Dhirubhai Ambani University
Technology

Formerly DA-IICT

DS605 – Fundamentals of machine learning

THE CONFUSED MATRIX

Name	Student ID
Xena Doris Pereira (leader)	202518014
Pronnati Tapaswi	202518052
Angel Manoj	202518033
Deep Patel	202518050

Abstract
This project aims to support YouTube content creators by predicting both the expected engagement rate of an upcoming video and the optimal video duration required to maximize that engagement. Using a dataset of over 50,000 videos scraped through the YouTube API, we built machine learning models that analyze video metadata and channel characteristics to forecast performance. The system provides creators with actionable insights for strategic content planning helping them understand how well a video is likely to perform and what duration is most effective for their audience. Several regression models were evaluated, with Gradient Boosting achieving the most accurate results.

Introduction
YouTube creators often struggle to estimate how well a video will perform before uploading. Engagement driven by views, likes, comments, video duration, audience behavior, and channel factors vary widely. Predicting this performance manually is nearly impossible. This project proposes a predictive system powered by machine learning to forecast engagement score and guide creators toward optimized content decisions.

Problem Statement
YouTube content creators currently lack a systematic, data-driven approach to estimate how their videos will perform and to determine the optimal video characteristics for maximum engagement. This project addresses the problem by predicting a video's engagement rate based on both its metadata such as topic, duration, and quality and channel specific features, including subscriber count and historical performance trends. The system empowers creators to make informed decisions for content planning and audience targeting, improving the likelihood of higher engagement.

Dataset Description
The dataset consists of 50160 YouTube videos scraped using the YouTube API. It has diverse categories such as Parenting Tips, Meditation, Robotics,

Electronics Repair, Ethical Hacking, and others. Short videos under 60 seconds and videos with fewer than 1,000 views were filtered out to ensure quality.
Key columns include:
Video metadata: title, topic, category_id, published_at, duration_seconds, captions_flag, quality.
Channel metadata: subscriber_count, channel_age_days, total_videos.
Performance stats: views, likes, comments.
Engineered features: duration_minutes, publish hour/day, cyclic time encodings, subscriber_log, previous channel averages.

Data Preprocessing

The raw dataset of YouTube videos underwent several pre-processing steps to ensure quality and suitability for modelling. Key steps included:

1. **Data Cleaning**
   - Duplicate records were removed based on video_id.
   - Videos with zero views or likes exceeding views were filtered out.
   - Columns such as captions_flag, video_definition, and category_id were converted to categorical types.
2. **Numeric Clipping and Transformation**
   - Video durations were clipped between 5 and 7,200 seconds to remove unrealistic outliers.
   - Duration was also converted to minutes for easier interpretation.
   - Subscriber counts were log-transformed to reduce skewness.
3. **Target Variable Construction**
   - The engagement rate was computed as:
   - engagement_rate_log = (log(likes+1) + log(comments+1)) / log(views+1)

     This was used as the regression target.
4. **Time Feature Engineering**
   - Extracted publish_hour and publish_dayofweek from the publication timestamp.

    - Cyclical encoding applied using sine and cosine transformations to capture periodicity in posting times.
5. **Additional Feature Engineering**
    - Video-related: HD flag (is_hd), title length, word count, and punctuation counts (exclamation/question marks).
6. **Leakage Prevention**
    - Columns containing post-publication metrics or identifiers such as video_id, views, likes, comments, and raw category IDs were removed to prevent data leakage.
7. **Train-Test Split and Target Encoding**
    - Group-based splitting ensured videos from the same channel did not appear in both train and test sets.
    - Target encoding was applied to categorical features based on mean engagement rates within training groups.
8. **Outlier Clipping**
    - Outliers in numeric columns like total_videos, channel_age_days, and subscribers_per_category were clipped using an interquartile range-based method.
9. **Final Dataset for Modeling**
    - All numeric and categorical features were finalized, with categorical features explicitly marked for LightGBM training.
    - The dataset was now ready for regression modeling to predict engagement rates.

Model Training

A regression pipeline was built to predict engagement score. Train-test split, hyperparameter tuning etc were used. Models tested:

- Gradient Boosting
- XGBoost
- LightGBM
- CatBoost

Model Evaluation

Gradient Boosting:

	TEST PERFORMANCE	TRAIN PERFORMANCE
MAE	0.15497050132639917	0.1435711861455554
RMSE	0.2117415667307127	0.1952141081690276
R2	0.25861998312955525	0.35534870536460705

XGBoost:

	TEST PERFORMANCE	TRAIN PERFORMANCE
MAE	0.15058916864625246	0.1318864331357275
RMSE	0.20587921000986514	0.1832572093017762
R2	0.29910394151129227	0.43190025473561156

LightGBM:

	TEST PERFORMANCE	TRAIN PERFORMANCE
MAE	0.15365056808249036	0.13852243171442824
RMSE	0.21045154780423053	0.1905519605433147
R2	0.26762606435984604	0.38577243854307386

CatBoost:

	TEST PERFORMANCE	TRAIN PERFORMANCE
MAE	0.15204589042670164	0.13370693441005715
RMSE	0.20823041251799765	0.18483668126343047
R2	0.28300364071088824	0.4220652859186843

Analysis

Several gradient boosting-based models were tested, including Gradient Boosting, XGBoost, LightGBM, and CatBoost. Based on overall performance, **Gradient Boosting was selected as the best model**. This choice was motivated by its **minimal difference between training and test R^2**, indicating better generalization and reduced overfitting. Additionally, Gradient Boosting effectively achieved the **main objective of minimizing MAE**, ensuring accurate engagement rate predictions. Overall, the model provides a reliable and robust approach for predicting video engagement while supporting actionable insights for creators.

Application and Impact

- Helps creators optimize video strategy before upload.
- Reduces uncertainty and planning time.

• Useful for digital agencies, influencer marketers, and brand content teams.
• Scalable to other platforms such as Instagram Reels and TikTok.

Conclusion

This project demonstrates a machine learning system capable of predicting YouTube video engagement and recommending optimal video durations. Using a dataset of over 50,000 videos, multiple models were evaluated, and **Gradient Boosting was selected as the best-performing model**, providing low MAE and strong generalization with minimal R^2 difference between training and test sets. The system offers actionable insights for creators to optimize content strategy and maximize audience engagement. Future improvements could include text-based sentiment analysis, thumbnail evaluation, and multi-task learning to jointly predict engagement and optimal duration.