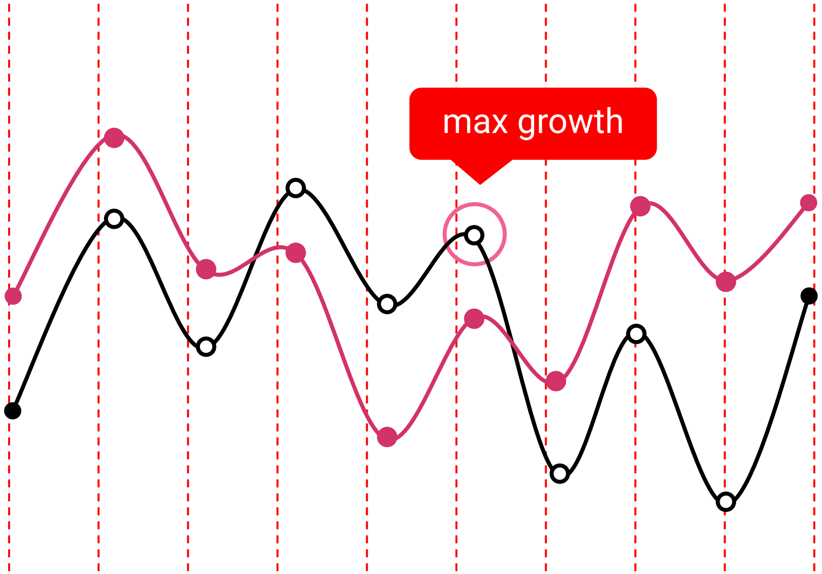
Trending Youtube Video Analysis

By Zirui(Mary) Guo
April 4th, 2023



About Me

A Passionate Data Engineer

- Senior at UC Berkeley
- Major: **DS** & Econ Minor: **CS**
- Incoming Master student at Northwestern University
- Past experience: **Data Engineering Intern** at Rimble; **Data Analyst Intern** at Wing Assistant (Marketing Team)

Overview

01 Background & Data Collection

02 Data Exploration

03 Feature Engineering

04 Model Selection

05 Evaluation & Conclusion

06 Deploy the Model

The Background

Platform Overview

YouTube is a video-sharing platform where users can upload, watch, and share videos. It was founded in 2005, acquired by Google in 2006, and has over 2 billion monthly active users.

Problem statement

- We want to explore what factors make a video become trending
- For the marketing campaign reason, we want select a model, train a model to predict the exact views of a video based on other metadata of the video

Data Collection

Web Scraping: Run the script to web scrape US and Canada Trending Youtube Video at 12:00am every day

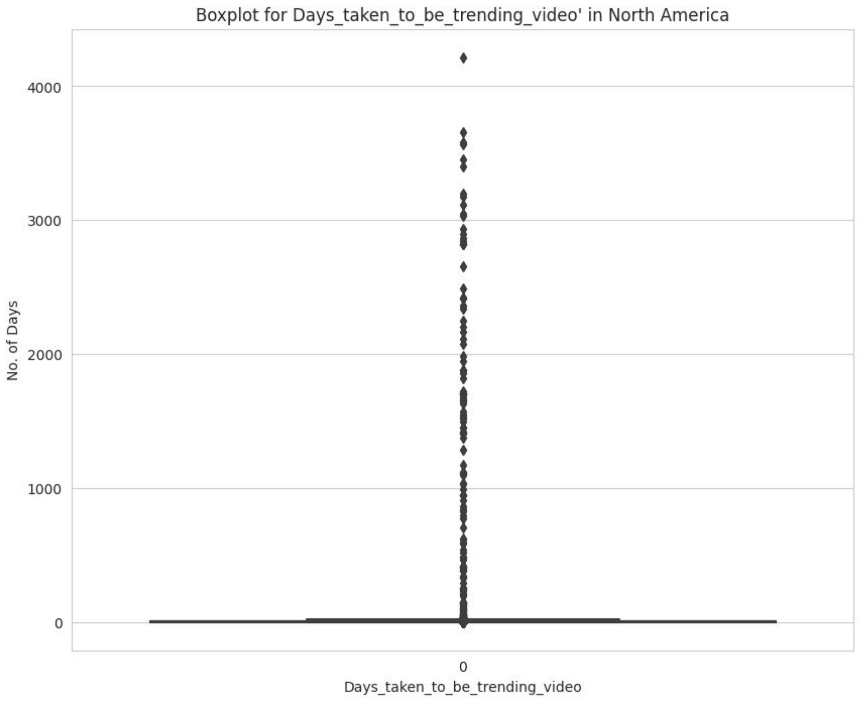
DataSet: Kaggle US and Canada Trending Youtube Video Statistics

Initial Columns: video id, trending date, video title, channel title, category id, publish time, tags, views, likes, dislikes, description, comment count, thumbnail link, comments disabled,, ratings disabled, video error or removed

Data Exploration

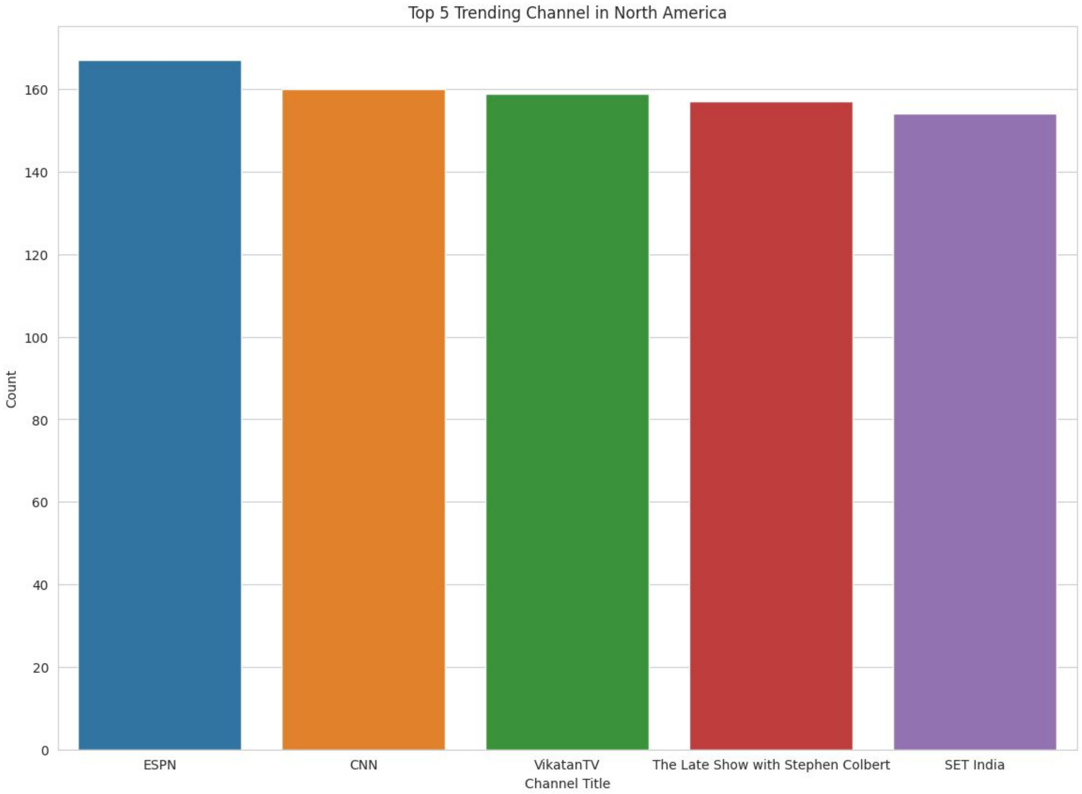
Exploratory Data Analysis - General Observations

Maximum Trending Days V.S. Maximum Days to be Trending



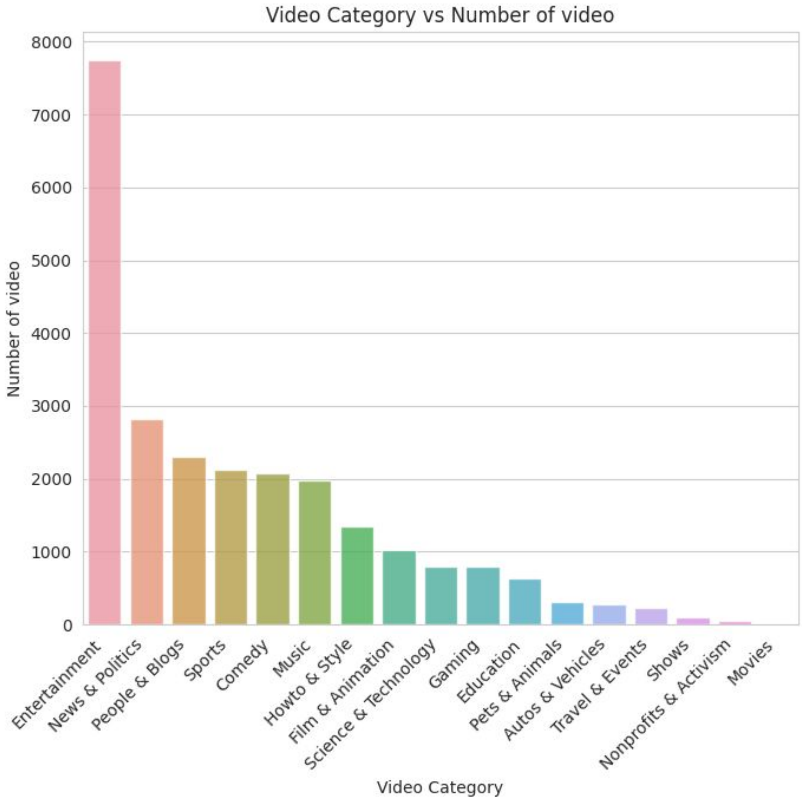
Top 5 Channels with Most of Trending Videos

- Top 5 Channels with most of trending videos in North America are ESPN, CNN, Vikatan TV, The Late Show with Stephen Colbert, SET India
- All of them produced more than 150 trending videos



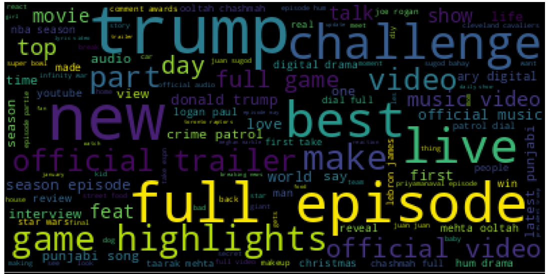
The Distribution of Trending Videos' Categories

- The Entertainment category has the most number of trending videos, followed by News & Politics, then Sports, People & Blogs, and Comedy of all the trending videos

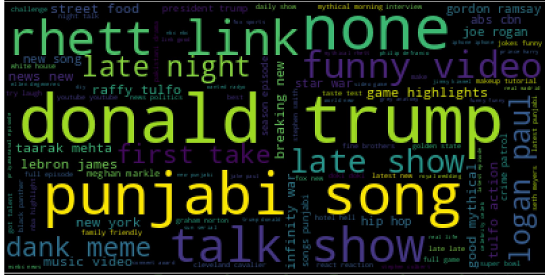


Common words in Trending Video Titles/Tags/Descriptions

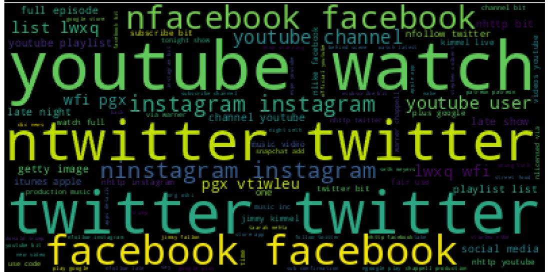
- Most frequently-used words in Titles: Episode, full, trump, official, video, game, new
- Most frequently-used words in Tags: news, show, funny, new, trump, video, comedy
- Most frequently-used words in Descriptions: Youtube, bit, twitter, facebook, instagram, watch, video



(CW in titles)



(CW in tags)



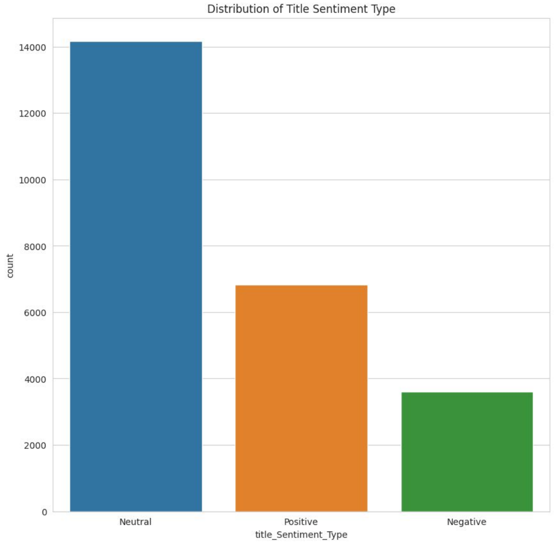
(CW in descs)

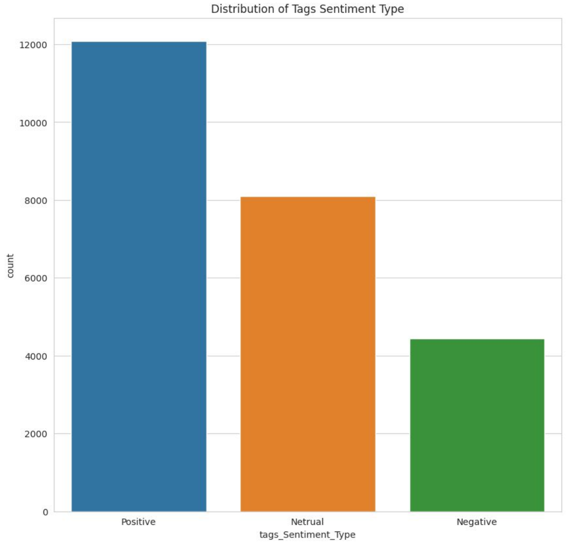
Video Title/ Tags/ Description Sentiment Type Distribution

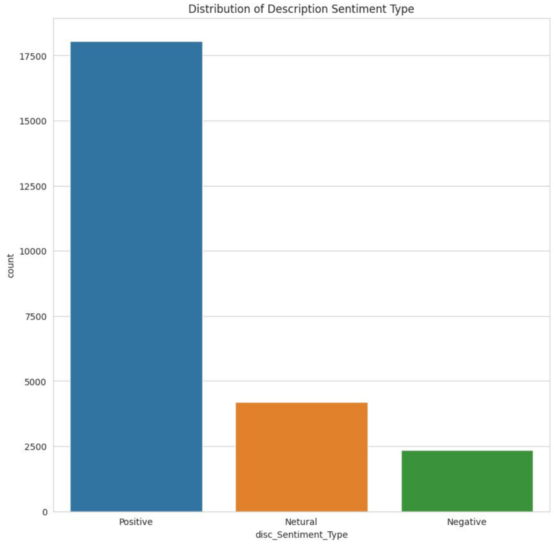
- Trending Video Title Sentiment Type order:
  - Neutral --> Positive --> Negative

- Trending Video Tag Sentiment Type order:
  - Positive --> Neutral --> Negative

- Trending Video Description Sentiment Type order:
  - Positive --> Neutral --> Negative







Feature Engineering

Feature Engineering

Potential Training features

- Country
- Number of Tags
- Length of Description & Length of Title
- Publishing Month/Weekday/Hour
- Sentiment of Title & Tag & Description
- Polarity of Title & Tag & Description

Potential Evaluation features

- Views, likes, dislikes, and Comment counts
- Ratio of views and likes
- Ratio of views and dislikes
- Ratio of views and comment counts
- Ratio of likes and dislikes

Feature Engineering

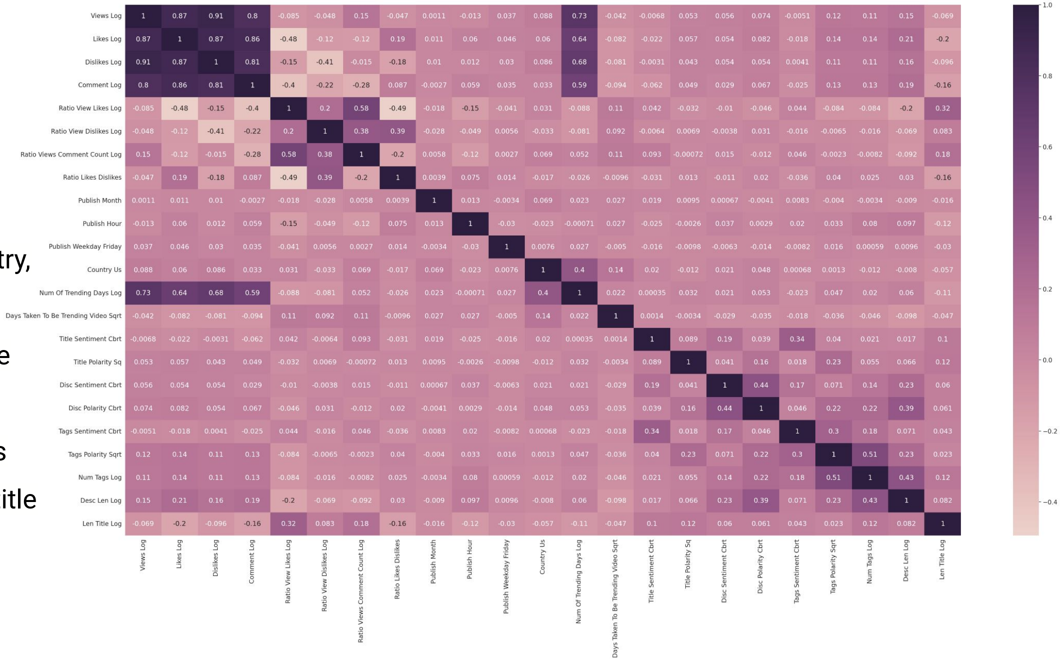
Data Transformation

- **Convert time format**: Trending_Date and Publish Time
- **One Hot Encoding** on categorical variables: Published Weekdays, Months & Hours, Country
- **Log transformation:** views & Likes & Comment Counts & Dislikes
- **Various transformation** (square, square root, cubic, cubic root, log): # of trending days, # of Days taken to be trending video, Title Sentiment, Title Polarity, Desc Sentiment, Disc Polarity, tags Sentiment , tags Polarity, # of tags, length of desc, length of title, Ratio of views and likes, Ratio of View and Dislikes, Ratio of views and comment counts, Ratio of likes and dislikes

Feature Engineering

Feature Importance

- Run Decision Tree Recursive Feature Elimination to select important features
- Top 15 features for log of views prediction (publish month, publish hour, Friday publication, US country, log of # of trending days, square root of days taken to be trending video ,cubic root of title sentiment, square of title polarity, cubic root of disc sentiment, cubic root of disc Polarity, cubic root of tags sentiment, square root of tags polarity, log of number of tags, log of desc length, log of title length)
- Run correlation heatmap

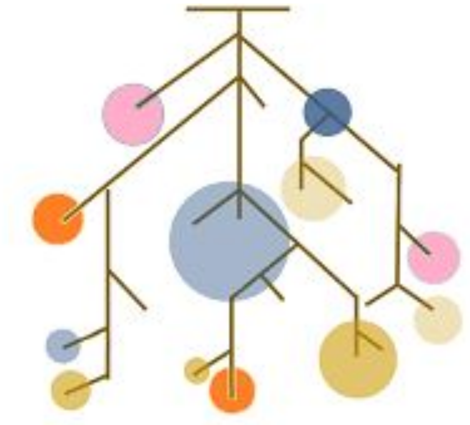


Finally…
Our Data is Ready for Modeling!

Model Selection

Modeling Choices

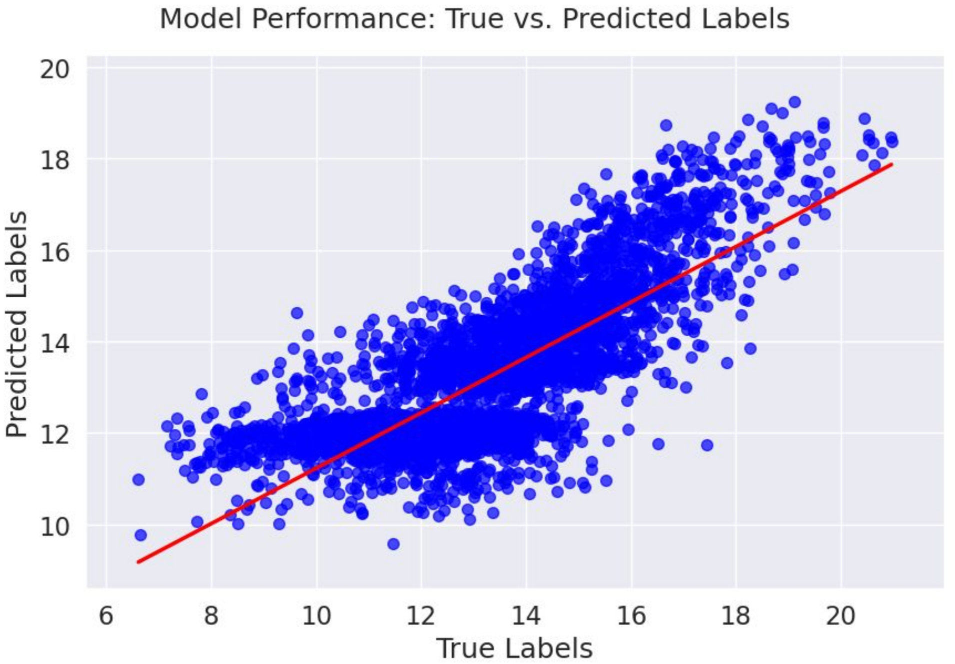
- Training and Test Splitting at 80:20 Ratio
- Modelling choice
    - Supervised
        - Linear Regression
        - Decision Tree
        - Random Forest



Predict Log of Views

Linear Regression

RMSE 1.263
R^2: 0.61
Accuracy: 0.614

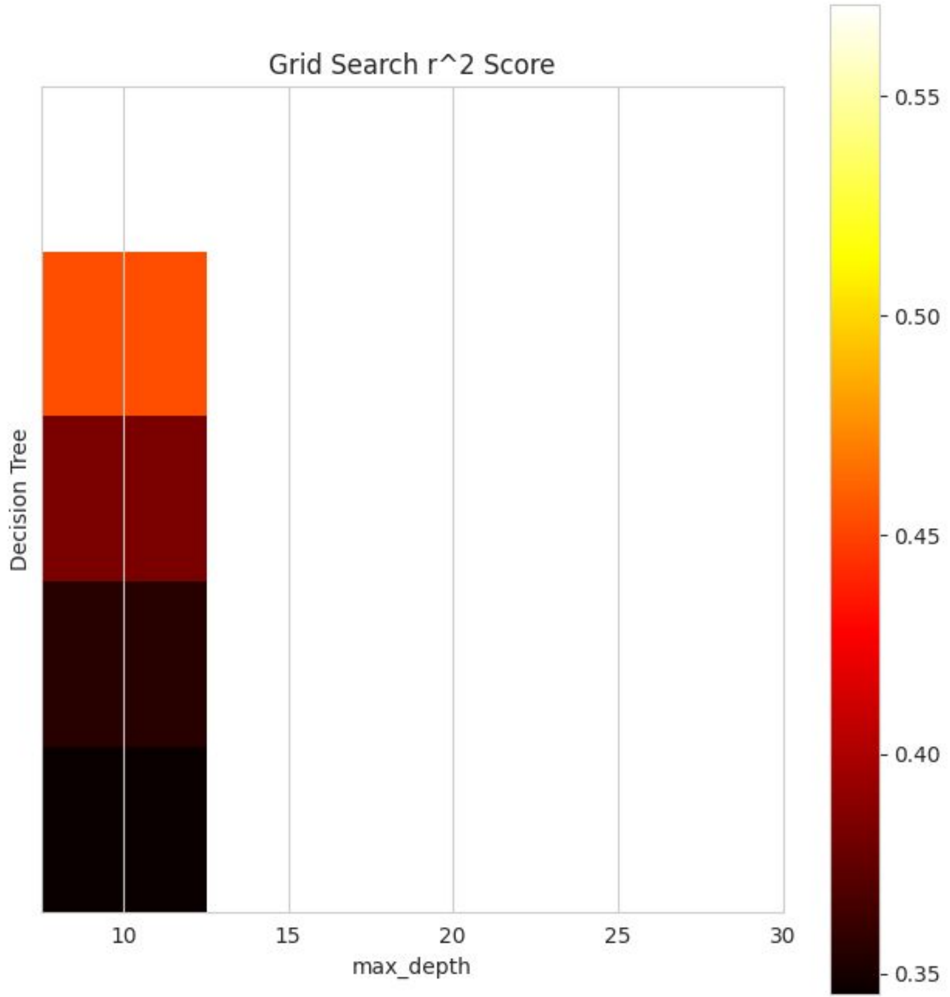


Decision Tree

(Hyper-parameter Tuning)

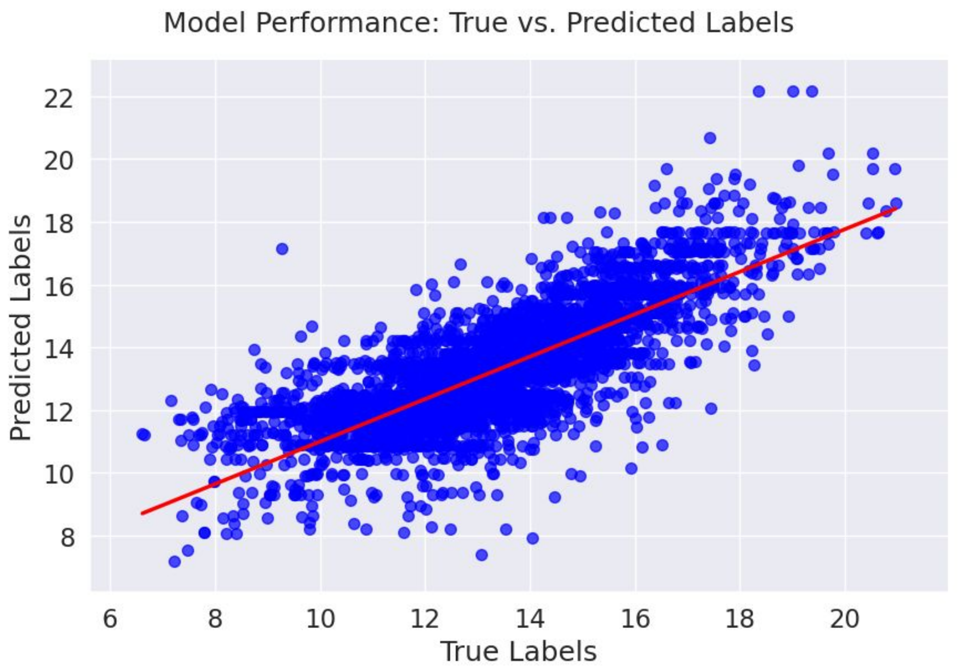
Best HyperParameter:

- max_depth: 10



Decision Tree

RMSE 1.259
R^2: 0.62
Accuracy: 0.617

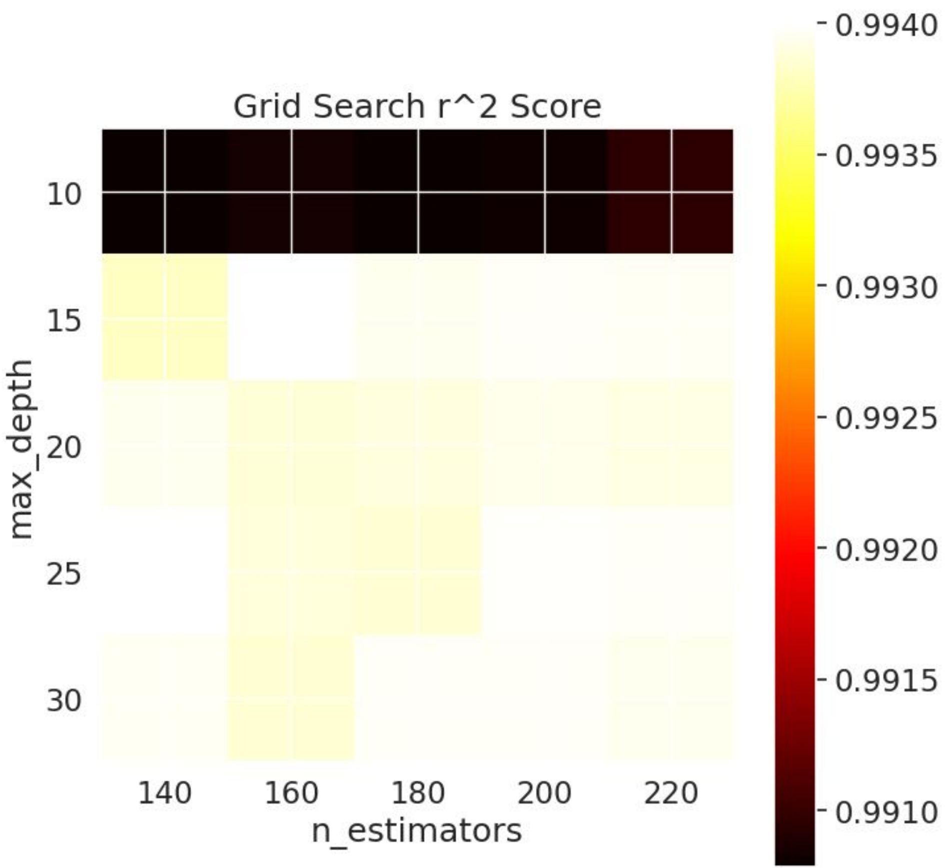


Random Forest

(Hyper-parameter Tuning)

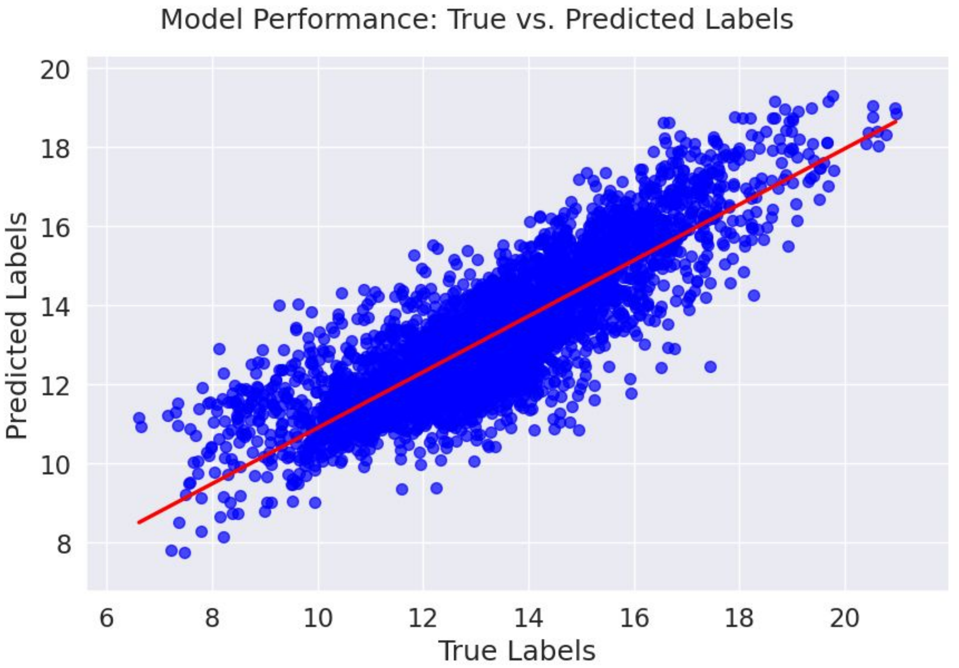
Best HyperParameter:

- max_depth: 25
- n_estimators: 200



Random Forest

RMSE 1.077
R^2: 0.72
Accuracy: 0.719



Model Trade off

Linear Regression

Performance:
Accuracy: 0.61, R^2: 0.61, RMSE: 1.26

Pros:

- Simple Linear Assumption
- Fast to train & Less Computation

Cons:

- Lower Accuracy

Decision Tree

Performance:
Accuracy: 0.62, R^2: 0.62, RMSE: 1.259

Pros:

- Ability to capture non-linear relationships
- Easy to interpret and visualize

Cons:

- Chance of Overfitting
- Accuracy was marginally better than Linear Regression, but not significantly

Random Forest

Performance:
Accuracy: 0.72, R^2: 0.72, RMSE: 1.077

Pros:

- Ensemble Method: Best Performance
- Low chance to overfitting
- Capture complex relationships & robust to outliers

Cons:

- Require expensive computational power

Conclusion

Conclusion

Final Model Choice for this case:

- **Random Forest**
- highest accuracy, highest R^2, and lowest RMSE

If we value interpretability more:

- Linear Regression
- Decision Tree

Limitation

Limiting Data:

- Covers a specific time period
- Limited locations
- Missing data points
- Limited features

Dynamic environment:

- Not account for change overtime

Future Improvement

- Web Scrape more updated metadata related to trending video (Code Appendix)
- Included more features to train the models
- Try other models like gradient boosting and neural network
- Conducted time series analysis for the dynamic changes
- Data Pipeline on Google BigQuery

Future Steps

Deploy the Model

- Save the best trained model to a file
- Create an API for future interactions
- Do local testing through Postman
- Deploy the API to a server like AWS

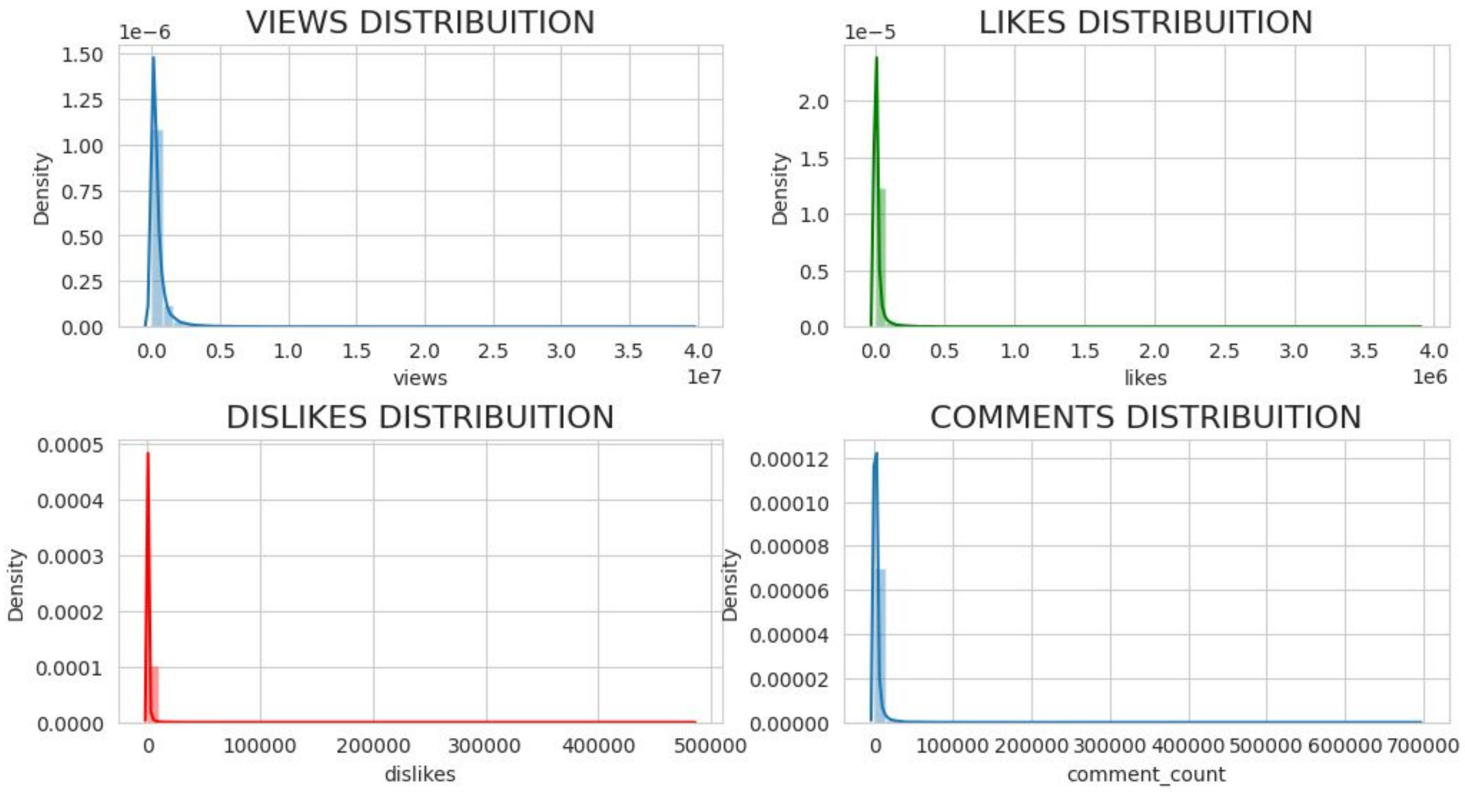
Thank You!

Q&A

Appendix

- Analysis:https://colab.research.google.com/drive/1w_bBQ6SFDWLAmfe5V5XNDH8TySf83XKk?usp=sharing
- Web Scraping: https://colab.research.google.com/drive/1q3ebjryOFRh3nIOP8e3UwduNecCe6mwe?usp=sharing

Views/ Likes/ Comment counts/ Dislikes Distribution



Views/Likes/Comment counts/Dislikes Log Transformation Distribution