Youtube Trending Video Analytics

Problem Statement: The project aims to predict key factors that influence youtube video popularity which is crucial for content creators, advertisers and platform managers because trending videos drive high engagement, ad revenue and audience growth.

Project Phases:

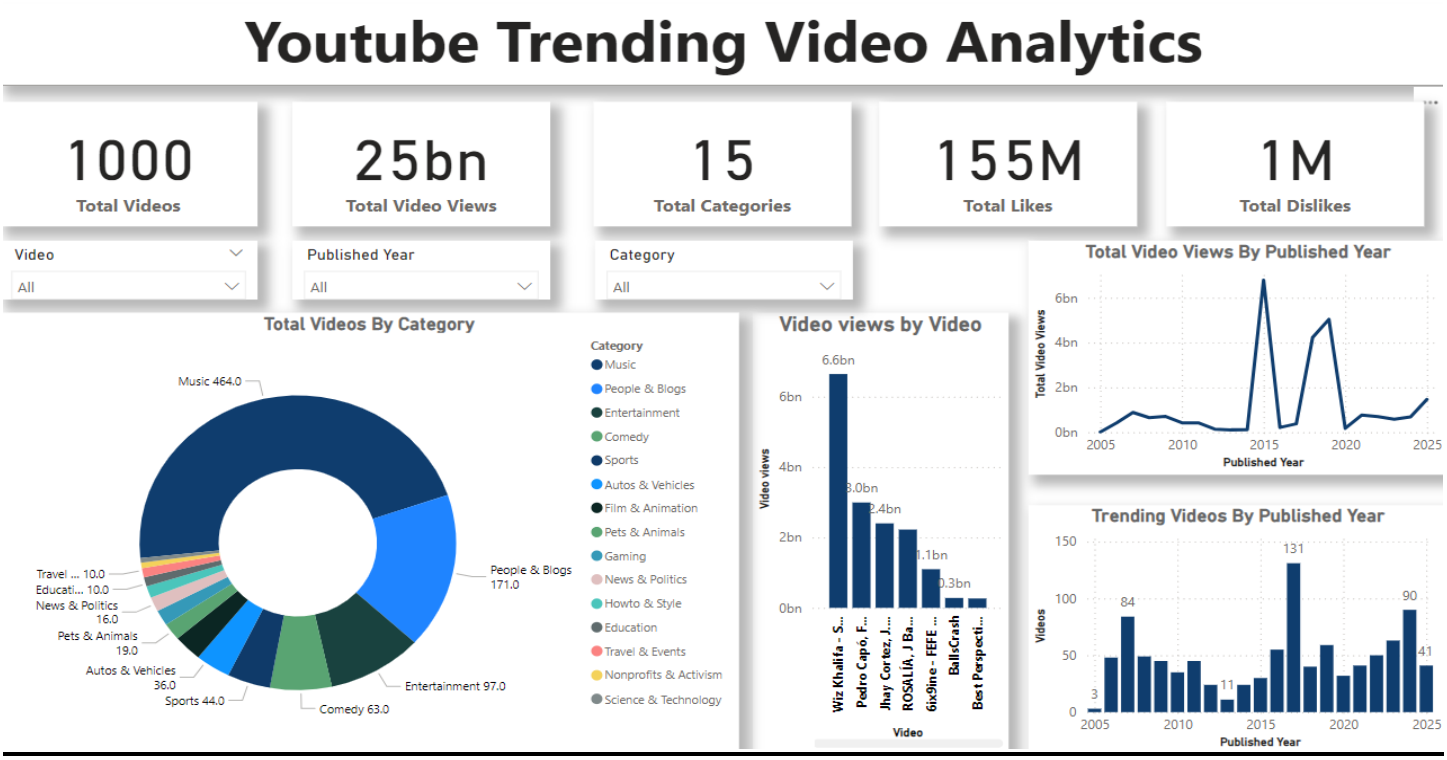
1. Data Pre-Processing.
2. Data Visualization.
3. Exploratory Data Analysis.
4. Feature Engineering.
5. Model Selection, Training And Evaluation.
6. Hyperparameter Tuning.
7. Recommendations.

1.Data Pre-Processing:

- In Data Pre-Processing firstly the dataset is loaded by using pandas library.
- There are missing values, wrong datatypes in the data.
- Dealing with wrong datatypes:
    1. Converted video views from object to float datatype.
    2. Converted Likes from object to float datatype.
    3. Converted Dislikes from object to float datatype.

- Dealing with missing values:
    1. The missing values of Likes and Dislikes columns are replaced with Median values.
    2. The missing values of Category column are replaced with Mode values.

2. <u>Data Visualization</u>: Created a Power BI report to know about data.



Summary:

1. The dataset contains **1,000 trending YouTube videos**, accumulating over **25 billion total views**.
2. These videos span across **15 categories**, with **Music** dominating at **464 videos (46.4%)**.
3. The platform has seen a collective **155 million likes** and **1 million dislikes** on the videos.
4. Peak video popularity was observed around **2015**, with total views nearing **6.7 billion** that year.
5. A sharp dip in trending uploads is seen during **2016**, with only **11 videos**, followed by a rebound.
6. The top video—by Wiz Khalifa—gained **6.6 billion views**, leading the list of individual hits.
7. Categories like **People & Blogs** (171 videos) and **Entertainment** (97 videos) also show strong engagement.
8. A filter system allows dynamic breakdown by **video**, **category**, and **published year** for deep dives.
9. Total views by year highlight content lifecycle patterns and shifts in user engagement trends.
10. The dashboard reveals a clear concentration of popularity in music-related content and specific peak years.

3. Exploratory Data Analysis:

- Transformations: Log-transform Video views, Likes, and Dislikes to handle skewness.
- Feature Engineering: Create video_age (2025 - published) to capture the effect of time on views. Encode Category using one-hot or target encoding.
- Category Imbalance: Address the dominance of "Music" by either grouping rare categories or using techniques like target encoding to reduce dimensionality.
- Feature Selection: Likes is a strong predictor (high correlation with Video views), while Dislikes may be less impactful. We'll assess feature importance during modeling.
- Modeling Strategy: Given the correlations and non-linear relationships (e.g., views by category), tree-based models like Random Forest or XGBoost are likely to perform well.

4. Feature Engineering:

- Log transformed Video views, Likes, and Dislikes.
- "Video_age" feature is created to capture the effect of time on views.
- Scale numerical features (log_likes, log_dislikes, video_age).

The dataset is now fully numerical, scaled, and ready for modeling.

log_likes and category_encoded are likely to be strong predictors based on EDA correlations and category effects.

5. Model Selection, Training and Evaluation:

- Import libraries required for model training.

```
# Import required libraries
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
!pip install xgboost
from xgboost import XGBRegressor
from sklearn.metrics import mean_squared_error, r2_score
```

```
# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

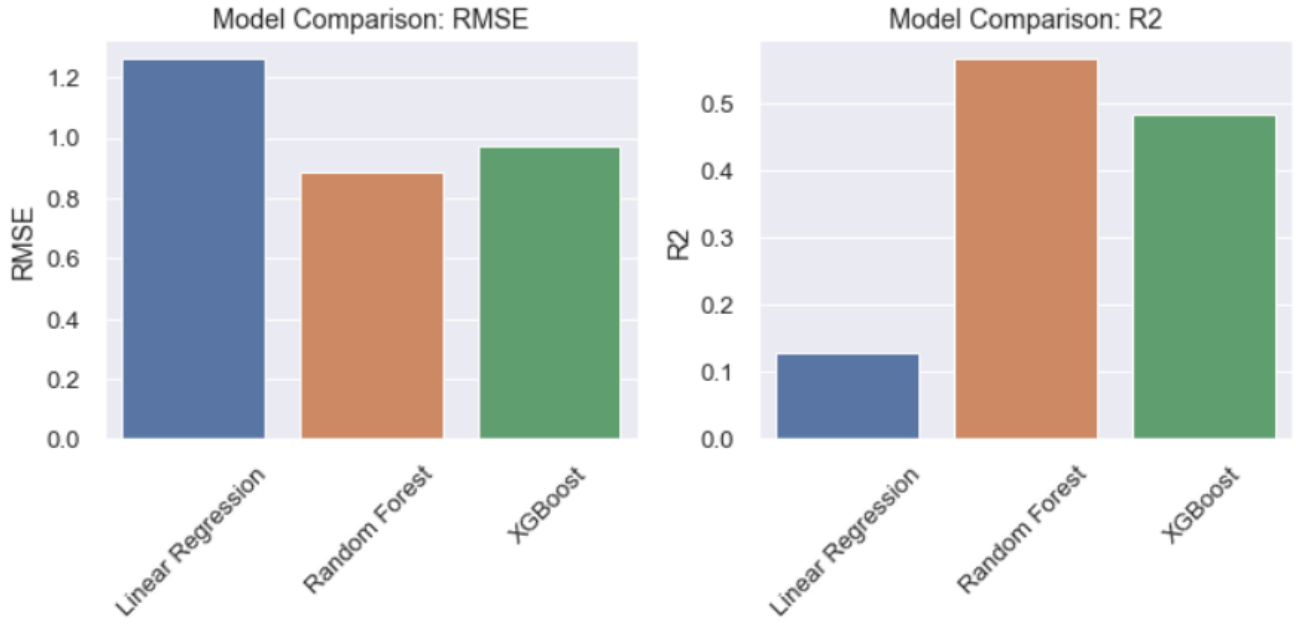
- Model Training Results

```
Linear Regression Results:
RMSE: 1.2634
R2: 0.1283

Random Forest Results:
RMSE: 0.8897
R2: 0.5677

XGBoost Results:
RMSE: 0.9721
R2: 0.4840
```

- Model Comparison:



- Random Forest is the best model with RMSE = 0.8897 and R^2 = 0.5677, and will be the focus of hyperparameter tuning.
- The model explains 56.77% of the variance, which is a good starting point but can be improved through tuning.

6. Hyperperameter Tuning:

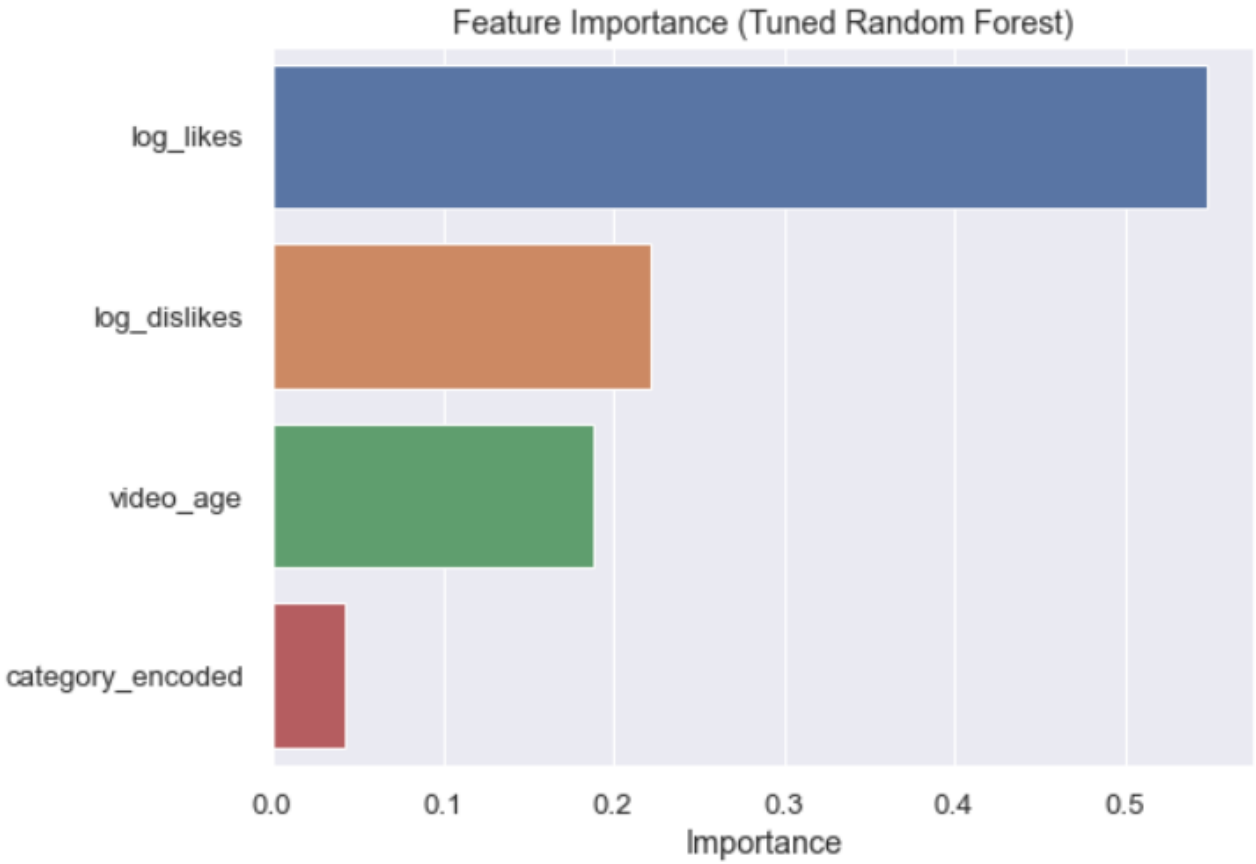
- Tuned Random Forest Results:

```
Tuned Random Forest Results:
RMSE: 0.8622
R2: 0.5940
```

Baseline Random Forest Results VS Tuned Random Forest Results

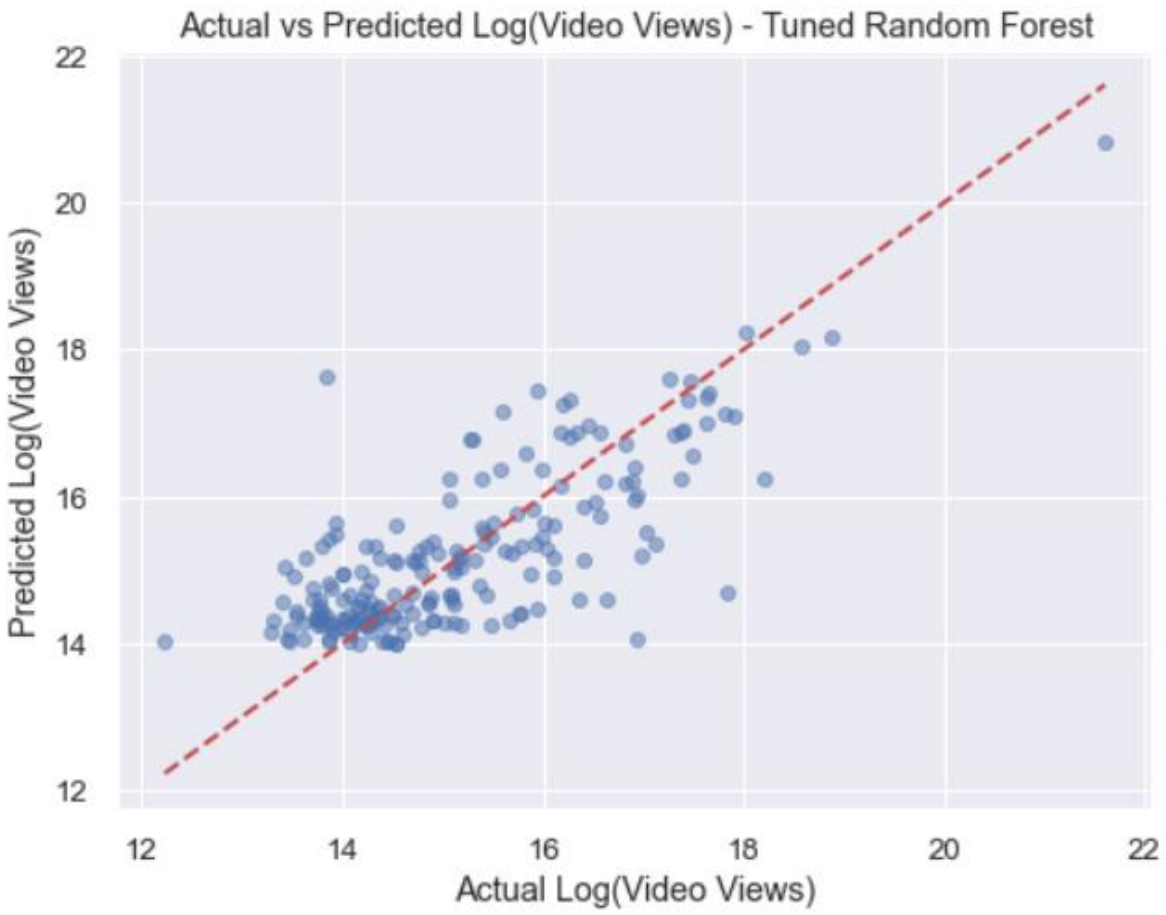
```
Baseline Random Forest Results (for comparison):
RMSE: 0.8897
R2: 0.5677
```

```
Tuned Random Forest Results:
RMSE: 0.8622
R2: 0.5940
```



- **log_likes** is the most important predictor, contributing over 50% to the model's decision-making, indicating a strong relationship with the target variable.
- **log_dislikes** also holds significant predictive power but is notably less important than log_likes.
- **video_age** has moderate importance, suggesting the age of a video has some impact on the outcome.
- **category_encoded** contributes the least, implying that the video category plays a minor role in the model.
- Overall, engagement metrics (likes) are more influential than video metadata in predicting the target variable.

Scatter plot of Actual vs Predicted log(video_views) for Tuned Random Forest Model



- **The scatterplot shows that most predicted log(video_views) values align closely with the actual values, following the red diagonal line, indicating good model performance.**
- **Hyperparameter tuning improved the Random Forest's performance, achieving an RMSE of 0.8622 and R^2 of 0.5940, making it a solid model for predicting YouTube video views.**

7. Recommendations:

For Content Creators:

- Focus on maximizing viewer engagement, particularly likes, as they strongly correlate with higher views.
- Encourage interaction through calls-to-action (e.g., “Like this video if you enjoyed it!”).
- While older videos accumulate more views over time, newer videos can still perform well with strong initial engagement.
- Prioritize creating high-quality, engaging content to boost likes early on.
- Categories like "Music" and "Entertainment" tend to have higher views, but the model’s low reliance on
- category_encoded suggests that engagement matters more than category. Focus on content quality over category choice.

For Marketers:

- Identify videos with high like-to-view ratios for advertising campaigns, as these are likely to have broader reach and engagement.
- Target older videos in popular categories (e.g., "Music") for long-term campaigns, as they have accumulated significant views over time.
- Invest in content that drives immediate engagement (likes) to maximize short-term viewership, especially for newer videos.

Conclusion:

- This project successfully developed a **Random Forest model** to predict **YouTube video views**, which achieved an **RMSE of 0.8622** and **R^2 of 0.5940** after **hyperparameter tuning**.
- The model highlights the **importance of viewer engagement (likes)** as the primary driver of views.
- While the model provides actionable insights for content creators and marketers, future work could enhance performance by incorporating additional features, improving category encoding, and addressing imputation biases.