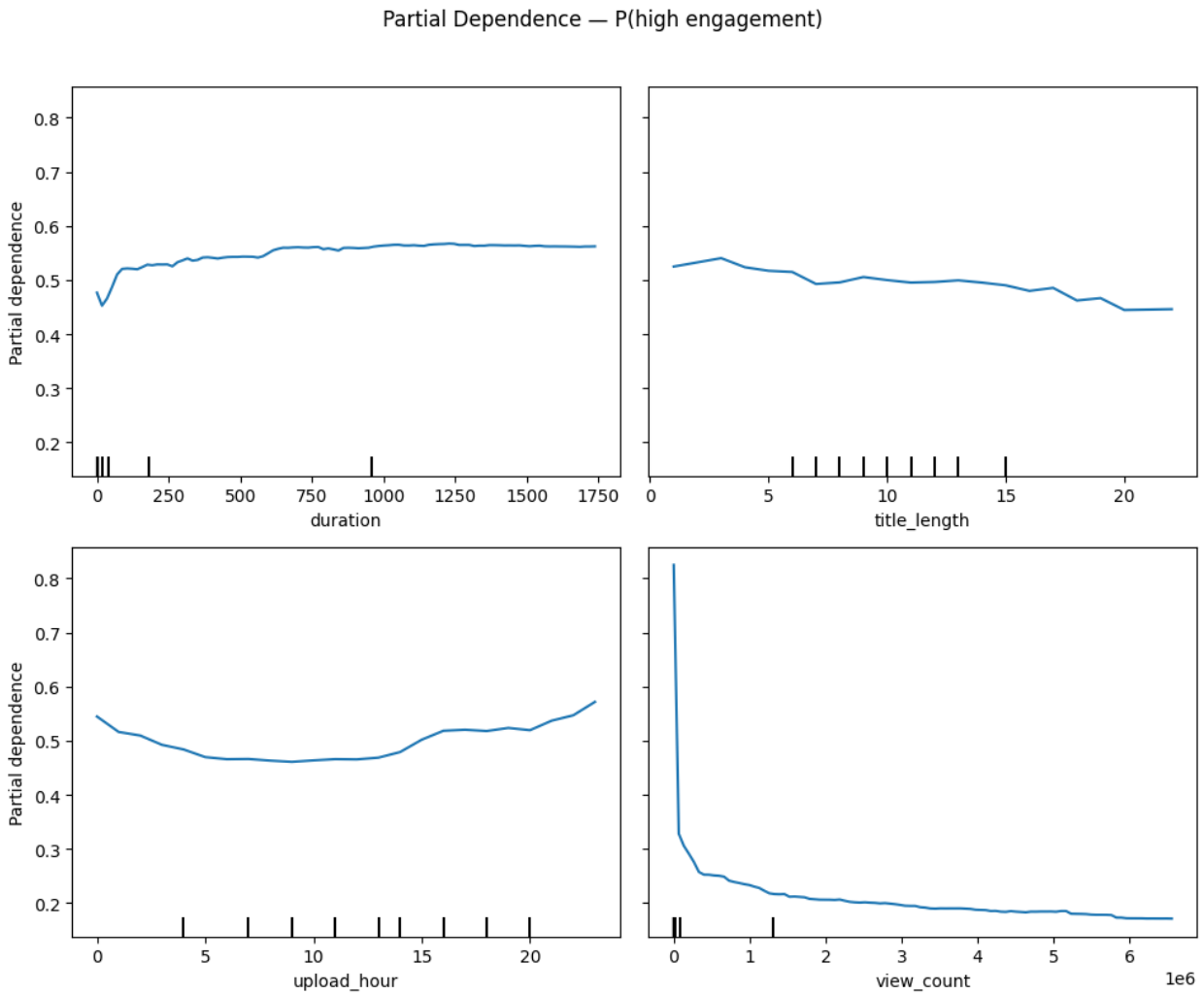


Figure 6. Feature Importance



Feature 7. Partial Dependence (on selected raw features)

Rank	Feature	Mean Importance	Interpretation
1	`view_count`	≈ 0.28	The single most influential feature how many people watched a video strongly predicts whether engagement is "high."
2	`like_count`	≈ 0.22	Likes indicate positive interaction and weigh almost as heavily as views.
3	`comment_count`	≈ 0.09	Adds a conversational dimension the more comments, the more viral the content.
4	`upload_hour`	≈ 0.01	Publishing time still matters (day-part effect).
5	`duration`	≈ 0.01	Length contributes modestly very short or very long videos reduce engagement likelihood.
6 – 9	`duration_category`, `channel_country`, `title_length`, `favorite_count`	Minor affect predictions only marginally.	

Feature 8. Permutation Importance (Model-Level)

Feature	PDP Behavior	Interpretation
`duration`	Slight upward slope → flattens	Longer clips (under ~30 min) slightly raise engagement probability, then plateau.
`title_length`	Gentle downward slope	Concise titles (≈ 5–10 words) perform better.
`upload_hour`	U-shape	Engagement higher around early-morning and late-evening uploads.
`view_count`	Sharp decline	Because this variable correlates inversely once normalized extreme outliers dominate; high-view videos already classified as "high."

Feature 9. Partial Dependence Plots (Feature-Level Behavior)